THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

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

Regrets optimaux dans les processus de décisions Markoviens

Optimal Regrets in Markov Decision Processes

Présentée par :

Victor BOONE

Direction de thèse :

Bruno GAUJAL DIRECTEUR DE RECHERCHE, INRIA CENTRE GRENOBLE-RHONE-ALPES	Directeur de thèse

Rapporteurs :

Aurélien GARIVIER
PROFESSEUR, ENS de Lyon
Ronald ORTNER
ASSOCIATE PROFESSOR, Montanuniversität Leoben

Thèse soutenue publiquement le **29 novembre 2024**, devant le jury composé de :

Bruno GAUJAL, DIRECTEUR DE RECHERCHE, Centre de l'INRIA de l'Université Grenoble Alpes	Directeur de thèse
Aurélien GARIVIER, PROFESSEUR, ENS de Lyon	Rapporteur
Ronald ORTNER, ASSOCIATE PROFESSOR, Montanuniversität Leoben	Rapporteur
Eric GAUSSIER, PROFESSEUR DES UNIVERSITES, Université Grenoble Alpes	Examinateur
Pierre GAILLARD, CHARGE DE RECHERCHE, Centre INRIA de l'Université Côte d'Azur	Examinateur
Tor LATTIMORE, SENIOR SCIENTIST, DeepMind London	Examinateur

Remerciements

L'écriture d'un manuscrit est un tel investissement que beaucoup y confinent l'intimité de leur auteurice; Ce sont aussi, de fait, des documents pour lesquels il est toléré de prendre certaines libertés qui auraient toujours dues rester inviolées. Cela en fait souvent des documents bien plus intéressants à lire et parcourir que les papiers de recherche, y compris d'un point de vue scientifique. Tout manuscrit qui s'impose contient une section *Remerciements* qui est probablement la plus crue de l'ensemble, où l'auteurice laisse quelques mots à l'intention de ses proches, ami.e.s ou famille, collègues ou parfois juste quelques personnalités croisées sur le chemin. Ces remerciements viennent généralement à la fin du voyage quand bien même ce sont les premiers mots lus à la découverte d'un ouvrage.

Et c'est la fin du voyage pour moi.

Le visage présagé de ce manuscrit a beaucoup évolué lors de cette dernière année et, il y a de cela 10 mois, la majorité des résultats majeurs ci-présents n'existaient tout simplement pas. Leur mise au point est l'aboutissement de trois ans de travail, durant lesquels j'ai forgé une expertise qui m'a permis de venir à bout des preuves les plus épuisantes que j'ai jamais écrites. Ce travail n'aurait pas été possible sans l'indéfectible soutien de mon directeur de thèse, Bruno Gaujal, à qui cette thèse doit beaucoup. Un soutien qui dépasse le cadre strictement professionel, puisque complété d'une forme de complicité dans le rythme de travail, de snobisme culinaire et d'une forme d'amitié née d'une confiance mutuelle. Par ailleurs, je tiens tout autant à remercier Panayotis Mertikopoulos qui aura parfois fait office de co-directeur officieux à défaut d'officiel, et qui m'aura accompagné et épaulé durant de multiples périodes de doutes. Je tiens également à remercier Annie Simon, pour sa compétence et son humanité dans mes heures d'instabilité. Ces trois dernières années, presque quatre, ont été aussi bien courtes que longues, et j'ai eu l'immense chance d'avoir des co-bureaux qui ont été des collègues autant que des amis. Louis-Sébastien, pour m'avoir écouté jacasser à propos de science autant que de tant d'autres sujets et m'avoir épargné moultes journées trop longues en me mettant un café dans les mains. Romain, pour m'avoir traîné tant de fois de mon bureau pour passer du temps à plusieurs autour d'une table et d'un jeu de cartes. Joël, pour toutes ces balades en quête de caféine et pour m'avoir accompagné dans tous mes tâtonnements infructueux à rendre la boisson moins mauvaise. Sebastian, pour sa bonne humeur inébranlable et son sens de la courtoisie. Davide, qui aura veillé maintes fois à ce bureau derrière moi, lorsque les échéances ont été les plus rudes. Beaucoup de gens sont passés par ce bureau 424, pour y vivre plusieurs mois ou années, avant de partir pour ne jamais revenir; et je dois dire que ces départs me pèsent beaucoup plus aujourd'hui, où je cherche désespéremment une forme de stabilité dans le cocon fébrile de mon quotidien. Mes pensées vont à tous mes voisins d'autrefois qui sont partis poursuivre leur route : Louis-Sébastien, Sebastian, Joël, Séhane, Kimang, Chen. Maintes fois ai-je fait le tour des couloirs, à vagabonder dans mes pensées, pour être toujours rappelé par le visage fuyant d'un laboratoire de recherche dont la moitié de l'effectif ne survit guère plus de trois années.

En dernière note, je tiens à remercier tous les collègues que j'ai croisé durant cette aventure, avec qui j'ai travaillé ou parfois simplement partagé quelques pauses café : Zihan, Odalric-Ambrym, Ronan, Nicolas, Mathieu, Aina, Samuel, Flora, Jonatha, Greg et toustes les habitant.e.s

du bâtiment IMAG. Mes remerciements vont également à mes relecteurs : Ronald Ortner, dont les papiers, que j'ai lu avec grande minutie, m'ont permis de construire des intuitions essentielles sur l'apprentissage dans les processus de décision Markoviens, sans lesquelles mes travaux n'auraient pas vu le jour; Et Aurélien Garivier, qui a aussi été mon professeur il y a quelques années, dont l'écriture et la rigueur de la technique m'ont servi maintes fois d'exemples pour la mienne. Vos travaux me servent encore d'exemple aujourd'hui, et je vous suis immemsément reconnaissant d'avoir révisé mon manuscrit. Vos retours signifient beaucoup pour moi.

Et bien évidemment, je ne peux remercier assez mes amis, ma famille, compagnons et compagnes de route et de doutes, sans qui ce manuscrit n'aurait jamais été possible.

Merci à toustes,
Le 20 novembre 2024,
Victor Boone

Contents

Introduction

This document is devoted to learning theory of Markov decision processes.

Say that someone, that shall be called Charlie, interacts with their environment by playing actions. We assume that upon every decision, Charlie has perfect knowledge of the state of the environment that they observe in its entirety. Upon playing, Charlie observes the effect of their played actions by collecting rewards and observing how the system changes states. This evolution (the state change and the produced reward) is stochastic yet memoryless. Specifically, if Charlie sees the state S_t of the environment and decides to play some action A_t, then they see a reward and the new state:

$$R_t \sim r(S_t, A_t) \quad \text{and} \quad S_{t+1} \sim p(S_t, A_t) \tag{1}$$

where $r(S_t, A_t)$ and $p(S_t, A_t)$ are respectively the reward distribution and transition kernel associated to S_t upon playing A_t. The index t tracks the evolution of time. In other words, the reward and state produced by the interaction, despite their random nature, are generated independently of the past and only depend on the current state and played action: This is what is meant when we say that the system is memoryless.

The goal of Charlie is to score maximally, to wander their environments and to do their best at choosing actions. With time, one expects Charlie to play better and better actions, that Charlie *learns* to play optimally. The question that we will investigate throughout this manuscript is: How should Charlie behave? How does one learn such a system as efficiently as possible?

The system that we have just described with (1) is a *Markov decision process* and this will be what models the environment of Charlie in the background. In the sequel, we will make a strong assumption: The underlying environment exists and unflaggingly generates rewards and states for Charlie to observe. This assumption is, in a few areas, so commonly adopted that it is easily forgotten that it is even one. This is a *frequentist* formulation of the learning task, stating that the ground truth exists before Charlie interacts with it, instead of being the by-product of the interaction of Charlie with the environment; The latter is the *Bayesian* formulation. This assumption will help to streamline the discussion and come along with its load of paperwork, painting all this manuscript with the color of frequentism. The possibility of revoking frequentism shall eventually remain open to escape mathematical blinkers, but will mostly be unquestionably adopted to deepdive into the theory.

This pitfall being now clearly sidelined, we may reformulate our prior question: How should Charlie behave in this frequentist formulation of a learning environment?

What this manuscript is about

This manuscript focuses on the motivated learning task. No restriction is put on what Charlie can do, although Charlie may only play legal actions and can never reset the state of their environment (unless there exists an action that does so) that keeps running forever. The rewards that Charlie gathered are compared to what an all knowing planner could achieve with all the

available computational power in the world to prepare their optimal strategy. Hence, the **online** performance of Charlie are compared to those of an optimal **offline** strategy; The difference between these two quantities is called the **regret** and is the reference benchmark in this document. Accordingly and formally, the environment is modeled by an average reward Markov decision process and Charlie's quality of play is measured by the regret. Regret minimization for Markov decision process is the one and only problem that is investigated in the following pages, where we provide a fairly complete (and self-contained) treatment of model dependent and independent settings. We also suggest new directions.

I have learned, over the past few years, that it is foolish to extensively explain what a document does not do; The literature is too vast, communities are too broad and shattered, hence you never know where your reader is coming from. Leaving a few keywords is the best I can do: This manuscript is about (0) the theory of (1) frequentist online learning of (2) average reward Markov decision processes (3) with finite state-action spaces in the (4) model dependent and (5) model independent settings, and mostly with (6) model-based algorithms. The only benchmark is (7) expected regret minimization. It also brings up a few notions such as (8) the regret of exploration and (9) the sliding regret. Every single point has existing alternatives with considerable literature. (0) Experimental and theoretical approaches to reinforcement learning are quite different, and so are the communities. A nice entry to old-school methods that are commonly used in reinforcement learning is Sutton and Barto (2018); these methods are nowadays commonly coupled with deep neural networks to find hidden structures, because practical applications commonly face tremendously large environments. (1) Frequentism is a choice of design for the learning setting, assuming that the underlying model actually exists. The main alternative is the Bayesian approach, where the underlying model is produced by the interaction between the learner and their environment, or equivalently, that the environment is random rather than fixed. (2) The average-reward criterion is not the only one for Markov decision processes, and we can mention the finite horizon criterion, the discounted criterion, total reward criterion, β-entropies, risk-aware settings and so on; again, all of them lead to different learning problems. (3) By choosing the state-action space to be finite, we mean that it is finite and small. This assumption is fragile and fails in many scenarios. For instance the state-space may be infinite (in queuing theory), the state-action space may not be discrete (e.g., the space of policies is parameterized by a parameter living in a continuous space) or the state-action space may be larger than the number of atoms in the universe (e.g., in go or chess). (6) Model-based algorithms learn their environment by estimating its structure. They are not the only learning solution. Other approaches directly estimate the structure of the optimal policy, which is computationally lighter, especially when the underlying environment is large. Such approaches include Monte Carlo (MC) methods, Temporal Difference learning (TD) (see Sutton and Barto (2018)) and the celebrated Q-learning (see Watkins and Dayan (1992); Watkins (1989)), and won't ever be discussed in this document. Lastly, (7) I have chosen to stick to regret guarantees in expectation, which corresponds to Charlie's learning problem, described above. In parallel of regret minimization exists a significative research line on optimal policy identification, and these works are close enough to our setting to be discussed here when times come. So, in our regret minimization setting, the learner-environment system runs until the end of times and the learner is trying to maximize the accumulated rewards. Optimizing the regret in expectation when the system is only run once may sound strange. It is indeed not the only existing choice, and many work studies regret minimization in high probability, a setting that I like less for the simple reason that there is rarely any canonical choice of an acceptable probability of error. Asking for regret guarantees in expectation has links with Bayesian formulations of the learning task that, regrettably, I do not have the time to develop in these pages.

The reference that is the closest in spirit to this manuscript is the book of Lattimore and Szepesvári (2020) which is dedicated to multi-armed bandits.

Now, why this subject in particular? This is a troublesome question. The most candid answer is perhaps that I read the seminal paper of Auer et al. (2009), that I didn't understand it and that my misunderstandings resulted in fractious and inconclusive attempts at improving their algorithm from every angle I could. To this extent, Auer et al. (2009) is a great paper, because it is very well-written and self-contained yet leaves many opportunities for improvements, that I grew obsessed with. Now, I could say that average reward Markov decision processes are a natural generalization of multi-armed bandits, that themselves are a recognized problem that finds many real-word applications. And I could list all these real-worlds applications that Markov decision processes can handle while multi-armed bandits cannot. But let me be honest in the first place. This manuscript is not applied and by no mean ought to be applicable, although many of the presented structural results should inevitably stand as the foundations of any serious efficient and deployable algorithm for Markov decision processes. While some will claim that the lack of applications is a weakness, I would say what they talk about is a different job, a job that requires skills of a different range than those required to establish the results below. And that such skills should be considered specifically rather than additionally. Furthermore, just like it is easy to forget that the frequentist assumption is even an assumption, the question of applications is an assertion in disguise; that a theoretical work is better if directly addressing an application or a real-world problem. Behind this assertion hides a dogma: that theory, or more generally science, should systematically find an application. Or said differently, because this is really the question here, that every work of science should *product* something, that there is a return on investment. We all see how thorny this is all getting. The question of a work's application is in essence following a capitalist and productivist dogma that we should all, as people of science, be at the very least troubled with. Because this question is unquestionably incredibly biased.

Overall, I would just say that this problem is interesting, incredibly rich and doesn't seem detached from reality. I hope you will find it interesting as well.

Outline of the manuscript and highlighted results

The manuscript is split in parts, themselves split in chapters again split into sections and subsections. The graph of dependencies between chapters is given in Figure 1.

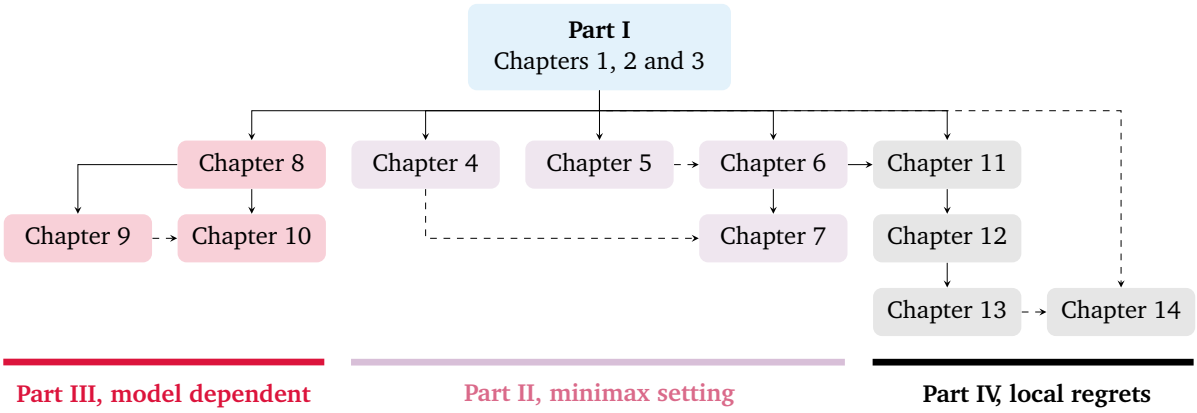


Figure 1: Outline of the manuscript. A plain line arrow means "you should read it before" while a dashed arrow means "it's better to read it before".

The manuscript begins with Part I that introduces the **basic material and concepts**. It starts with a self-contained treatement of average reward Markov decision processes (Chapter 1),

introducing the notion of policies, gain, bias, Poisson and Bellman equations, etc.; It then continues with concepts relative to statistical learning (Chapter 2) and quickly overviews the state-of-the-art in the model dependent and independent settings. The last chapter (Chapter 3) overviews standard tools of the literature, such as changes of measures and concentration inequalities.

Things start with Part II, that treats the model independent setting, also called the **minimax setting**. We provide a state-of-the-art lower bound scaling with the bias span independently of the diameter in Chapter 4. Chapter 5 prepares the next two chapters by taking the time to explain a technique used to bound the gain deviations. Chapter 6 is a more detailed overview of existing algorithms and a complete introduction to the optimism-in-face-of-the-uncertainty principle for Markov decision processes. It pinpoints a few issues in the literature and explains the main challenges of the domain, later addressed in Chapter 7. In the last Chapter 7, we provide a general method that is guaranteed to reach minimax optimal regret, called `PMEVI`.

We continue with Part III, that treats the **model dependent setting**. In Chapter 8, we provide the first general regret lower bound for communicating Markov decision processes. It is shown to be tight in Chapter 10, where we provide an algorithmic scheme (`ECoE`) that reaches the lower bound arbitrarily close. This scheme is not implemented however, and we show in Chapter 9 that the regret lower bound is intractable in general.

Finally, in Part IV, we investigate a new direction: the **local behaviors** of algorithms. We notice that standard methods may play suboptimally for arbitrarily long periods of times, even when they have already enough data to correctly identify the optimal policy. It means that algorithms overshoot the duration of their exploration phases. In Section 11.3, we introduce a new learning metric that measures the performance of learning algorithms when they start exploration phases. In Chapter 12 and Chapter 13, we provide several ways of fixing the existing algorithms, by changing the way episodes are managed, and show that these fixes do not harm the minimax regret guarantees. In the last Chapter 14, we deepdive into the local trajectorial behavior of algorithms in stochastic bandits. We show that the burst of suboptimal plays observed in the beginning of Part IV cannot be removed for optimistic methods and that a form of randomization is required.

Typographic conventions and writing choices

This manuscript is content heavy, especially regarding proofs and techniques. Parts used to be chapters, and chapters used to be sections. It happens that many reader expect sections to be streamlined, when I usually take the time for a few detours, to point subtleties and proof techniques and ideas. Multiplying the number of chapters cleared the required room to embrace my writing style. I estimate that a chapter corresponds to a reading session. This being said, some chapters are short and easy to read (e.g., Chapter 5) and others are very technical and should be approached carefully (e.g., Chapter 10).

There are overall many chapters, many notions and results in this manuscript. To ease the eye, blocks and colors have been used.

> A **blue block** contains **novel** content, that was developped by one of my recent works and is not standard in the literature.

> A **pink block** contains **folklore** content, i.e., results that are known in the literature.

A **plain block** contains **technical** or **secondary** content, i.e., results or notions that are of less importance than emphasized ones. They can be folklore or novel.

A **black block** contains a **remark of first order importance**, i.e., an important assumption or notation or anything that, if missed, can perturb the understanding of the sequel.

Most results of this manuscript are proved, even when they are well known by the community. There are three reasons for this. The first, truthful reason, is that everything is proved out of obsession. I am simply incapable of using a result that I do not understand at least barely. The second, perhaps more valuable to the reader, is that I could uniformize, unify and systematize a technique while providing a fairly complete view of undiscounted infinite horizon reinforcement learning. This technique is a systematic use of Poisson equations, Bellman equations, reward and transition transforms, and careful choices of stopping times. Combined, these four ingredients usually provide better results that what I could obtain with algebraic methods, that I ended up dropping entirely. The third reason is obsession again, since by following the second reason alone, I ended up with 90% of proved results. So I went for a near 100%, to obtain a complete, self-contained document on the current state-of-the-art of small sized and undiscounted infinite horizon reinforcement learning.

Even after all this work, with a satisfying conclusion to optimistic methods, after establishing the first complete regret lower bound in communicating Markov decision processes, I feel that a lot of work has still to be done. Yet every PhD ends at some point. It ends by freezing a view of a comprehensible problem in a voluminous manuscript. So I've put aside many aspects of my work that are too immature to see the light of day.

When one writes a scientific document of such size, one hopes that it will be of help for others and that it will survive for a while, at least a little bit. So, what is the scientific purpose of a PhD manuscript? I simply cannot accept that this is merely a proof of knowledge and skills.

Part I

Learning Markov Decision Processes

In this part, we present the fundamental material of the manuscript: Markov decision processes and statistical learning theory. Chapter 1 is dedicated to Markov decision processes in the average reward setting, introducing the notion of policies, gain, bias, Bellman gaps, optimal policies, as well as algorithms to compute them. This introduction is self-contained and only requires basic knowledge on probability theory and martingales. In Chapter 2, we provide concepts from statistical decision theory, eventually instantiated to Markov decision processes. We provide a formal definition of the regret, discuss the state-of-the-art in the model independent (or minimax) and model dependent settings, and further pinpoint the main contributions of my work regarding these two domains. The last Chapter 3 gathers a few classical results that are nowadays the mandatory tools to design and analyze any learning algorithm for Markov decision processes.

General Notations. Throughout the manuscript, the standard fields of numbers (natural integers, signed integers, rational, reals) are denoted $\mathbf{N}, \mathbf{Z}, \mathbf{Q}, \mathbf{R}$ respectively. Calligraphic letters $\mathscr{A}, \mathscr{B}, \ldots$ denote sets, the power set of $\mathscr{A}$ is denoted $2^{\mathscr{A}}$ and if $\mathscr{A}$ is a Borel set, $\mathscr{P}(\mathscr{A})$ denotes the Borel probability distributions on $\mathscr{A}$. Probability and expectation operators are written $\mathbf{P}(-), \mathbf{E}[-]$. $\|-\|_p$ is the p-norm, $\mathrm{KL}(-||-)$ is the Kullback-Leibler divergence and $\mathbf{1}(-)$ is the indicator function. The letter e denotes the vector full of ones, e_i is the i-th element of the canonical basis, and the dot product between a co-vector p and a vector u is written pu, $p \cdot u$ or even $\langle p, u \rangle$ depending on the sensibility of the surrounding equation to ambiguities. The typography of function arguments is semantic rather than positional and a given letter will stick to a unique semantic as much as possible. A function argument can momentarily move from parenthesis to index if typographically more convenient although disclaimers are added as much as possible to avoid confusion.

Chapter 1

Foundations of Markov Decision Processes

This chapter introduces general material for Markov decision process in the average reward setting. This introduction is far from exhaustive and more is omitted than the converse. This lack of exhaustiveness is a choice. The content of this section, together with the presented technique, depicts a landscape of the theory of average reward Markov decision processes that is complete enough to support the addition of learning considerations. We begin with a definition.

Definition I.1 (Markov decision process)**.** *A Markov decision process $M = (\mathscr{S}, \mathscr{A}, p, r)$ is given by (1) a state space $\mathscr{S}$, (2) an action space $\mathscr{A} \equiv \bigcup_{s\in\mathscr{S}} \mathscr{A}(s)$ together inducing a pair space $\mathscr{Z} := \bigcup_{s\in\mathscr{S}} \{s\} \times \mathscr{A}(s)$, (3) a transition kernel $p : \mathscr{Z} \to \mathscr{P}(\mathscr{S})$ and (4) a reward function $r : \mathscr{Z} \to \mathscr{P}(\mathbf{R})$.*

We say that a Markov decision process (shortened as **MDP** or **model**) is **finite** if its pair space is finite. It is said **tabular** if $|\mathscr{A}(s)|$ is the same for all $s \in \mathscr{S}$. From now on, if not specified otherwise, all the considered models will be finite. Compact action spaces will be a concern in Parts II and IV but are cast out for now.

The model is typically controlled by a mechanism called "the controller" that picks legal actions at each time step. The t-step state, action and reward are respectively denoted

$$S_t, A_t, R_t \tag{I.1}$$

and we use the shorthand $Z_t := (S_t, A_t)$ for the t-step played pair, with $t \in \mathbf{N}$. The **history of play** at time t is the aggregation of all the observations prior to time t, and it is written $O_t := (S_0, A_0, R_0, \dots, S_t)$. The played action A_t is a function of O_t plus a possible extra-randomness ω. Formally, A_t is $\sigma(O_t, \omega)$-measurable. The observed reward and states satisfy the **Markov property** (I.2), meaning that they only depend on the current played state and action, rather than on the full history and future ahead. Accordingly, the underlying stochastic model satisfies:

$$\begin{aligned} \mathbf{P}\big((R_t, S_{t+1}) \in \mathscr{U} \times \mathscr{S}' \mid O_t, A_t\big) &= \mathbf{P}\big((R_t, S_{t+1}) \in \mathscr{U} \times \mathscr{S}' \mid S_t, A_t\big) \\ &= \int_{\mathscr{U}\times\mathscr{S}'} r(u|S_t, A_t) p(s'|S_t, A_t)\, d(u, s'). \end{aligned} \tag{I.2}$$

Remark that in (I.2), we implicitly assume that rewards and states are sampled independently. This is not important. Instead, we could define $q : \mathscr{Z} \to \mathscr{P}(\mathbf{R} \times \mathscr{S})$ a joint probability distribution on rewards and states with marginals r and p and all the results of this manuscript would hold

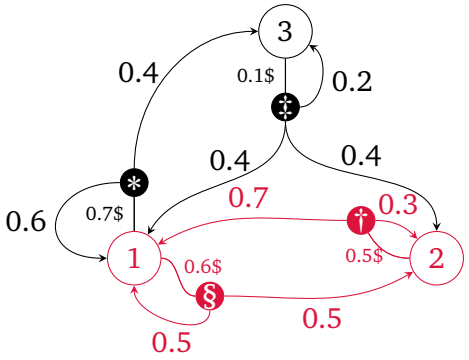


Figure 1.1: This is how MDPs will typically be represented in this manuscript. States are circled nodes, actions are filled symbolic disks going out from states. Transition probabilities are represented with arrows weighted by the probability of the transition, and mean rewards are the weight (in dollars) on undirected edges between states and actions. The dollar symbol ($) is used to break ambiguities between transition probabilities and rewards.

similarly. It is well known that, if rewards take finitely many values, a model transform can convert back rewards and transitions to be independent, see Puterman (1994).

> **Important remark.** All throughout Chapter 1, the underlying Markov decision process, M, is fixed. Also, the dependency in M of the dynamics and various quantities that we introduced is ignored within notations. This is out of typographic convenience.

1.1 Policies, gain, bias and the Poisson equation

The notion of policy is a simple way to model "a way to select actions". In this section, we provide a complete and self-contained introduction to the material required to properly understand the behavior of a policy's iterates.

> **Definition I.2** (Policies and randomized policies)**.** *A **policy** ($\pi \in \Pi$) is a map $\pi : \mathcal{S} \to \mathcal{A}$. A **randomized** policy ($\pi \in \Pi^{\mathrm{SR}}$) is a map $\pi : s \in \mathcal{S} \mapsto \pi(-|s) \in \mathscr{P}(\mathcal{A}(s))$. If ambiguous, the term policy always refer to deterministic policies.*

It is also possible to make policies dependent on the histroy of play, making them similar to the controller mentioned upstream. These **history dependent policies** will be discussed in more detail in Chapter 2.

When the dynamics of the model are driven by a fixed (possibly randomized) policy $\pi \in \Pi^{\mathrm{SR}}$, the laws of the sequence of states, actions and rewards are completely determined by π and the initial state s. When the distribution of the initial state is ν, the associated probability and expectation operators under the dynamics imposed by π will be denoted $\mathbf{P}_\nu^\pi(-)$ and $\mathbf{E}_\nu^\pi[-]$. By driving the dynamics with a fixed policy, the process becomes a **Markov reward process**, which is merely a Markov chain on $\mathcal{S}$ with an additive functional. Because it defines a Markov chain, the adequate terminology can be imported from this theory, see Levin and Peres (2017) for a general reference. This terminology will be crucial in the sequel.

Definition I.3 (Hitting time, recurrence, classification)**.** *Fix a randomized policy $\pi \in \Pi^{\mathrm{SR}}$.*

(1) *The **hitting time** to $\mathcal{S}' \subseteq \mathcal{S}$ is $\tau^0_{\mathcal{S}'} := \inf\{t \geq 0 : S_t \in \mathcal{S}'\}$;*

(2) *The **positive hitting time** to $\mathscr{S}' \subseteq \mathscr{S}$ is $\tau_{\mathscr{S}'} := \inf\{t \geq 1 : S_t \in \mathscr{S}'\}$;*
(3) *A state s is **recurrent** under π if $\mathbf{P}_s^\pi(\forall n, \exists m > n : S_m = s) = 1$, or equivalently for finite state space models, if $\mathbf{E}_s^\pi[\tau_s] < \infty$; Recurrent states form **components** induced by the equivalence relation $s \sim s'$ if $\mathbf{E}_s^\pi[\tau_{s'}] < \infty$; Non-recurrent states are said **transient**;*
(4) *A policy is **recurrent** if all its states are recurrent and in the same component; A policy is **unichain** if it has a unique recurrent component; A policy is **multi-chain** otherwise.*

The claims that are inherent to the above definition are standard and fairly easy to prove, hence their proofs are omitted. Refer to Levin and Peres (2017) for further details.

A policy also induces a kernel and a reward vector. The definition below introduces notations.

Definition I.4 (Policy kernel and reward function)**.** *For $\pi \in \Pi^{\mathrm{SR}}$, we define*

(1) *its **kernel** by $p^\pi(s'|s) = \sum_{a \in \mathscr{A}(s)} p(s'|s,a)\pi(a|s)$; and*
(2) *its **reward function** is given by $r^\pi(s) := \sum_{a \in \mathscr{A}(s)} r(s,a)\pi(a|s)$.*

We further write P^π the transition matrix associated to the kernel p^π.

1.1.1 The gain of a policy

In this manuscript, we are interested in the **total rewards**, consisting in the expected aggregate rewards $\mathbf{E}_s^\pi[\sum_{t=0}^{T-1} R_t]$ when the horizon T goes to infinity, without discount. This sum, as shown downstream, is known to behave like $Tg^\pi(s) + h^\pi(s) + \mathrm{o}(1)$ where $g^\pi(s), h^\pi(s) \in \mathbf{R}$ are quantities respectively characterizing the first and second order growth of aggregate rewards and are respectively called the **gain** and **bias** of the policy. We show first that the gain is well defined.

Definition I.5. *The **gain** of a policy $\pi \in \Pi^{\mathrm{SR}}$ is the vector $g^\pi \equiv g^\pi(M)$ given by:*

$$g^\pi(s) := \lim_{T\to\infty} \frac{1}{T}\mathbf{E}_s^\pi\left[\sum_{t=0}^{T-1} R_t\right] = \lim_{T\to\infty} \frac{1}{T}\mathbf{E}_s^\pi\left[\sum_{t=0}^{T-1} r(Z_t)\right]. \tag{I.3}$$

This is not clear that the limit exists, so we provide a proof below. Remark that, if well defined, the gain satisfies $P^\pi g^\pi = g^\pi$, so is within the (right) kernel of $I - P^\pi$. The above result can also be written in matrix form with using that $\mathbf{E}_s^\pi[\sum_{t=0}^{T-1} R_t] = (\sum_{t=0}^{T-1}(P^\pi)^t r^\pi)(s)$.

Proof. We show first that the limit exists for recurrent states, so pick $s \in \mathscr{S}$ recurrent. Let $g_+^\pi(s) := \limsup \frac{1}{T}\mathbf{E}_s^\pi[\sum_{t=0}^{T-1} R_t]$ the upper gain and $g_-^\pi(s) := \liminf \frac{1}{T}\mathbf{E}_s^\pi[\sum_{t=0}^{T-1} R_t]$ the lower gain. Let $\epsilon > 0$. There exists T_ϵ, as large as desired, such that

$$\mathbf{E}_s^\pi\left[\sum_{t=0}^{T_\epsilon-1} R_t\right] \geq T_\epsilon(g_+^\pi(s) - \epsilon).$$

Observe the following: If $\nu \in \mathscr{P}(\mathscr{S})$ is supported on the component containing s, then

$$\begin{aligned}\mathbf{E}_\nu^\pi\left[\sum_{t=0}^{T_\epsilon-1} R_t\right] &\geq \mathbf{E}_\nu^\pi\left[\sum_{t=\tau_s}^{T_\epsilon-1} R_t\right] - \mathbf{E}_\nu^\pi[\tau_s]\|r\|_\infty \geq \mathbf{E}_\nu^\pi\left[\sum_{t=\tau_s}^{T_\epsilon+\tau_s-1} R_t\right] - 2\mathbf{E}_\nu^\pi[\tau_s]\|r\|_\infty \\ &\overset{(\dagger)}{=} \mathbf{E}_s^\pi\left[\sum_{t=0}^{T_\epsilon-1} R_t\right] - 2\mathbf{E}_\nu^\pi[\tau_s]\|r\|_\infty \geq T_\epsilon\left(g_+^\pi(s) - \epsilon - \frac{2\mathbf{E}_\nu^\pi[\tau_s]\|r\|_\infty}{T_\epsilon}\right)\end{aligned} \tag{I.4}$$

where (†) follows from the Markov property (I.2). Remark that $\mathbf{E}_\nu^\pi[\tau_s] \leq \max_{s'} \mathbf{E}_{s'}^\pi[\tau_s] =: D_s < \infty$ where s' goes other the states in the same recurrent component than s. With this in mind,

let $T \in \mathbf{N}$ and do the euclidian division of T with T_ϵ, i.e., $T = nT_\epsilon + m$ with $m < T_\epsilon$. Using (I.4) n times, we see that for all s' in the same component than s, we have:

$$\mathbf{E}_{s'}^{\pi}\left[\sum_{t=0}^{T-1} R_t\right] \geq nT_\epsilon\left(g_+^\pi(s) - \epsilon - \frac{2D_s\|r\|_\infty}{T_\epsilon}\right) \sim T\left(g_+^\pi(s) - \epsilon - \frac{2D_s\|r\|_\infty}{T_\epsilon}\right). \tag{I.5}$$

Let $T \to \infty$, then $T_\epsilon \to \infty$ and $\epsilon \to 0$. We conclude that $\liminf \mathbf{E}_s^\pi[\sum_{t=0}^{T-1} R_t] \geq g_+^\pi(s)$, hence the infemum limit and the limsup are equal and the gain is well defined for s recurrent. In fact, from (I.5) follows that $g_-^\pi(s') \geq g_+^\pi(s)$ for every s' in the same recurrent component than s. By symmetry on s, s', we deduce that $g^\pi(s) = g^\pi(s')$.

To extend the well-definition of (I.3) to transient states, the gain from transient states is expressed with respect to the gain on recurrent states. Let $\mathcal{S}_1, \dots, \mathcal{S}_m \subseteq \mathcal{S}$ the disjoint recurrent components under π and pick a witness $s_i \in \mathcal{S}_i$ for each. Let $\tau := \tau_{\{s_1,\dots,s_m\}}$ the hitting time to $\{s_1, \dots, s_m\}$ and observe that $\mathbf{E}_s^\pi[\tau] < \infty$ for all $s \in \mathcal{S}$, and that $\mathbf{P}_s^\pi(S_\tau \in \mathcal{S}_i)$ is the probability that S_t eventually gets trapped in the component $\mathcal{S}_i$ under π starting from s. We have:

$$\frac{1}{T}\mathbf{E}_s^\pi\left[\sum_{t=0}^{T-1} R_t\right] = \frac{1}{T}\mathbf{E}_s^\pi\left[\sum_{t=0}^{\tau\wedge T-1} R_t + \sum_{i=1}^{m} \mathbf{1}(S_\tau \in \mathcal{S}_i)\sum_{t=\tau}^{T-1} R_t\right] = \sum_{i=1}^{m} \mathbf{P}_s^\pi(S_\tau \in \mathcal{S}_i) g^\pi(s_i) + \mathrm{O}(1) \tag{I.6}$$

concluding the proof. □

1.1.2 Invariant measures, regeneration and empirical measures

The gain of a policy is intimately linked to the invariant measures of the same policy. By changing the reward to the probing function $r(s) = \mathbf{1}(s = s_0)$ for some fixed $s_0 \in \mathcal{S}$, the associated gain becomes the asymptotic ratio of visits which is a "natural" invariant measure of the policy.

Definition I.6. *A (state-wise) **invariant measure** of a policy $\pi \in \Pi^{\mathrm{SR}}$ is a co-vector $\mu \in \mathbf{R}^{\mathcal{S}}$ such that $\mu \cdot P^\pi = \mu$, or equivalently, such that $\sum_{s\in\mathcal{S}} \mu(s)p^\pi(s'|s) = \mu(s')$ for all $s' \in \mathcal{S}$. The **asymptotic empirical measure** of a policy $\pi \in \Pi^{\mathrm{SR}}$ from an initial state $s \in \mathcal{S}$ is given by:*

$$\mu^\pi(s'|s) := \lim_{T\to\infty} \frac{1}{T}\mathbf{E}_s^\pi\left[\sum_{t=0}^{T-1} \mathbf{1}(S_t = s')\right]. \tag{I.7}$$

In other words, invariant measures are elements of the (left) kernel of $I - P^\pi$ which a linear space with dimension equal to the number of recurrent components. Observe that well-definition of $\mu^\pi(-|s)$ is a consequence of Definition I.5, proving by the meantime that invariant measures exist. The above limit can also be written in matrix form by using the formula $\mathbf{E}_s^\pi[\sum_{t=0}^{T-1} \mathbf{1}(S_t = s')] = (\sum_{t=0}^{T-1}(P^\pi)^t e_{s'})(s)$. We provide a few useful results on invariant measures below.

Proposition I.1 (Levin and Peres (2017))**.** *Fix a randomized policy $\pi \in \Pi^{\mathrm{SR}}$. Let $\mathcal{S}_1, \dots, \mathcal{S}_m$ its (disjoint) recurrent components.*

1. *Given $i \in \{1, \dots, m\}$, the unique probability invariant measure supported in $\mathcal{S}_i$ is $\mu^\pi(-|s_i)$ for $s_i \in \mathcal{S}_i$ chosen arbitrarily; It is written $\mu^\pi(-|\mathcal{S}_i)$;*
2. *For all $s \in \mathcal{S}$, $\mu^\pi(-|s) = \sum_{i=1}^m \mathbf{P}_s^\pi(\tau_{\mathcal{S}_i} < \infty)\mu^\pi(-|\mathcal{S}_i)$;*
3. *(Regeneration property) For $i \in \{1, \dots, m\}$ and $s_i \in \mathcal{S}_i$, we have*

$$\forall s \in \mathcal{S}_i, \quad \mu^\pi(s|s_i) = \frac{1}{\mathbf{E}_{s_i}^\pi[\tau_{s_i}]}\mathbf{E}_{s_i}^\pi\left[\sum_{t=0}^{\tau_{s_i}-1} \mathbf{1}(S_t = s)\right]. \tag{I.8}$$

Proof. Using $\mu P^\pi = \mu$, by induction on T follows that $\mathbf{E}^\pi_\mu[\sum_{t=0}^{T-1} \mathbf{1}(S_t = s)] = T\mu(s)$ for all $s \in \mathscr{S}$. Statement (1) then follows as such: Fix $s_i \in \mathscr{S}_i$ and remark that for all $s \in \mathscr{S}_i$,

$$T\mu(s) = \mathbf{E}^\pi_\mu\left[\sum_{t=0}^{\tau_{s_i}\wedge T-1} \mathbf{1}(S_t = s) + \sum_{t=\tau_{s_i}}^{T-1} \mathbf{1}(S_t = s)\right] = \mathrm{O}(1) + T\mu^\pi(s|s_i) + \mathrm{o}(T).$$

Statement (2) is a rewriting of (I.6). For the regeneration property (Statement (3)), introduce $\mu(s) := \mathbf{E}^\pi_{s_i}[\sum_{t=0}^{\tau_{s_i}-1} \mathbf{1}(S_t = s)]$. We show that $\mu P^\pi = \mu$. Let $s \in \mathscr{S}$. We have:

$$\begin{aligned}
\sum_{s'\in\mathscr{S}} p^\pi(s|s')\mu(s') := \sum_{s'\in\mathscr{S}} p^\pi(s|s')\mathbf{E}^\pi_{s_i}\left[\sum_{t=0}^{\tau_{s_i}-1} \mathbf{1}(S_t = s')\right] &= \mathbf{E}^\pi_{s_i}\left[\sum_{t=0}^{\tau_{s_i}-1}\sum_{s'\in\mathscr{S}} p^\pi(s|s')\mathbf{1}(S_t = s')\right] \\
&= \mathbf{E}^\pi_{s_i}\left[\sum_{t=0}^{\tau_{s_i}-1} \mathbf{1}(S_{t+1} = s)\right] \\
&= \mathbf{E}^\pi_{s_i}\left[\sum_{t=0}^{\tau_{s_i}-1} \mathbf{1}(S_t = s)\right] + \mathbf{E}^\pi_{s_i}[\mathbf{1}(S_{t=1} = s) - \mathbf{1}(S_0 = s)] \\
&= \mathbf{E}^\pi_{s_i}\left[\sum_{t=0}^{\tau_{s_i}-1} \mathbf{1}(S_t = s)\right] = \mu(s).
\end{aligned}$$

By (1), μ is therefore proportional to $\mu^\pi(-|\mathscr{S}_i)$. We conclude by normalizing. □

The next result links asymptotic empirical measures to the gain.

Proposition I.2. *For every randomized policy, we have $g^\pi(s) = \mu^\pi(-|s) \cdot r^\pi$.*

Proof. We have:

$$g^\pi(s) := \lim_{T\to\infty} \frac{1}{T}\mathbf{E}^\pi_s\left[\sum_{t=0}^{T-1} R_t\right] = \lim_{T\to\infty} \frac{1}{T}\mathbf{E}^\pi_s\left[\sum_{t=0}^{T-1}\sum_{s'\in\mathscr{S}} \mathbf{1}(S_t = s') \cdot r^\pi(s')\right] = \mu^\pi(-|s) \cdot r^\pi \qquad \text{(I.9)}$$

proving the claim. □

1.1.3 The bias of a policy and the Poisson equation

In Definition I.5, it has been shown that the gain is the first term in the expansion of rewards collected by a policy, with the formula $\mathbf{E}^\pi_s[\sum_{t=0}^{T-1} R_t] = Tg^\pi(s) + \mathrm{o}(T)$. For several reasons, it is quite insufficient to provide a satisfying understanding of how $\mathbf{E}^\pi_s[\sum_{t=0}^{T-1} R_t]$ behaves. Justice cannot be made to the importance of the second order term of the expansion without advancing the theory a little bit. However, its foreground role is foreshadowed by the example of Figure 1.2.

In Figure 1.2, there are two policies that have both the same gain $g^\pi(s) = 1$ for $s \in \{1, 2\}$. However, the policy taking the action (†) is clearly better than the one taking the action (‡) from a transient viewpoint, by scoring 10 instead of 0 in the first round. This transient advantage is encoded by the **bias** (sometimes called **potential**) of the policy, that encodes what the policy scores in addition to the gain. Together, the gain and the bias of a policy form what is known as the **Poisson equation**.

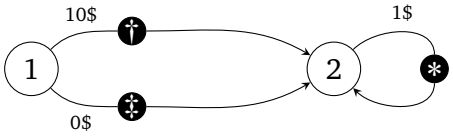


Figure 1.2: Beyond the gain: (†) is undoubtedly a better action than (‡).

Definition I.7. *Let $\pi \in \Pi^{\mathrm{SR}}$. There exists a **bias function** h^π such that $\mu^\pi(-|s)\cdot h^\pi = 0$ for all $s \in \mathcal{S}$ that satisfies the **Poisson equation**:*

$$\forall s \in \mathcal{S}, \quad g^\pi(s) + h^\pi(s) = r^\pi(s) + p^\pi(s)h^\pi \tag{I.10}$$

Proof. Let $\mathcal{S}_1, \dots, \mathcal{S}_m$ the disjoint recurrent components of π and pick a witness $s_i \in \mathcal{S}_i$ for each. Let τ the positive hitting time to $\{s_1, \dots, s_m\}$, i.e., $\tau := \inf\{t \geq 1 : S_t \in \{s_1, \dots, s_m\}\}$. Let $\alpha \in \mathbf{R}^m$ and introduce the following quantity:

$$h^\pi_\alpha(s) := \mathbf{E}^\pi_s\left[\sum_{t=0}^{\tau-1}(R_t - g^\pi(S_t)) + \sum_{i=1}^{m}\alpha_i \mathbf{1}(S_\tau = s_i)\right]. \tag{I.11}$$

The vector h^π_α is a linear function of α. We check that (I.10) is satisfied by h^π_α.

$$\begin{aligned}
(*) &:= p^\pi(s)h^\pi_\alpha - h^\pi_\alpha(s) \\
&= \sum_{s'\in\mathcal{S}} p^\pi(s'|s)h^\pi_\alpha(s') - \mathbf{E}^\pi_s\left[\sum_{t=0}^{\tau-1}(R_t - g^\pi(S_t)) + \sum_{i=1}^{m}\alpha_i\mathbf{1}(S_\tau = s_i)\right] \\
&= g^\pi(s) - r^\pi(s) + \sum_{s'\in\mathcal{S}} p^\pi(s'|s)\left(h^\pi_\alpha(s') - \mathbf{E}^\pi_s\left[\sum_{t=1}^{\tau-1}(R_t - g^\pi(S_t)) + \sum_{i=1}^{m}\alpha_i\mathbf{1}(S_\tau = s_i)\middle| S_1 = s'\right]\right).
\end{aligned}$$

The summand is null. Indeed, if $s' \notin \{s_1, \dots, s_m\}$ then

$$\mathbf{E}^\pi_{s'}\left[\sum_{t=0}^{\tau-1}(R_t - g^\pi(S_t)) + \sum_{i=1}^{m}\alpha_i\mathbf{1}(S_\tau = s_i)\right] = \mathbf{E}^\pi_s\left[\sum_{t=1}^{\tau-1}(R_t - g^\pi(S_t)) + \sum_{i=1}^{m}\alpha_i\mathbf{1}(S_\tau = s_i)\middle| S_1 = s'\right].$$

If $s' = s_i$, then by (I.11) we see that:

$$\begin{aligned}
h^\pi_\alpha(s_i) &= \mathbf{E}^\pi_{s_i}\left[\sum_{t=0}^{\tau-1}(R_t - g^\pi(S_t)) + \sum_{i=1}^{m}\alpha_i\mathbf{1}(S_\tau = s_i)\right] \\
&\overset{(\dagger)}{=} \mathbf{E}^\pi_{s_i}\left[\sum_{t=0}^{\tau_{s_i}-1}(R_t - g^\pi(S_t))\right] + \alpha_i \\
&= \sum_{s\in\mathcal{S}_i}\mathbf{E}^\pi_{s_i}\left[\sum_{t=0}^{\tau_{s_i}-1}\mathbf{1}(S_t = s)\right](r^\pi(s) - g^\pi(s)) + \alpha_i \\
&\overset{(\ddagger)}{=} E^\pi_{s_i}[\tau_{s_i}]\mu^\pi(-|\mathcal{S}_i)\cdot(r^\pi - g^\pi) + \alpha_i \overset{(\S)}{=} \alpha_i
\end{aligned}$$

where (†) follows from the observation that when starting from $\mathcal{S}_i$, the dynamics are trapped in $\mathcal{S}_i$ under π; (‡) follows by regeneration (Proposition I.1); and (§) follows from Proposition I.2. We conclude by setting $\alpha_i := -\mu^\pi(-|\mathcal{S}_i)\cdot h^\pi_0$ and $h^\pi := h^\pi_\alpha$. □

The formula (I.11) is important and shows that the bias of a policy is obtained by normalizing the rewards collected on pieces of the trajectory. An alternative definition, which is usually more standard, is $h^\pi(s) := \lim \mathbf{E}_s^\pi[\sum_{t=0}^{T-1}(R_t - g^\pi(S_t))]$. The limit is however not guaranteed to exist, and this issue is circumvented by taking the Cesàro limit instead. Drazin inverses and Laurent series are sometimes also used, but I prefer (I.11) to all these approaches, that I believe is true to the stochastic nature of Markov reward processes.

Regarding the example of Figure 1.2, the bias of the policy choosing (†) is $h^\dagger = (9, 0)$ while the bias of the one choosing (‡) is $h^\ddagger = (-1, 0)$. We observe that $h^\ddagger \leq h^\dagger$.

1.2 Classification of Markov decision processes

Obviously, some policies are better than others. A policy may have better gain than another, or two may have the same gain but one has better bias, or two may gave the same gain and different biases, yet none has better bias than the other. In any case, this raises the question of **optimal policies**. The difficulty of this question is a matter of the underlying structure of the model. Markov decision processes are therefore classified into several categories.

Definition I.8. *Markov decision processes are classified as follows:*

(0) ***Ergodic*** *models, where every policy is ergodic (recurrent and aperiodic);*
(1) ***Recurrent*** *models, where every policy is recurrent;*
(2) ***Unichain*** *models, where every policy is unichain;*
(3) ***Communicating*** *models, where fully supported randomized policies are recurrent, or equivalently, where every state is reachable from any other under the right policy;*
(4) ***Weakly communicating*** *models, where fully supported randomized policies are unichain, or equivalently, if the model is the union of a communicating model and a bunch of states that are transient under every policy;*
(5) ***Multi-chain*** *models: All Markov decision processes.*

(Note: A randomized policy is fully supported if $\pi(a|s) > 0$ for all $(s, a) \in \mathscr{Z}$.)

The inclusion of classes is pictured in Figure 1.3.

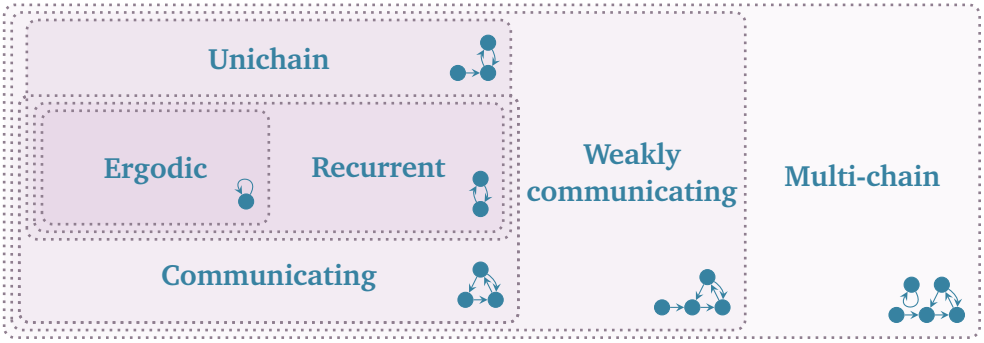


Figure 1.3: The classification of Markov decision processes.

In the reinforcement learning literature, the terminology *ergodic* Markov decision process is often wrongly used in place of *recurrent*, making it inconsistent with the theory of Markov chains. In Markov decision processes, the aperiodicity property is not very important and can always be artificially achieved via model transformations leaving the gain and the bias invariant.

Most of the manuscript focuses on the **communicating** classes with the exception of Part II that steps within the world of weakly communicating models.

1.3 The Bellman equation and optimal policies

We start with one of the most important result of this theory.

Theorem I.3 (First order Bellman theorem)**.** *Assume that M is communicating. There exists a policy $\pi \in \Pi$ with constant gain $g^\pi \in \mathbf{R}e$ solving the **Bellman equations**:*

$$\forall s \in \mathscr{S}, \quad g^\pi(s) + h^\pi(s) = \max_{a \in \mathscr{A}(s)} \{r(s,a) + p(s,a)h^\pi\}. \tag{I.12}$$

Proof. We use an argument belonging to the large family of **policy improvement** mechanisms. Pick π_0 arbitrarily then construct a sequence of policy (π_n) following the rules below:

(1) If $g^{\pi_n} \notin \mathbf{R}e$ then, using the communicativity of M, construct π_{n+1} as a deterministic policy converging to the component of π_n with maximal gain, from where π_{n+1} copies π_n.
(2) If $g^{\pi_n} \in \mathbf{R}e$, then consider $\Delta^{\pi_n}(s,a) := g^{\pi_n}(s) + h^{\pi_n}(s) - r(s,a) - p(s,a)h^{\pi_n}$. If there exists a pair $(s_n, a_n) \in \mathscr{Z}$ such that $\Delta^{\pi_n}(s_n, a_n) < 0$, then pick π_{n+1} as the copy of π_n changed with $\pi_{n+1}(s_n) = a_n$; Otherwise, set $\pi_{n+1} = \pi_n$.

We claim the following. $(*)$ If π_{n+1} is obtained via (1), then $g^{\pi_n} < g^{\pi_{n+1}}$ for the product order on $\mathbf{R}^{\mathscr{S}}$; and $(**)$ if $\pi_{n+1} \neq \pi_n$ is obtained via (2), then either $g^{\pi_n} < g^{\pi_{n+1}}$, or $g^{\pi_n} = g^{\pi_{n+1}}$ and $h^{\pi_n} < h^{\pi_{n+1}}$. In other words, the pair (g^{π_n}, h^{π_n}) is increasing for the lexicographic order unless $\pi_n = \pi_{n+1}$. Since Π is finite, it follows from $(*, **)$ that the sequence (π_n) is eventually stationary, converging to a $\pi \in \Pi$ satisfying $g^\pi \in \mathbf{R}e$ and (I.12), proving the result.

We are therefore left with proving our claims $(*)$ and $(**)$. $(*)$ is obvious and we focus on $(**)$. For all $s \in \mathscr{S}$, we have

$$\begin{aligned}\mathbf{E}_s^{\pi_{n+1}}\left[\sum_{t=0}^{T-1} R_t\right] &= \mathbf{E}_s^{\pi_{n+1}}\left[\sum_{t=0}^{T-1}\big(g^{\pi_n}(S_t) + \big(e_{S_t} - p(S_t,A_t)\big)h^{\pi_n} - \Delta^{\pi_n}(S_t,A_t)\big)\right] \\ &= Tg^{\pi_n}(s) - \mathbf{E}_s^{\pi_{n+1}}\left[\sum_{t=0}^{T-1} \mathbf{1}(S_t = s_n)\right]\Delta^{\pi_n}(s_n,a_n) + \mathbf{E}_s^{\pi_{n+1}}[h^{\pi_n}(s) - h^{\pi_n}(S_T)].\end{aligned} \tag{I.13}$$

We immediately see from (I.13) that $g^{\pi_n} \leq g^{\pi_{n+1}}$. If s_n is recurrent under π_{n+1} then the expected visit counts of s_n satisfy $\mathbf{E}_s^{\pi_{n+1}}[\sum_{t=0}^{T-1} \mathbf{1}(S_t = s_n)] \sim T\mu^{\pi_{n+1}}(s_n|s_n)$ hence grow linearly with T; We deduce that $g^{\pi_n}(s_n) < g^{\pi_{n+1}}(s_n)$. Otherwise, s_n is transient under π_{n+1} so $g^{\pi_n} = g^{\pi_{n+1}}$ and $h^{\pi_n}(s) = h^{\pi_{n+1}}(s)$ when s is recurrent under π_{n+1}. For all $s \in \mathscr{S}$, we have

$$\mathbf{E}_s^{\pi_{n+1}}\left[\sum_{t=0}^{T-1} R_t\right] = \mathbf{E}_s^{\pi_{n+1}}\left[\sum_{t=0}^{T-1} r^{\pi_{n+1}}(S_t)\right] \overset{(\dagger)}{=} Tg^{\pi_n}(s) + \mathbf{E}_s^{\pi_{n+1}}[h^{\pi_{n+1}}(s) - h^{\pi_{n+1}}(S_T)] \tag{I.14}$$

where $(\dagger)$ invokes the Poisson equation of π_{n+1} (Definition I.7). Since $h^{\pi_{n+1}}(s) = h^{\pi_n}(s)$ for s a recurrent state of π_{n+1}, we further have $\mathbf{E}_s^{\pi_{n+1}}[h^{\pi_{n+1}}(S_T)] = \mathbf{E}_s^{\pi_{n+1}}[h^{\pi_n}(S_T)] + o(1)$ when $T \to \infty$. Combined with (I.13) and (I.14), we obtain

$$h^{\pi_{n+1}}(s) = h^{\pi_n}(s) + \mathbf{E}_s^{\pi_{n+1}}\left[\sum_{t=0}^{T-1} \mathbf{1}(S_t = s_n)\right]\Delta^{\pi_n}(s_n,a_n) + o(1) \tag{I.15}$$

for all $s \in \mathscr{S}$. We conclude accordingly. □

The consequences of this result are striking. The Bellman equations guarantee the existence of optimal policies of the first order, that there exists deterministic optimal policies and motivate the introduction of the Bellman operator. However, the very strength of the result is rather that it gets rid of the necessity to take the history of play into account: Policies are not required to depend on time or on the history of play to achieve optimality. That is, the smartest learner cannot get higher asymptotic average reward than $\max_{\pi\in\Pi} g_s^\pi$, even if they make use of the whole history of play. This will be the basis of the regret benchmark in the next Chapter 2

Theorem I.3 is incomplete however, because the result is too coarse to provide a satisfying treatment of what happens at the second order, i.e., at the order of the bias h^π. This is the subject of the next paragraph.

1.3.1 Optimal policies, gain and bias and higher order Bellman equations

In this manuscript, we distinguish between three forms of optimalities: gain optimality, Bellman optimality and bias optimality. The assumption "M is communicating" is technical and can be removed up to fixing the definition of Bellman optimal policies.

Definition I.9. *Let $\pi \in \Pi^{\mathrm{SR}}$ and assume that M is communicating.*

(1) *π is **gain-optimal** ($\pi \in \Pi^*$) if $g^\pi(s) \geq \max_{\pi'\in\Pi^{\mathrm{SR}}} g^{\pi'}(s)$ for all $s \in \mathscr{S}$;*
(2) *π is **Bellman-optimal** ($\pi \in \Pi^*_{\mathrm{Bell}}$) if $g^\pi \in \mathbf{R}e$ and π satisfies (I.12);*
(3) *π is **bias-optimal** ($\pi \in \Pi^*_{\mathrm{bias}}$) if $\pi \in \Pi^*$ and $h^\pi(s) \geq \max_{\pi'\in\Pi^*} h^{\pi'}(s)$ for all $s \in \mathscr{S}$.*

Gain optimal policies are policies with maximal asymptotic average reward whatever the initial state. This is the first order optimality. At second order is bias optimality, consisting in gain optimal policies with maximal bias among gain optimal policies (from every state). Sitting right in-between gain and bias optimalities is the notion of **Bellman optimality**, which is not standard but was pointed out as an important refinement of gain optimality in the definition of Whittle and Gittins index in Markovian bandits, see Gast et al. (2023). Bellman optimality is closer to bias optimality than gain optimality however, as a Bellman optimality can be made bias optimal up to infinitesimal reward perturbation (see Boone and Gaujal (2023a)). The three are formally related as follows.

Proposition I.4. *For M communicating, we have $\Pi^*(M) \supseteq \Pi^*_{\mathrm{Bell}}(M) \supseteq \Pi^*_{\mathrm{bias}}(M)$. In general, the inclusions are strict.*

Proof. This proof echoes the proof of Theorem I.3. Let π a Bellman optimal policies. Then its gaps $\Delta^\pi(s,a) := g^\pi(s) + h^\pi(s) - r(s,a) - p(s,a)h^\pi$ are non-negative, so whatever the learner and initial state, we have

$$\mathbf{E}\left[\sum_{t=0}^{T-1} R_t\right] = \mathbf{E}\left[\sum_{t=0}^{T-1}\left(g^\pi(S_t) + \left(e_{S_t} - p(S_t,A_t)\right)h^\pi - \Delta^\pi(S_t,a_t)\right)\right] \leq T g^\pi(s) + \mathbf{E}[h^\pi(S_0) - h^\pi(S_T)] \tag{I.16}$$

where $s \in \mathscr{S}$ is arbitrary. Instantiating the learner to whatever policy and taking the limit in T, we deduce that g^π is maximal among $\pi \in \Pi^{\mathrm{SR}}$.

Now pick a bias optimal policy π. It has optimal gain by definition, so $g^\pi \in \mathbf{R}e$. Assume ad absurdum that π is not Bellman optimal. Then, one enters the case (2) of the improvement process of the proof Theorem I.3 that constructs a policy π' with better gain or bias than π, which is not possible by definition. A contradiction.

To see that the inclusion are strict in general, observe first that the example provided by Figure 1.2 provides an example of gain optimal policy (‡) which is not Bellman optimal. Figure 1.4 provides an example of Bellman optimal policy which is not bias optimal. □

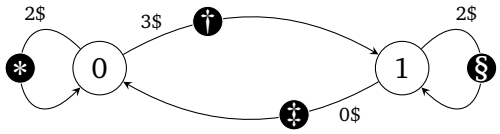


Figure 1.4: A Markov decision process (with deterministic transitions) where Bellman and bias optimalities differ. The policy $(*, ‡)$ is Bellman optimal with bias $(0, -2)$ while the policy $(†, §)$ is bias optimal with bias $(1, 0)$, hence $(*, ‡)$ is not bias optimal.

From Proposition I.4 follows that the Bellman equations (I.12) are incapable of capturing bias optimality completely, hence the optimal bias (Definition I.11) cannot be defined right from Theorem I.3 and will have to wait for the finer Theorem I.5. Ever since the paper of Blackwell (1962), it is known that Bellman equations can be pushed to higher orders and Theorem I.3 can be generalized. By seeing the bias of a policy h^π as a cost, we can define the next order bias as the bias of the Markov reward process $(-h^\pi, P^\pi)$.

Definition I.10. *Let $\pi \in \Pi^{\mathrm{SR}}$. The **0-th order bias** of π is the standard bias $h_0^\pi := h^\pi$. For $n \geq 1$, the **n-th order bias** of π is the bias (see Definition I.7) of the Markov reward process $(-h_{n-1}^\pi, P^\pi)$. In particular, the higher order biases are related by the Poisson equation:*

$$\forall s \in \mathcal{S}, \quad h_n^\pi(s) = -h_{n-1}^\pi(s) + p^\pi(s) \cdot h_n. \tag{I.17}$$

We can prove a stronger version of Theorem I.3.

Theorem I.5 (Second order Bellman theorem)**.** *Assume that M is communicating. There exists a policy $\pi \in \Pi$ with constant gain $g^\pi \in \mathbf{R}e$ satisfying the **nested** Bellman equations:*

$$\begin{aligned} \forall (s,a) \in \mathcal{Z}, \quad & g^\pi(s) + h^\pi(s) \geq r(s,a) + p(s,a)h^\pi \\ \forall (s,a) \in \mathcal{Z}_0^\pi, \quad & -h_1^\pi(s) \geq h^\pi(s) + p(s,a)h_1^\pi \end{aligned} \tag{I.18}$$

where $(s,a) \in \mathcal{Z}_0^\pi$ if $g^\pi(s) + h^\pi(s) = r(s,a) + p(s,a)h^\pi$.

Proof. The proof is a replicata of the one of Theorem I.3 with increased difficulty. Pick π_0 arbitrarily then construct a sequence of policy (π_n) following the rules below, where we denote the 0-th order gaps $\Delta^{\pi_n}(s,a) := g^{\pi_n}(s) + h^{\pi_n}(s) - r(s,a - p(s,a)h^{\pi_n}$ and the 1-th order gaps $\Delta_1^{\pi_n}(s,a) := h_1^{\pi_n}(s) + h^{\pi_n}(s) - p(s,a)h_1^{\pi_n}$.

(1) If $g^{\pi_n} \notin \mathbf{R}e$ then, using the communicativity of M, construct π_{n+1} as a deterministic policy converging to the component of π_n with maximal gain, from where π_{n+1} copies π_n.

(2) Else, if there exists a pair $(s_n, a_n) \in \mathcal{Z}$ such that $\Delta^{\pi_n}(s_n, a_n) < 0$, then pick π_{n+1} as the copy of π_n changed with $\pi_{n+1}(s_n) = a_n$;

(3) Else, if there exists a pair $(s_n, a_n) \in \mathcal{Z}_1^{\pi_n}$ with $\Delta_1^{\pi_n}(s_n, a_n) < 0$, then pick π_{n+1} as the copy of π_n changed with $\pi_{n+1}(s_n) = a_n$;

(4) Otherwise set $\pi_n = \pi_{n+1}$.

The argument is very similar. We show that unless $\pi_n = \pi_{n+1}$, the triplet $(g^{\pi_n}, h^{\pi_n}, h_1^{\pi_n})$ is increasing for the lexicographic order and conclude by finiteness of Π. Because (1) and (2) have already been analyzed in the proof of Theorem I.3, we focus on (3). Under (3), π^{n+1} is only

playing pairs z such that $\Delta^{\pi_n}(z) = 0$, so

$$\mathbf{E}_s^{\pi_{n+1}}\left[\sum_{t=0}^{T-1} R_t\right] = T g^{\pi_n}(s) + h^{\pi_n}(s) - \mathbf{E}_s^{\pi_{n+1}}[h^{\pi_n}(S_T)]$$
$$\overset{(\dagger)}{=} T g^{\pi_n}(s) + h^{\pi_n}(s) - \mathbf{E}_s^{\pi_{n+1}}[\mathbf{1}(S_T = s_n)]\Delta_1^{\pi_n}(s_n, a_n) + \mathbf{E}_s^{\pi_{n+1}}\left[h_1^{\pi_n}(S_T) - h_1^{\pi_n}(S_{T+1})\right]$$

where (†) follows by definition of $\Delta_1^{\pi_n}$ and by the observation that $\Delta_1^{\pi_n}(s, \pi_n(s)) = 0$ by the higher order Poisson equation (Definition I.10). Summing for $T = 1, \dots, T'$ provides

$$\mathbf{E}_s^{\pi_{n+1}}\left[\sum_{T=0}^{T'-1}\sum_{t=0}^{T-1} R_t\right] = \frac{T'(T'-1)}{2} g^{\pi_n}(s) + T' h^{\pi_n}(s) + \mathbf{E}_s^{\pi_{n+1}}\left[\sum_{T=0}^{T'-1} \mathbf{1}(S_T = s_n)\right]\Delta_1^{\pi_n}(s_n, a_n) + h_1^{\pi_n}(s) - \mathbf{E}_s^{\pi_{n+1}}\left[h_1^{\pi_n}(S_{T'})\right]. \tag{I.19}$$

We recognize $\mu^{\pi_{n+1}}(s_n|s)$ with $\frac{1}{T'}\mathbf{E}_s^{\pi_{n+1}}[\sum_{T=0}^{T'-1} \mathbf{1}(S_T = s_n)]$ (for T' large enough). Meanwhile, using the two first order Poisson equations, we have the analogous formula:

$$\mathbf{E}_s^{\pi_m}\left[\sum_{T=0}^{T'-1}\sum_{t=0}^{T-1} R_t\right] = \frac{T'(T'-1)}{2} g^{\pi_m}(s) + T' h^{\pi_m}(s) + h_1^{\pi_m}(s) + \mathrm{o}(1) \tag{I.20}$$

for $m \in \{n, n+1\}$, where the $\mathrm{o}(1)$ is $\mathbf{E}_s^{\pi_m}[h_1^m(S_{T'})]$. Remark that the quadratic term in (I.19) and (I.20) must be the same, hence $g^{\pi_n} = g^{\pi_{n+1}}$. Now, similarly to the analysis of (2) in the proof of Theorem I.3, we distinguish between $\mu^{\pi_{n+1}}(s_n|s) > 0$ or zero.

If $\mu^{\pi_{n+1}}(s_n|s) > 0$, then $\mathbf{E}_s^{\pi_{n+1}}[\sum_{T=0}^{T'-1} \mathbf{1}(S_T = s_n)]$ grows linearly with T' and in (I.19), the sublinear terms in T' can be ignored. By combining (I.19) and (I.20) for $m = n, n+1$, we conclude that $h^{\pi_n}(s) < h^{\pi_{n+1}}(s)$ for every s such that $\mu^{\pi_{n+1}}(s_n|s) > 0$.

If $\mu^{\pi_{n+1}}(s_n|s) = 0$, then $\mathbf{E}_s^{\pi_{n+1}}[\sum_{T=0}^{T'-1} \mathbf{1}(S_T = s_n)] = \mathrm{O}(1)$ so $h^{\pi_n}(s) = h^{\pi_{n+1}}(s)$ necessarily. If in addition s_n is reachable from s, then $\mathbf{E}_s^{\pi_{n+1}}[\sum_{T=0}^{T'-1} \mathbf{1}(S_T = s_n)] = \Theta(1)$ and $h_1^{\pi_n}(s) < h_1^{\pi_{n+1}}(s)$.

Overall, the triplet $(g^{\pi_n}, h^{\pi_n}, h_1^{\pi_n})$ is increasing for the lexicographic order, and we conclude by using that $|\Pi| < \infty$. □

Proposition I.6. *Any policy satisfying the nested Bellman equation* (I.18) *is bias optimal. In particular, bias optimal policies exist.*

The proof of the above is skipped because bias optimal policies won't appear much in this manuscript. The required material is already ready to be extracted from the proof of Theorem I.5. In the end, the optimal bias and gain are well defined. The result of Theorem I.5 can be extended at arbitrary order, leading to higher and higher orders of optimalities. At the very top of this hierarchy sits the **Blackwell optimality** that goes back to Blackwell (1962) that is optimal at every order, see Puterman (1994). However, beyond bias optimality, these optimality refinement are hard to interpret; Also, they are impossible to learn Boone and Gaujal (2023a), so we won't extend much on the subject. Gain and bias optimalities are enough for the scope of this manuscript.

> **Definition I.11.** *We define the **optimal gain** g^* and **optimal bias** h^* as the gain and bias function of any gain-optimal and bias-optimal policies respectively.*

1.3.2 The Bellman operator and Value Iteration

The description of optimal policies provided so far is qualitative. We have shown that optimal policies can be obtained by solving the Bellman equations, of first order to obtain gain optimal policies, and of up to second order to obtain bias optimal policies. Moreover, the proof

of Theorem I.3 provides a method that converges to Bellman optimal policies with a policy improvement scheme. This pseudo-algorithm is a variant of the famous **Policy Iteration** (PI), due to Howard (1960), and has been studied throughout although it is still the subject of many open questions, see Christ and Yannakakis (2023) and references therein for a recent overview of the literature. However, PI is not the numerical scheme that we will build upon in this manuscript. Instead, we mainly work with **Value Iteration** (VI), initially due to Bellman (1957) for which the literature is also immense, see Goyal and Grand-Clement (2023) and references therein for recent references. There is a third approach based on **Linear Programming** that can be traced back to d'Epenoux (1960), showing that gain-optimal policies can be obtained in polynomial time; An approach that we won't cover in this manuscript. Puterman (1994) covers the basics of all three.

1.3.2.1 The Bellman operator

The operator that is underlying Value Iteration is the **Bellman operator**.

Definition I.12. *The **Bellman operator** is the map $L : \mathbf{R}^{\mathscr{S}} \to \mathbf{R}^{\mathscr{S}}$ given by:*

$$\forall u \in \mathbf{R}^{\mathscr{S}}, \quad Lu(s) := \max_{a \in \mathscr{A}(s)} \{r(s,a) + p(s,a)u\} \tag{I.21}$$

*For a given $u \in \mathbf{R}^{\mathscr{S}}$, any randomized policy $\pi \in \Pi^{\mathrm{SR}}$ supported in actions achieving the above maximum is called a **greedy response** to u.*

Definition I.13. *The **span** semi-norm of $u \in \mathbf{R}^{\mathscr{S}}$ is $\mathrm{sp}(u) := \max(u) - \min(u)$.*

In light of Theorem I.3, first order Bellman equations (I.12) can simply be stated as the existence of $u \in \mathbf{R}^{\mathscr{S}}$ such that $Lu - u \in \mathbf{R}e$, or in other words, as the existence of a **span fixpoint** of L. Conversely, span fixpoints of the Bellman operators are solutions of the first order Bellman equation. The Bellman operator is computed in time $\mathrm{O}(|\mathscr{S}|^2)$ which is cheaper than computing the gain and the bias of a policy. It has many remarkable algebraic properties.

Proposition I.7. *Below, $\leq$ denotes the product order on $\mathbf{R}^{\mathscr{S}}$, u, v are generic vectors of $\mathbf{R}^{\mathscr{S}}$ and $\lambda \in \mathbf{R}$ is a generic scalar.*

(1) *L is **monotone**: $u \leq v \Rightarrow Lu \leq Lv$;*
(2) *L is **non span-expansive**: $\mathrm{sp}(Lu - Lv) \leq \mathrm{sp}(u - v)$;*
(3) *L is **linear**: $L(u + \lambda e) = Lu + \lambda e$.*

Proof. (1) and (3) are immediate. For (2), remark that $Lu = r^{\pi_u} + P^{\pi_u}u$ for some $\pi_u \in \Pi$. So,

$$Lu - Lv \leq (r^{\pi_u} + P^{\pi_u}u) - (r^{\pi_v} + P^{\pi_v}v) \leq P^{\pi_u}(u - v).$$

Symmetrically, $Lv - Lu \leq P^{\pi_v}(v - u)$. Hence, $\mathrm{sp}(Lv - Lu) \leq \mathrm{sp}((P^{\pi_u} - P^{\pi_v})(u - v)) \leq \mathrm{sp}(u - v)$. □

Proposition I.8. *Assume that $\mathrm{sp}(Lu - u) \leq \epsilon$ for some $\epsilon \geq 0$. Then every greedy response π to u has ϵ-optimal gain, i.e., $g^{\pi} \geq g^* - \epsilon e$.*

Proof. By assumption, there exists $g \in \mathbf{R}e$ and $w \in \mathbf{R}^{\mathscr{S}}$ satisfying $0 \leq w \leq \epsilon e$ such that $u + g \leq Lu \leq u + g + w$. By choice of π, we have $r^{\pi} + P^{\pi}u = Lu$, so $r^{\pi} \geq g + (I - P^{\pi})u$. By induction on $T \geq 0$, we derive:

$$g^{\pi}(s) \sim \frac{1}{T}\mathbf{E}_s^{\pi}\left[\sum_{t=0}^{T-1} R_t\right] = e_s \cdot \frac{1}{T}\sum_{t=0}^{T-1} r^{\pi} \geq e_s \cdot \left(g^{\pi} + \frac{1}{T}\left(I - (P^{\pi})^T\right)u\right) \sim g(s).$$

So $g^{\pi} \geq g$. Now, let π^* a bias-optimal policy. We have $r^{\pi^*} + P^{\pi^*}u \leq Lu \leq u + g + w$, so

$$g^*(s) \sim \frac{1}{T}\mathbf{E}_s^{\pi^*}\left[\sum_{t=0}^{T-1} R_t\right] = e_s \cdot \frac{1}{T}\sum_{t=0}^{T-1} r^{\pi^*} \leq e_s \cdot \left(g^{\pi^*} + \sum_{t=0}^{T-1}(P^{\pi^*})^t w + \frac{1}{T}\big(I - (P^{\pi^*})^T\big)u\right).$$

In the RHS, we have $\sum_{t=0}^{T-1}(P^{\pi^*})^t w \leq \sum_{t=0}^{T-1}(P^{\pi^*})^T \epsilon e \leq T\epsilon e$, hence $g^* \leq g + \epsilon e$. □

1.3.2.2 Value Iteration and Lazy Value Iteration

In summary, by Proposition I.8, greedy responses to near span fixpoints of the Bellman operator are nearly gain optimal. By the mean time, the Bellman operator is non-expansive in span by Proposition I.7. If it where span contracting, one would be guaranteed to converge to a span fixpoint by iterating it. This is the idea behind **Value Iteration** (Algorithm I.1): iterate L until a near span fixpoint is reached.

Algorithm I.1 Value Iteration (`VI`)

Parameters: A precision $\epsilon > 0$, an (optional) initialization u_0;

if u_0 is not initialized **then** $u_0 \leftarrow 0 \cdot e$;
for $n = 1, 2, \ldots,$ **do**
  $u_n \leftarrow Lu_{n-1}$;
  if $\mathrm{sp}(u_n - u_{n-1}) < \epsilon$ **then break;**
end for
return u_n.

Algorithm I.2 Lazy Value Iteration (`LVI`)

Parameters: A precision $\epsilon > 0$, an (optional) initialization u_0;

if u_0 is not initialized **then** $u_0 \leftarrow 0 \cdot e$;
for $n = 1, 2, \ldots,$ **do**
  $u_n \leftarrow \frac{1}{2}Lu_{n-1} + \frac{1}{2}u_{n-1}$;
  if $\mathrm{sp}(u_n - u_{n-1}) < \frac{1}{2}\epsilon$ **then break;**
end for
return u_n.

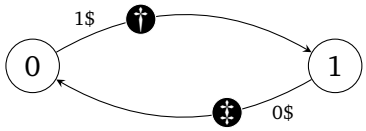


Figure 1.5: A single policy Markov decision process where Value Iteration (Algorithm I.1) is not converging. From $u_0 = (0, 0)$ we find that $u_n := L^n u_0 = (n, n-1)$ with $\mathrm{sp}(u_n) = 1$.

The stopping condition "$\mathrm{sp}(u_n - u_{n-1}) < \epsilon$" only makes sense if the Markov decision process is communicating, otherwise the Bellman operator is not even guarantee to have a span fixpoint. Even if a span fixpoint exists, Value Iteration is not guaranteed to converge in general and the main issue is the periodicity of optimal policies. An example is given with Figure 1.5. If all optimal policies are aperiodic however, then the algorithm converges.

Proposition I.9. *If all optimal policies are aperiodic and the model is communicating, then, for all* $u_0 \in \mathbf{R}^{\mathscr{S}}$,

$$w_\infty := \lim_{n\to\infty}(L^n u_0 - ng^* - h^*) \tag{I.22}$$

exists and Value Iteration (Algorithm I.1) converges in finite time.

Proof. The proof is a simplified and streamlined version of Puterman (1994)'s. Introduce the error term $w_n := u_n - ng^* - h^*$. Unfolding the definition of u_n and rearranging terms, we find:

$$w_n = \max_{\pi\in\Pi}(r^\pi - g^* + (P^\pi - I)h^* + P^\pi(u_{n-1} - (n-1)g^* - h^*)) = \max_{\pi\in\Pi}\big(-\Delta_\pi^* + P^\pi w_{n-1}\big) \tag{I.23}$$

where $\Delta_\pi^* := g^* - r^\pi + (I - P^\pi)h^*$ is the first order Bellman error of π against π^*. By Theorem I.3, $\Delta_\pi^* \geq 0$ so $w_n \leq \max_{\pi\in\Pi} P^\pi w_{n-1}$ and by induction, $w_n \leq \max(w_0)e$. Also, by picking π^* a bias

optimal policy, we have $\Delta^*_{\pi^*} = 0$ and $w_n \geq P^{\pi^*} w_{n-1}$ so by induction, $w_n \geq \min(w_0)e$. It follows that the sequence (w_n) is bounded. Let $w^+ := \limsup w_n$ and $w^- := \liminf w_n$.

We start by showing that $w^+(s) = w^-(s)$ if there exists a policy $\pi \in \Pi$ under which s is recurrent. Let $\mathscr{S}'$ the corresponding recurrent class. For all $s' \in \mathscr{S}$, we have $w_n(s') \geq (P^\pi w_{n-1})(s')$ so by induction, $w_m(s') \geq ((P^\pi)^{m-n} w_n)(s')$. By aperiodicity of π, it can be shown Levin and Peres (2017) that $\lim_{T\to\infty} e_{s'} \cdot (P^\pi)^T = \mu^\pi(-|\mathscr{S}')$ for all $s' \in \mathscr{S}'$. Therefore, for $?, ! \in \{+,-\}$, we have

$$\forall s' \in \mathscr{S}', \quad w^? \geq \mu^\pi(-|\mathscr{S}') \cdot w^! \tag{I.24}$$

so $\mu^\pi(-|\mathscr{S}')(w^+ - w^-) = 0$. Since $w^+ - w^- \geq 0$ and $\mu^\pi(s'|\mathscr{S}') > 0$ for $s' \in \mathscr{S}'$, we have $w^+ = w^-$ on the support of $\mu^\pi(-|\mathscr{S}')$.

We now extend that result to all states. For all $\epsilon > 0$ and $s \in \mathscr{S}$, there exists n arbitrarily large such that $w^+(s) - \epsilon \leq w_n(s)$. Moreover, $w_n(s) = \max_\pi(-\Delta^*_\pi + P^\pi w_{n-1})(s) \leq \max_\pi(-\Delta^*_\pi + P^\pi w^+)(s) + \epsilon$ if n is large enough. So $w^+ \leq \max_\pi(-\Delta^*_\pi + P^\pi w^+) + 2\epsilon e$. Similarly, we show $w^- \geq \max_\pi(-\Delta^*_\pi + P^\pi w^-) - 2\epsilon e$. Accordingly,

$$w^+ \leq \max_{\pi\in\Pi}\big(-\Delta^*_\pi + P^\pi w^+\big) \quad \text{and} \quad w^- \geq \max_{\pi\in\Pi}\big(-\Delta^*_\pi + P^\pi w^-\big). \tag{I.25}$$

Let $\pi \in \Pi$ achieving $\max_{\pi\in\Pi}(-\Delta^*_\pi + P^\pi w^+)$. By (I.25), $r^\pi \geq g^* + (I - P^\pi)(h^* + w^+)$ so π is gain optimal and have an aperiodic transition matrix P^π. Moreover, by (I.25) again, we have:

$$0 \leq w^+ - w^- \leq P^\pi(w^+ - w^-). \tag{I.26}$$

So by induction, $0 \leq w^+ - w^- \leq \frac{1}{n}\sum_{k=0}^{n-1}(P^\pi)^n(w^+ - w^-)$ for all $n \geq 1$, so by Definition I.6, $0 \leq w^+(s) - w^-(s) \leq \mu^\pi(-|s) \cdot (w^+ - w^-)$. Since π is gain optimal, we already know that $w^+ = w^-$ on the support of $\mu^\pi(-|s)$, hence $w^+(s) = w^-(s)$. □

The aperiodicity condition is actually never a problem, because Markov decision processes can be forced to be aperiodic without changing the optimal gain, optimal bias nor optimal policies (of arbitrary order), and this **aperiodicity transform** can be artificially simulated by a lazy version of Algorithm I.1, called **Lazy Value Iteration** (Algorithm I.2). The question of the convergence speed of these algorithms is more difficult and won't be covered. Empirically, the convergence of Lazy Value Iteration is very fast.

Proposition I.10. *Given a Markov decision process M and $\lambda \in (0,1)$, its **λ-lazy transform** is the model M_λ with the same state and action spaces, with rewards $r_\lambda(s,a) := \lambda r(s,a)$ and transitions $p_\lambda(s,a) := \lambda p(s,a) + (1-\lambda)e_s$.*

1. *All policies of M_λ are aperiodic;*
2. *For all $\pi \in \Pi^{\mathrm{SR}}$, we have $\lambda g^\pi(M) = g^\pi(M_\lambda)$ and $h^\pi(M) = h^\pi(M_\lambda)$;*
3. *$\Pi^*(M) = \Pi^*(M_\lambda), \Pi^*_{\mathrm{Bell}}(M) = \Pi^*_{\mathrm{Bell}}(M_\lambda)$ and $\Pi^*_{\mathrm{bias}}(M) = \Pi^*_{\mathrm{bias}}(M_\lambda)$.*

In particular, if M is communicating, then Lazy Value Iteration (Algorithm I.2) stops in finite time on every entry with u_n such that $\mathrm{sp}(Lu_n - u_n) < \epsilon$.

The proof can be found in Puterman (1994).

1.4 Comments

We only have scratched the surface of the theory of average reward Markov decision processes. We could talk more deeply about further optimality notions, the characterization of optimal policies or their computation. However, I estimate that the material introduced so far will be sufficient for what follows in this document.

In this manuscript, the objective function is $\mathbf{E}[R_1 + \ldots + R_T]$ when $T \to \infty$ and the problem is attacked head on. This is not how it is usually done, and many classical references of the literature (such as Arapostathis et al. (1993); Bertsekas (2012); Bertsekas and others (2011); Kallenberg (2016); Puterman (1994)) treat the average reward criterion as the limit of easier reward criteria, for instance the finite horizon criterion $\mathbf{E}[R_1 + \ldots + R_T]$ for $T < \infty$, or the discounted criterion $\mathbf{E}[\sum_{t=1}^{\infty} \gamma^{t-1} R_t]$ for $\gamma < 1$. These three reward criteria are perhaps the big three of objective functions in Markov decision processes, that are all the basis of different reinforcement learning problems. However, the average reward criterion is arguably the most general among these three, because the finite horizon and discounted criteria can be rewritten as an average reward criterion up to simple Markov decision process transforms, while the converse is not true in general. Beyond gain optimality exist optimalities of higher order that we have already mentioned. There is the bias optimality (Definition I.9), higher order bias optimalities, but also higher order discounted optimalities Puterman (1994) and Blackwell optimality Blackwell (1962), the latter being obtained as the limit of discounted optimalities when the discount coefficient γ goes to 1. Perhaps *because* they are finer than gain-optimality, the learnability properties of these higher order criteria are limited, see Boone and Gaujal (2023a).

Regarding the computation of optimal policies, Value Iteration (Algorithm I.1) is far from the only way to compute gain-optimal policies. Such policies can also be computed with Policy Iteration (PI) with a mechanism which is similar in spirit to the proof of Theorem I.3. Value Iteration and policy iteration are iterative methods and their time complexities are difficult to control. Optimal policies can also be obtained as the solutions of a linear program, by looking for the invariant measure μ (on $\mathscr{Z}$) maximizing the linear function $\sum_z \mu(z) r(z)$, hence the complexity of finding a gain-optimal policy is polynomial in the size of the Markov decision process.

I would also add that there is more to finite state-action spaces. In many simple problems, the state space is countable rather than finite, or the action space is compact rather than discrete. Time may not be discrete either. Moreover, the tools presented above are not designed to handle an explosion of the number of states or actions, and when the underlying model is too complex, the set of policies is usually parameterized by a lower dimensional parameter that is latter optimized, using gradient-descent-like algorithms Sutton and Barto (2018).

Chapter 2

Foundations of Reinforcement Learning in MDPs

In Chapter 1, we have introduced the basic concepts and results for Markov decision processes under the average reward criterion, as well as algorithmic solutions to compute optimal policies. Given a communicating model, Lazy Value Iteration (Algorithm I.2) can be used to compute optimal policies that can be deployed to navigate the Markov decision process optimally. This can only be done if the Markov decision process is known, because Lazy Value Iteration requires access to the reward function r and the transition kernel p. What if none of them are accessible? Then, no optimal policy may be computed beforehand as one has to interact with the environment to have any idea of its structure. Whenever the underlying model is unknown, one would expect a good agent to try actions at first and eventually play better and better actions. In any case, in face of an unknown environment, the choice of actions shall depend on past observations.

Definition I.14. *A **planner**[a] is any **random** sequence of randomized policies $\pi = (\pi_t)_{t\geq 0}$ such that π_t is $\sigma(O_t, \omega)$-measurable where $O_t := (S_0, A_0, R_0, \dots, S_t)$ is the history of play and ω is the inner randomness of the algorithm. Their set is denoted Π^{HR}.*

[a] sometimes called **history dependent policy**, or **agent**, or **controller**, or **learner** depending on communities and the authors' preference.

The question of the design of efficient planners that converge to optimal play has a long history. Making a complete survey goes way beyond the intended scope of this manuscript, although dusting and sorting the past literature would definitively help to trace all the various ideas that shape the theory of reinforcement learning in Markov decision processes. I nonetheless willingly took the time to go back the timeline to find the oldest possible work that can be considered as learning problems in Markov decision processes. Unsurprisingly, it goes back to multi-bandits. Multi-armed bandits (single state models) can be traced back to the works of Robbins (1952); Thompson (1933), and partially known Markov decision processes to Fox and Rolph (1973) at least, themselves extending methods of Mallows and Robbins (1964). In the 70's, the main focus was on planners achieving optimal gain, i.e., making sure that $\sum_{t=0}^{T-1} R_t$ approaches $Tg^*(s)$ when T goes to infinity. For quite a few decades however, achieving optimal gain is considered insufficient and has been replaced with the question of the convergence speed. How close can $\sum_{t=0}^{T-1} R_t$ be to Tg^*(s)? Up to this day and for probably many years still, the difference is measured by the **regret** of Robbins (1952), here instantiated in the style of Auer and Ortner (2006), that measures the asymptotic performance of an optimal policy $Tg^*(s)$ to the actual performance of the planner $\sum_{t=0}^{T-1} R_t$.

Definition I.15 (Auer and Ortner (2006))**.** *Assume that M is communicating. The* ***regret*** *under M is given by*

$$\mathrm{Reg}(T;M) = T g^*(s_0) - \sum_{t=0}^{T-1} R_t \tag{I.1}$$

where the dynamics are driven by a planner $\pi = (\pi_t)$ navigating on M from some initial state $s_0 \in \mathscr{S}$. Whenever the underlying model M is clear in the context, the dependence in M is dropped.

The expected regret grows sub-linearly on M if and only if the gain of the planner (π_t) is at least g^* on M. We say that the planner is **no-regret** on M.

Important remark. From now on, the dependency in M of the dynamics, of various quantities such as the gain, bias and gaps has to be made cleared, because M is hidden to the learner and is only known to live in a plausible space of models. We will write $\mathbf{E}^{\pi,M}[-], \mathbf{P}^{\pi,M}(-), g^\pi(s,M)$ etc. to account for this dependency. When M is not ambiguous however, this dependency may be dropped. This is for typographic convenience only.

2.1 Regret, gaps and classification of pairs

The structure of the underlying model can be exploited to rewrite the regret into a quantity that is less subjected to stochasticity. Indeed, even if the planner only picks optimal actions, the quantity $\sum_{t=0}^{T-1} R_t$ will deviate because the rewards gathered from a given state-action pair won't exactly match the expected value, and the stochastic transitions will drive the planner to states that may differ from what the planner expects. Said differently, the regret may punish or grant the planner for events over which they have very little to absolutely no control. This is cumbersome even from the viewpoint of the overseer, because all this noise makes the actual planner's playing quality more difficult to evaluate. To overcome these difficulties, we introduce the **first order regret** which is much less noisy than the regret yet shadows its expectation, see Proposition I.11. The definition of the first order regret is based on the notion of **first order Bellman gaps** (sometimes **gaps** or **disadvantage**) that we introduce now.

Definition I.16. *Assume that M is communicating. The first order* ***Bellman gap****, or* ***gap****, of a state-action pair $(s,a) \in \mathscr{Z}$ is given by:*

$$\Delta^*(s,a) := g^*(s) + h^*(s) - r(s,a) - p(s,a)h^*. \tag{I.2}$$

By Theorem I.3, $\Delta^* \geq 0$.

These gaps are analogous to Q-values in finite horizon and discounted Markov decision processes. It quantifies how close is a pair to optimality. The aggregate gaps account for the dominant part of the regret in expectation, see Proposition I.11 below.

Proposition I.11. *Assume that M is communicating. Whatever the planner (π_t), the expected regret satisfies:*

$$\mathbf{E}^{(\pi_t)}[\mathrm{Reg}(T)] = \mathbf{E}^{(\pi_t)}\left[\sum_{t=0}^{T-1} \Delta^*(Z_t)\right] + \mathbf{E}^{(\pi_t)}[h^*(S_0) - h^*(S_T)]. \tag{I.3}$$

The term $\mathrm{FOReg}(T) := \sum_{t=0}^{T-1} \Delta^*(Z_t)$ *of the RHS of* (I.3) *is called the* ***first order regret****.*

Proof. We write:

$$\mathbf{E}^{(\pi_t)}\left[T g^*(s) - \sum_{t=0}^{T-1} R_t \right] = \mathbf{E}^{(\pi_t)}\left[\sum_{t=0}^{T-1} (g^*(S_t) - r(Z_t)) \right] = \mathbf{E}^{(\pi_t)}\left[\sum_{t=0}^{T-1} \big((e_{S_t} - p(Z_t)) h^* + \Delta^*(Z_t) \big) \right]$$

and we conclude using that $\mathbf{E}^{(\pi_t)}[\sum_{t=0}^{T-1}(e_{S_t} - p(Z_t))h^*] = \mathbf{E}^{(\pi_t)}[h^*(S_0) - h^*(S_T)]$. □

In (I.3), the term $\mathbf{E}^{(\pi_t)}[h^*(S_T) - h^*(S_0)]$ is bounded while the regret usually is not, hence this is a second order term that will be utterly ignored. Because the gaps are non-negative, the first order regret is non-decreasing and count the number of times the algorithm has picked suboptimal pairs weighted by how much they are suboptimal. Remark that the first order regret increases when, and only when pairs with positive gaps are picked. As a consequence, even in full knowledge of the model and history, the planner cannot obtain better gain than g^*. The same can also be proved at higher order (for the bias) by using Theorem I.5, showing that there is no need to rely on the history to score optimally and that comparing the performance of a planner to the performance of the best deterministic policy is relevant: The planner cannot best the optimal policy and no-regret planners are those achieving optimal gain.

We conclude this paragraph by providing a classification of pairs in the communicating setting, motivated by this discussion.

Definition I.17. *The pairs of a communicating Markov decision process are classified as:*

(1) ***Suboptimal pairs*** $\mathscr{Z}^{-} := \{z \in \mathscr{Z} : \Delta^*(z) > 0\}$*;*
(2) ***Weakly-optimal pairs*** $\mathscr{Z}^{*} := \{z \in \mathscr{Z} : \Delta^*(z) = 0\}$*;*
(3) ***Optimal pairs*** $\mathscr{Z}^{**} := \{(s,a) \in \mathscr{Z} : \exists \pi \in \Pi^*, \pi(s) = a \text{ and } \mu^\pi(s|s) > 0\}$. .

Namely, suboptimal pairs are pairs with positive gap, necessarily transient under gain-optimal policies and it is by playing those that the first order regret increases; Weakly-optimal pairs are pairs with null gap, that do not make the first order regret increase hence are "cost-free to play"; Optimal pairs are pairs that are recurrent under at least one gain-optimal policy. When the model is communicating, optimal pairs are weakly optimal too, i.e., $\mathscr{Z}^{**} \subseteq \mathscr{Z}^{*}$. The novel part of this classification is the distinction between weakly optimal pairs and optimal pairs. This distinction is shown to be very important in Part III, that develops the idea that the very structure of gain optimal policies is provided by $\mathscr{Z}^{**}$ rather than $\mathscr{Z}^{*}$.

2.2 Statistical decision theory, consistency and robustness

We have a big problem:

> *Fix a policy* π *and consider the planner* $(\pi_t) = (\pi, \pi, \ldots)$*. For any model* M *in which* π *is optimal,* (π_t) *has null first order regret and bounded expected regret on* M*.*

So that's it. For every model, there exists a planner with no regret on that model. So I can stop the manuscript here and we can all go home.

Well, of course not.

The planner introduced above plays π whatsoever and disregards accumulated observations whatsoever. So, if they happen to play in another environment $M^\dagger$ where π is not gain-optimal,

then the regret of (π_t) will grow linearly. This is an **overfitting** problem. The quality of play of the planner does not generalize to other models than M. To quote (Lattimore and Szepesvári, 2020, §34), "*the fundamental challenge in learning problems is that the true environment is unknown and [planners] that are optimal in one environment are not usually optimal in another*." This raises a few interesting questions: What defines an acceptable planner? Are there optimal planners? How small can the regret of an acceptable planner be? Can it be achieved?

This problem is formalized by borrowing concepts from **statistical decision theory**. Consider a space of Markov decision processes $\mathscr{M}$ of state-action space $\mathscr{Z}$ so that planners Π^{HR} can universally run on all elements on $\mathscr{M}$. $\mathscr{M}$ will be called the **space of plausible environments**, and plays the role of an assumption; One assumes that the true model M belongs to $\mathscr{M}$. The considered **loss function** $\ell : \mathscr{M} \times \mathbf{N} \times \Pi^{\mathrm{HR}} \to \mathbf{R}_+$ is the expected first order regret:

$$\ell(M, T, (\pi_t)) := \mathbf{E}^{(\pi_t), M}\left[\sum_{t=0}^{T-1} \Delta^*(Z_t; M)\right]. \tag{I.4}$$

The capacity of a planner to *learn* optimal actions is categorized depending on the behavior of their associated loss on the class of plausible environments.

Definition I.18. *Fix a state-action space $\mathscr{Z}$ and let $\mathscr{M}$ a space of models with state-action space $\mathscr{Z}$ (we write $\mathscr{M} \in \mathfrak{M}(\mathscr{Z})$). A planner (π_t) is*

(1) ***admissible*** *if, for every other planner (π'_t), there exists $M \in \mathscr{M}, T \in \mathbf{N}$ such that $\ell(M, T, (\pi_t)) \leq \ell(M, T, (\pi'_t))$, and* ***dominated*** *otherwise;*
(2) ***consistent*** *if, for every $M \in \mathscr{M}$, $\ell(M, T, (\pi_t)) = \mathrm{o}(T)$ when $T \to \infty$;*
(3) ***robust*** *if $\sup\{\ell(M, T, (\pi_t)) : M \in \mathscr{M}\} = \mathrm{o}(T)$ when $T \to \infty$.*

The concept of (1) **admissibility**, that comes from statistical decision theory, is not very satisfying regarding learning for two reasons. First, it does not exclude the overfitting concern mentioned upstream and, second, it is too binary. It labels as *dominated* many planners that, despite being barely dominated by other planners, are worth investigating for exogenous reasons, e.g., computational advantages. From a learning perspective, (2) **consistency** and (3) **robustness** are more pertinent. Consistency states that, whatever the model among plausible environments, the regret eventually grows sublinearly, hence the planner eventually converges to optimal play. The convergence speed is not prevented from varying greatly from a model to another and consistent algorithms are subjected to unstable performance, motivating robustness. Robustness is stronger, stating that convergence to optimal play is uniform in the class of plausible environments. A high level representation of these three classes of planners is represented in Figure 2.1. Remark that all these classes depends on $\mathscr{M}$. An algorithm that is consistent (or robust) on $\mathscr{M}$ is not guaranteed to be consistent (or robust) on a super-class of $\mathscr{M}$. This is very natural: Some planners may exploit an a priori structure of the environment to learn faster (if the environment is a queuing system, or has deterministic rewards, or has deterministic transitions, etc.), and generalize badly to environments where this structure is absent. To some extent, while overfitting a planner to a specific model is questionable, overfitting a planner to a class of plausible models is actually recommended.[1]

Regarding Definition I.18, robustness is a stronger property than consistency. Consistency is more flexible, making the design of consistent algorithms easier than robust algorithms and it is pretty common that, for a given model, the asymptotic regret of a consistent planner is better than the ones of their robust kins. The analysis is also so different that they pretty much live in

[1] This follows the well-known learning principle: "*Always make sure to use the problem's structure that you are aware of.*"

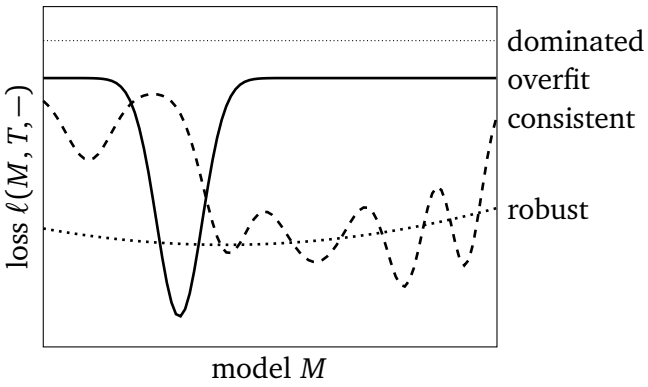


Figure 2.1: Artistic representation of overfitting, consistency and robustness. The x-axis is the underlying model and the y-axis is the associated loss of the planner for an arbitrary (fixed) time horizon T. Excepted the dominated planner, all planners are admissible.

different worlds. Consistency is linked to **model dependent settings**, where one takes a planner, fix a plausible model, then derives a regret bound that is specific for that model. Robustness is linked to **model independent settings**, where one takes a planner then derives a regret bound that holds simultaneously for all plausible models. For both family of settings, our work process is the following:

(1) Can we design a **lower bound** for this class of planners and plausible environments?
(2) How closely can this lower bound be approached (**tight upper bound**)?

Designing lower bounds is important. A lower bound provides insight in what makes the learning task difficult. It obviously depends on the considered class of plausible environments: The larger the class is, the larger is the lower bound. However, given an established lower bound, it is equally important to prove its tightness by providing a planner with a performance upper bound matching the lower bound. Otherwise, the possibility that the lower bound misses an important part of the learning task cannot be rejected. This being said, upper bounds are usually much harder to derive than lower bounds in reinforcement learning.

Important remark. In the sequel, we assume that $\mathcal{M}$ is a space of Markov decision spaces sharing the same (finite) state-action space $\mathcal{X}$, and assume that all elements of $\mathcal{M}$ have **Bernoulli rewards**. The second assumption could be changed to any single parameter exponential family of distributions, such as Gaussian distributions, Poisson distributions or finite-support distributions. Bernoulli rewards are rich enough to make the problem interesting, make a few (isolated) argument easier and spare a few additional notations.

2.3 The model independent setting, or minimax setting

I sometimes refer to the model independent setting as the **minimax setting**, because the goal is to find a planner (π_t) with **robust regret** guarantees, i.e., its regret on the worst possible instance $\sup_{M\in\mathcal{M}} \ell(M, T, (\pi_t))$ is as small as possible. This is about finding a planner approaching:

$$\inf_{(\pi_t)\in\Pi^{\mathrm{HR}}} \sup_{M\in\mathcal{M}} \mathbf{E}^{(\pi_t),M}[\mathrm{Reg}(T; M)] \tag{I.5}$$

hence the terminology "minimax". To approach (I.5), it must first be estimated. Especially, this is done using lower bounds, that tell how small the robust regret may be hoped to be.

Lower bounds of (I.5) can be traced back to Vogel (1960) at least, already providing a precise minimax analysis of two-armed bandits with Bernoulli rewards. As far as Markov decision processes are concerned, the first minimax lower bound is due to the seminal paper of Auer et al. (2009) for communicating MDPs, relating the minimax regret to the time horizon, the size of the state-action space and the **diameter** of the model, which is the expected time to travel from a state to another within the environment. That bound generalizes the minimax lower bound on multi-armed bandits.

Definition I.19 (Auer and Ortner (2006)). *The **diameter** of a Markov decision process is:*

$$D(M) := \max_{s \neq s'} \min_{\pi \in \Pi} \mathbf{E}_s^{\pi, M}[\tau_{s'}]. \tag{I.6}$$

The diameter is finite if, and only if M is communicating.

Theorem I.12 (Auer et al. (2009)). *Let $c > 0$ and let $\mathscr{M}_D(c)$ the class of all Bernoulli rewards models with state-action pairs $\mathscr{X}$ with diameter less than c. Then:*

$$\inf_{(\pi_t) \in \Pi^{\mathrm{HR}}} \sup_{M \in \mathscr{M}_D(c)} \mathbf{E}^{(\pi_t), M}[\mathrm{Reg}(T; M)] = \Omega\Big(\sqrt{(c \vee 1)|\mathscr{X}|T}\Big). \tag{I.7}$$

For tabular state-action space, this lower bound is more commonly written $\sqrt{DSAT}$, where S denotes the cardinal of the state space and $A := |\mathscr{A}(s)|$ is number of actions per state. What is important to remember from Theorem I.12 is that the dependency in the time horizon is $\sqrt{T}$, the dependency in the size of the model is $\sqrt{|\mathscr{X}|}$ and we have an extra dependency on the diameter, $\sqrt{D}$. This last dependency is related to the fact that the gain of a communicating model is D-Lipschitz and that the Lipschitz coefficient D is tight for *some* models. In the same paper Auer et al. (2009), the authors provide an algorithm (UCRL2) with minimax regret $DS\sqrt{AT\log(T)}$ and has been the basis for many works, including Bartlett and Tewari (2009); Bourel et al. (2020); Filippi et al. (2010); Fruit et al. (2020, 2018); Talebi and Maillard (2018); Tossou et al. (2019); Zhang and Ji (2019), improving the method of Auer et al. (2009) to obtain sharper regret upper bounds. However, not all works on robust planners originate from Auer et al. (2009). I provide in Table 2.1 a compendium of a few recent results.

In Table 2.1, we observe that bounds of many kind coexist. Some depend on the diameter, others on the span of the bias function, others on the mixing time and I have actually simplified a few exotic bounds that are relying on notions of distribution mismatch coefficient, local diameters, bias variances (etc.) to obtain finer regret bounds. Some algorithms use prior information, some do not; Some work for weakly communicating models, others only in the communicating setting, some need the underlying model to be ergodic. Observe that in the ergodic setting, the mixing time frequently appears in regret guarantees, while it is absent from the lower bound. The presence of this extra term is not mild because the mixing time cannot be bounded as a function of the diameter, meaning that the robust regret bound is not "compatible" with the lower bound; I will say that the bound is **not adapted** to the model class.

Another subsequent observation is that the dependency of all bounds with respect to the time horizon T is off by a polylogarithmic factor of T. The current tendency is to hide it under a $\widetilde{\mathrm{O}}(-)$ that I quite dislike, because by no means are these logarithmic factors mere artifacts of the analysis. Furthermore, the $\widetilde{\mathrm{O}}(-)$ further swallows *all* polylogarithmic factors, including $\log(D)$ and $\log(S)$ that we should keep track of. Going from $\sqrt{T\log^{\alpha}(T)}$ to $\sqrt{T}$ is not trivial and the analysis of algorithms has to be significatively reworked. While I am not aware of any attempt for Markov decision processes, the removal of the $\sqrt{\log(T)}$ is the subject of a few papers on

Algorithm	Robust regret	Class of models	Extra assumptions
REGAL, Bartlett and Tewari (2009)	$HS\sqrt{AT\log(T)}$	weakly communicating	intractable, knowledge of H
UCRL2, Auer et al. (2009)	$DS\sqrt{AT\log(T)}$	communicating	-
KL-UCRL, Filippi et al. (2010)	$DS\sqrt{AT\log(T)}$	communicating	-
KL-UCRL, Talebi and Maillard (2018)	$H\sqrt{SAT\log(T)} + S^2A^2 t_{\text{mix}}\log(T)$	ergodic	-
SCAL, Fruit et al. (2018)	$HS\sqrt{AT\log(T)}$	weakly communicating	knowledge of H
EBF, Zhang and Ji (2019)	$\sqrt{HSAT\log(T)}$	weakly communicating	intractable, knowledge of H
POLITEX, Abbasi-Yadkori et al. (2019)	$(t_{\text{mix}})^3 D(SA)^{\frac{1}{2}} T^{\frac{3}{4}}$	ergodic	-
UCRL2B, Fruit et al. (2020)	$S\sqrt{DAT\log^2(T)}$	communicating	-
UCRL3, Bourel et al. (2020)	$S\sqrt{DAT\log(T)}$	communicating	-
OPTIMISTIC-Q, Wei et al. (2020)	$H(SA)^{\frac{1}{3}} T^{\frac{2}{3}}$	ergodic	-
OSP, Ortner (2020)	$\sqrt{t_{\text{mix}}SAT\log^2(T)} + S^3A(\frac{t_{\text{mix}}}{\mu^2_{\min}})^2\log^2(T)$	ergodic	intractable
MDP-OOMD, Wei et al. (2020)	$\sqrt{(t_{\text{mix}})^3(D\vee S)AT\log(T)}$	ergodic	knowledge of t_{mix} and D
O-PSRL, Agrawal and Jia (2023)	$DS\sqrt{AT\log^6(T)}$	communicating	-
UCB-AVG, Zhang and Xie (2023)	$HS^5A^2\sqrt{T\log^2(T)}$	weakly communicating	-
PMEVI, Part II	$\sqrt{HSAT\log(T)}$	weakly communicating	-
Diameter lower bound	$\sqrt{DSAT}$	communicating	Auer et al. (2009)
Mixing lower bound	$\sqrt{t_{\text{mix}}SAT}$	ergodic	**see Part II**
Bias lower bound	$\sqrt{HSAT}$	weakly comm.	**see Part II**

Table 2.1: A compendium of algorithms with theoretical robust guarantees. We use the shorthand $H \equiv \text{sp}(h^*)$.

multi-armed bandits, see Audibert and Bubeck (2009); Garivier et al. (2022).

Among all the works listed in Table 2.1, only one method gets close to the lower bound: EBF of Zhang and Ji (2019); However EBF is completely intractable. The algorithm that I will present in Part II is, to some extent, an improved and tractable version of EBF.

Overall, Table 2.1 embodies the variety of robust regret guarantees and upon evoking mixing times, I have foreshadowed that the terms appearing in the upper bound do shape the class of models over which the robustness guarantees are valid. This leads to two definitions. The first notion that I introduce is **minimax complexity**, that quantifies the learning hardness of a class of Markov decision processes, see Definition I.20. The second definition, see Definition I.21 contains notions of **adapted learners** and **minimax optimality** and provide formal grounds to three ideas: (1) prior knowledge, (2) robust bounds correlated to the minimax complexity of a model class and (3) planners achieving minimax optimal regret up to multiplicative factors.

Definition I.20. *Let $\mathscr{X}$ a state-action space and let $\mathscr{M} \in \mathfrak{M}(\mathscr{X})$. The **minimax complexity** of $\mathscr{M}$ is:*

$$K(\mathscr{M}) := \liminf_{T\to\infty} \inf_{(\pi_t)\in\Pi^{\text{HR}}} \sup_{M\in\mathscr{M}} \frac{\mathbf{E}^{(\pi_t),M}[\text{Reg}(T;M)]}{\sqrt{T}} \tag{I.8}$$

Note that this definition anticipates on the $\sqrt{T}$ scaling of the regret.

Definition I.21. *Let $\mathscr{X}$ a state-action space. Let $\Theta \subseteq \mathbf{R}^d$ a parameter space and $\mathscr{M}(\Theta) \equiv (\mathscr{M}(\theta) : \theta \in \Theta)$ a parameterized family of model spaces with $\mathscr{M}(\theta) \in \mathfrak{M}(\mathscr{X})$. A planner $(\pi_t) \in \Pi^{\text{HR}}(\mathscr{X})$ is:*

(1) ***robust relatively to $\mathscr{M}(\Theta)$*** *if $\forall\theta \in \Theta, \sup_{M\in\mathscr{M}(\theta)} \mathbf{E}^{(\pi_t),M}[\text{Reg}(T;M)] = o(T)$;*

(2) **(α,β)-adapted to $\mathscr{M}(\Theta)$** *(where $\alpha \geq 0$ and $\beta \geq 0$) if*

$$\sup_{\theta\in\Theta} \limsup_{T\to\infty} \frac{\sup_{M\in\mathscr{M}(\theta)} \mathbf{E}^{(\pi_t),M}[\mathrm{Reg}(T;M)]}{K(\mathscr{M}(\theta))^{\alpha}\sqrt{T\log^{\beta}(T)}} < C < \infty; \tag{I.9}$$

(3) **β-minimax optimal on $\mathscr{M}(\Theta)$** *if it is $(1,\beta)$-adapted;*

Above, C is a generic constant that is ***independent of*** *$|\mathscr{X}|$.*

The content of Definition I.21 is new hence must be discussed. The definition considers classes of models parameterized by a parameter θ, that will typically track the maximum diameter, bias or mixing times, or even several at once. To understand this parametrization, let us focus on (1), the notion of **parametric robustness**. This strengthens the prior notion of robustness (Definition I.18) that is incapable of explaining the difference between UCRL2, a knowledge-free algorithm, and SCAL, that requires knowledge on $\mathrm{sp}(h^*)$ to have sub-linear regret guarantees. For instance, the regret upper bound of UCRL2 from the original work of Auer et al. (2009) can be written as: *Fix $\mathscr{S} := \{1,\dots,S\}$ and $\mathscr{A}(s) := \{1,\dots,A\}$. If $\mathscr{M}_D(c)$ denotes the class of all Bernoulli reward models with state-action space $\mathscr{X}$ and with diameter less than c, then the algorithm* UCRL2 *satisfies:*

$$\forall c > 0, \quad \sup_{M\in\mathscr{M}_D(c)} \mathbf{E}^{\mathtt{UCRL2},M}[\mathrm{Reg}(T;M)] = \mathrm{O}\Big(cS\sqrt{AT\log(T)}\Big). \tag{I.10}$$

Accordingly, the performance bound of UCRL2 is valid for several classes of models at once: It is valid for a family of classes parameterized by the maximal diameter. This property is characteristic of algorithms that do not require prior information. Meanwhile, the regret upper bound of SCAL from Fruit et al. (2018) can be written as: *Fix $\mathscr{S} := \{1,\dots,S\}$ and $\mathscr{A}(s) := \{1,\dots,A\}$. If $\mathscr{M}_{h^*}(c)$ denotes the class of all Bernoulli reward models with state-action space $\mathscr{X}$ and with $\mathrm{sp}(h^*) \leq c$, then, for all $c > 0$, the algorithm* SCAL(c) *satisfies:*

$$\sup_{M\in\mathscr{M}_{h^*}(c)} \mathbf{E}^{\mathtt{SCAL}(c),M}[\mathrm{Reg}(T;M)] = \mathrm{O}\Big(cS\sqrt{AT\log(T)}\Big). \tag{I.11}$$

It means that UCRL2 is robust relatively to $\mathscr{M}_D(\mathbf{R})$ while SCAL(c) is robust relatively to $\mathscr{M}_{h^*}(c)$ but not relatively to $\mathscr{M}_{h^*}(\mathbf{R})$. For instance, PMEVI is robust relatively to $\mathscr{M}_{h^*}(\mathbf{R})$.

The second notion, **(α,β)-adaptivity** builds on the idea that the robust regret guarantees grow with the minimax complexity of the class, hence that the parameterized family of model classes is sound to express the robustness properties of the planner, although the dependency may be sub-optimal ($\alpha > 1$). When $\alpha = 1$, the algorithm is said **β-minimax optimal** and the robust regret guarantees of the planner can only be improved up to numerical factors. For instance, the regret bound of KL-UCRL provided by Talebi and Maillard (2018) is not adapted to the class of ergodic models parameterized by their bias span, because there exist models with small bias span but arbitrarily large mixing time. The compendium of Table 2.1 is enriched in light of Definition I.20 and Definition I.21 in Table 2.2.

Remark that planners with regret $\mathrm{O}(T^{2/3})$ or $\mathrm{O}(T^{3/4})$, such as POLITEX or OPTIMISTIC-Q are not considered to be adapted. This is because if their dependency in the time horizon essentially means the their upper bound is of a different nature than the lower bound, hence the two are hard to compare. In Part II, I will further refine the lower bound of Theorem I.12 and describe the algorithm PMEVI: where it comes from and why it is minimax optimal.

Contributions of the manuscript. The main contributions of this manuscript regarding the minimax setting are summarized with Theorems I.13 and I.14. The first Theorem I.13 shows

Algorithm	Adapted parametrized class	Range of θ	α, β
REGAL(c)	wkly com., $H \leq \theta$	$\theta \leq c$	untractable, $\alpha = 2, \beta = 1$
UCRL2	communicating, $D \leq \theta$	$\theta < \infty$	$\alpha = 2, \beta = 1,$
KL-UCRL	communicating, $D \leq \theta$	$\theta < \infty$	$\alpha = 2, \beta = 1,$
KL-UCRL	ergodic, $H \leq \theta_1, t_{\text{mix}} \leq \theta_2$	$\theta_1, \theta_2 < \infty$	$\alpha = 2, \beta = 1$
SCAL(c)	wkly com., $H \leq \theta$	$\theta \leq c$	$\alpha = 2, \beta = 1$
EBF(c)	wkly com., $H \leq \theta$	$\theta \leq c$	untractable, $\boldsymbol{\alpha = 1}, \beta = 1$
POLITEX	-	-	-
UCRL2B	communicating, $D \leq \theta$	$\theta < \infty$	$\alpha = 2, \beta = 2$
UCRL3	communicating, $D \leq \theta$	$\theta < \infty$	$\alpha = 2, \beta = 1$
OPTIMISTIC-Q	-	-	-
OSP	ergodic, $t_{\text{mix}} \leq \theta_1, \frac{t_{\text{mix}}}{\mu_{\min}} \leq \theta_2$	$\theta_1, \theta_2 < \infty$	$\alpha = 1, \beta = 2$
MDP-OOMD(c_1, c_2)	ergodic, $t_{\text{mix}} \leq \theta_1, D \vee S \leq \theta_2$	$\theta_1 \leq c_1, \theta_2 \leq c_2$	$\alpha = 3, \beta = 1$
O-PSRL	communicating, $D \leq \theta$	$\theta < \infty$	$\alpha = 2, \beta = 6$
UCB-AVG	wkly com., $H \leq \theta$	$\theta < \infty$	$\alpha = 10, \beta = 2$
PMEVI	wkly com., $H \leq \theta$	$\theta < \infty$	$\boldsymbol{\alpha = 1}, \beta = 1$

Table 2.2: Extra information on the compendium of Table 2.1.

that the regret complexity of out the class of communicating models with bounded bias span (i.e., such that $\text{sp}(h^*) \leq c$) is scaling with the square root of the minimum between the bias span and the diameter. More specifically, if $\mathscr{M}(c,d)$ denotes the set of communicating Markov decision processes satisfying $\text{sp}(h^*) \leq c$ and $D \leq d$, we show that $K(\mathscr{M}(c,d))) = \Omega(\sqrt{c \wedge d})$. This result is more universal than the lower bound of Auer et al. (2009). In Auer et al. (2009), the authors show that the regret complexity of a class of models with bounded diameter (i.e., such that $D \leq d$) scales with the square root of the diameter. Their result can be written as $K(\mathscr{M}(\infty,d)) = \Omega(\sqrt{d})$. It is well known (and it will be discussed in Chapter 4) that the bias span is bounded by the diameter, hence $\mathscr{M}(\infty,d) \subseteq \mathscr{M}(d,d)$ hence the lower bound of Auer et al. (2009) is a result on the diagonal of the parametrized family of models spaces $(\mathscr{M}(c,d))_{c,d \geq 0}$. Theorem I.13 extends their lower bound outside of the diagonal. Theorem I.14 provides an 1-minimax algorithm for the same parametrized family of model spaces, proving that the regret complexity is indeed $\sqrt{c \wedge d}$ up to a $\sqrt{\log(T)}$ term, that I conjecture to be artificial. [2]

Theorem I.13 (Regret lower bound). *Let $\mathscr{Z} := \bigcup_{s \in \mathscr{S}} \{s\} \times \{1, \dots, A\}$ a tabular pair space with $S \in 3\mathbf{N}$ and $A \geq 3$. Denote $\mathscr{M}(c,d)$ the set of communicating Markov decision processes M with pair space $\mathscr{Z}$ satisfying $\text{sp}(h^*(M)) \leq c$ and $D(M) \leq d$. Then:*

$$K(\mathscr{M}(c,\infty)) = \Omega\big(\sqrt{cSA}\big) \quad \textit{when} \quad c \to \infty. \tag{I.12}$$

Theorem I.14 (Regret upper bound). *Let $\mathscr{Z} := \bigcup_{s \in \mathscr{S}} \{s\} \times \{1, \dots, A\}$ a tabular pair space. Denote $\mathscr{M}(c,d)$ the set of communicating Markov decision processes M with pair space $\mathscr{Z}$ satisfying $\text{sp}(h^*(M)) \leq c$ and $D(M) \leq d$. Then:*

$$\limsup_{T \to \infty} \sup_{M \in \mathscr{M}(c,\infty)} \frac{\mathbf{E}^{\text{PMEVI},M}[\text{Reg}(T)]}{\sqrt{T \log(T)}} = \text{O}\big(\sqrt{cSA}\big) \quad \textit{when} \quad c \to \infty. \tag{I.13}$$

[2] Theorem I.14 is actually stronger, because the bound which is provided is uniform on the worst value of the diameter.

2.4 The model dependent setting

In parallel of the model independent setting, robust minimax guarantees and minimax optimal algorithms, the other existing frequentist approach is model dependent. Given a model, what are the best performance that a learner may hope to achieve? In Section 2.2, we have discussed that this question is only interesting if the focus is restricted to **consistent** planners, otherwise nothing prevents the planner from over-specifying their performance to a given model at the price of suffering from catastrophic performance for another. Similarly to the minimax approach, the history of model dependent analysis of unknown Markov decision processes has a long history. Among a long line of works, the paper of Lai and Robbins (1985) stands out as the first model dependent lower bound, written just as in the style of what is done today. This paper, despite being specific to multi-armed bandits, settled a lower bound machinery that has been applied to various generalizations of multi-armed bandits, including our own in Part III, hence worth elaborating.[3]

What is arguably the most important baseline of all lower bounds is a strengthened version of consistency (Definition I.18).

Definition I.22 (Lai and Robbins (1985))**.** *Fix a state-action space $\mathcal{Z}$ and let $\mathcal{M} \in \mathfrak{M}(\mathcal{Z})$ a space of models. A planner (π_t) is said **strongly consistent** on $\mathcal{M}$ if, for all $M \in \mathcal{M}$ and all $\epsilon > 0$, $\mathbf{E}^{(\pi_t),M}[\mathrm{Reg}(T;M)] = o(T^\epsilon)$.*

For bandits with Bernoulli rewards, the result of Lai and Robbins (1985) can equivalently be formulated as such.

Theorem I.15 (Lai and Robbins (1985))**.** *Assume that $\mathcal{Z} := \{1\} \times \{1,\dots,A\}$, i.e., defines the structure of a multi-armed bandit and let $\mathcal{M}$ the space of all models with state-action spaces $\mathcal{Z}$ and Bernoulli rewards. For every strongly consistent planner (π_t) and $M \in \mathcal{M}$ with **interior rewards** (i.e., $0 < r(z) < 1$ for all $z \in \mathcal{Z}$), we have:*

$$\liminf_{T\to\infty} \frac{\mathbf{E}^{(\pi_t),M}[\mathrm{Reg}(T;M)]}{\log(T)} \geq \sum_{z\in\mathcal{Z}} \frac{\Delta^*(z;M)}{\mathrm{kl}(r(z), \max_{z^*\in\mathcal{Z}} r(z^*))}. \tag{I.14}$$

The bound is tight, in the sense that there exists a strongly consistent planner achieving it. Lai and Robbins (1985) provide one. Such planners are said **asymptotically optimal** and many optimal planners are known to this date; Many are simpler than Lai and Robbins (1985)'s solution. We can find the famous Thompson Sampling (`TS`) of Kaufmann et al. (2012); Thompson (1933), `KLUCB` of Garivier and Cappé (2011); Maillard et al. (2011), `MED` of Honda and Takemura (2010), `IMED` of Honda and Takemura (2015) or `RB-SDA` of Baudry et al. (2020, 2023) to list a few.

In the world of Markov decision processes, the works of Lai and Robbins (1985) have been generalized in several directions but interestingly, a big part of the literature seems to hit a significant barrier: recurrent models. Agrawal et al. (1988) generalizes the lower bound and the planner of Lai and Robbins (1985) to ergodic models, obtaining a tight lower bound. The approach of Agrawal et al. (1988) is policy-wise, in the sense that every policy is seen as an arm, each explored in turn. The number of policies growing exponentially in $|\mathcal{Z}|$, this approach suffers from obvious combinatorial drawbacks. This is addressed by Burnetas and Katehakis (1997), that decomposes the lower bound pair-wisely instead (see Theorem I.16), and provides

[3] This is slightly incorrect, because the lower bound machinery has been modernized since. Nowadays, lower-bounds are systematically obtained with information theoretic techniques of various kind.

a pair-wise indexed asymptotically optimal planner. Specifically, at time t, the planner associates an index $I_t(S_t, a)$ to every playable action, then plays the action with highest index. The indexed algorithm of Burnetas and Katehakis (1997) was reworked and mixed with more modern bandit methods a few decades later by Pirutinsky (2020), that apparently drove very little attention. Meanwhile, this indexed strategy was mixed to the works of Honda and Takemura (2015) and improved by Pesquerel and Maillard (2022), producing the algorithm `IMED-RL`. This algorithm is asympotically optimal for ergodic models just as Burnetas and Katehakis (1997) while displaying great empirical performance. Following the works of Burnetas and Katehakis (1997), the state-of-the-art stays frozen for about a decade. Close to simultaneously, Auer and Ortner (2006) and Tewari and Bartlett (2007) provide `UCRL` (a prototype ancestor to `UCRL2`, Auer et al. (2009)) and `OLP` respectively, both achieving $O(\log(T))$ regret on recurrent models, but both methods are sub-optimal.

Theorem I.16 (Burnetas and Katehakis (1997))**.** *Fix $\mathscr{Z}$ a state-action space and let $\mathscr{M}_{\text{ergodic}}$ be the space of all ergodic models with state-action space $\mathscr{Z}$. For $M \equiv (\mathscr{Z}, r, p) \in \mathscr{M}_{\text{ergodic}}$ and $z \in \mathscr{Z}$, introduce the informational coefficient $C(z; M)$ given by:*

$$\inf\Big\{\mathrm{kl}(r(z)||r^\dagger(z)) + \mathrm{KL}(p(z)||p^\dagger(z)) : r^\dagger(z) + p^\dagger(z)h^*(M) \geq g^*(z; M) + h^*(z; M)\Big\}.$$

*For every strongly consistent planner (π_t) and $M \equiv (\mathscr{Z}, r, p) \in \mathscr{M}_{\text{ergodic}}$ with **interior rewards** (i.e., $0 < r(z) < 1$ for all $z \in \mathscr{Z}$), we have:*

$$\liminf_{T\to\infty} \frac{\mathbf{E}^{(\pi_t),M}[\mathrm{Reg}(T; M)]}{\log(T)} \geq \sum_{z\in\mathscr{Z}} \frac{\Delta^*(z; M)}{C(z; M)}. \tag{I.15}$$

Regarding communicating models, there is the remarkable serie of works Agrawal (1990, 1991) that introduces a technique called "forcing" together with an algorithm with regret $O(f(T)\log(T))$ where $f(T)$ is an hyper-parameter of their method, that can be any function increasing to infinity. The first works to approach the communicating setting with $O(\log(T))$ regret bounds are the seminal papers of `UCRL2` Auer et al. (2009) and `REGAL` Bartlett and Tewari (2009), but none of these works provide model dependent lower bounds. The same holds for `KL-UCRL` Filippi et al. (2010). In fact, all three algorithms are sub-optimal by failing to achieve the lower bound that we provide in Part III. We have to wait for Ok et al. (2018) to find lower bounds again, generalizing Burnetas and Katehakis (1997) to broader settings but keeping the recurrence assumption. Finally, Tranos and Proutiere (2021) escapes the world of recurrent models by providing lower bounds for deterministic transition models,[4] and algorithms that are asymptotically optimal for a few of those models. Meanwhile, for model classes with fixed transitions, Saber et al. (2024) provides `IMED-KD` with a model dependent analysis and convincing empirical performance, but no lower bound is provided to the corresponding model classes.

This leaves us to where we are today.

The current state-of-the-art of model dependent analysis has three diverging branches, consisting in (1) **recurrent** models, with a lower bound and convincing asymptotically optimal planners; (2) **communicating** models, with a few algorithms shown to have regret of order $O(\log(T))$ with explicit constants, but no guarantees of asymptotic optimality; and (3) **deterministic transition** models, with a lower bound and partially optimal algorithms.

[4] We postpone the comparison of the works of Tranos and Proutiere (2021) and our main lower bound to Part III to keep the discussion streamlined.

Contributions of the manuscript. In Part III, we merge all three branches by providing the first model dependent lower bound for communicating models (Theorem I.17). The lower bound is shown to be tight in Chapter 10, by providing a planner (ECoE) whose asymptotic regret upper bound matches the lower bound (Theorem III.14). Our main results are summarized below.

Definition I.23. *Let $M \equiv (\mathcal{Z}, r, p)$ a Markov decision process. The set* $\mathrm{Inv}(M)$ *of* ***pair-wise invariant measures****, or more simply* ***invariant measures****, are elements $\mu \in \mathbf{R}_+^{\mathcal{Z}}$ such that $\sum_{a \in \mathcal{A}(s)} \mu(s,a) = \sum_{z \in \mathcal{Z}} \mu(z) p(s|z)$ for all $s \in \mathcal{S}$.*

Definition I.24. *For $\mathcal{M} \in \mathfrak{M}(\mathcal{Z})$ a model space and $M \in \mathcal{M}$, the* ***confusing models*** $\mathrm{Cnf}(M; \mathcal{M})$ *for M relatively to $\mathcal{M}$ is the set of $M^\dagger \in \mathcal{M}$ such that (1) $M = M^\dagger$ on $\mathcal{Z}^{**}(M)$ and (2) $g^*(M^\dagger) > g^*(M)$. When unambiguous,* $\mathrm{Cnf}(M; \mathcal{M})$ *is abbreviated* $\mathrm{Cnf}(M)$.

Theorem I.17. *Fix $\mathcal{Z}$ a state-action space and let $\mathcal{M} \in \mathfrak{M}(\mathcal{Z})$ any space of communicating models. For every strongly consistent planner (π_t) and $M \equiv (\mathcal{Z}, r, p) \in \mathcal{M}$ with* ***interior rewards*** *(i.e., $0 < r(z) < 1$ for all $z \in \mathcal{Z}$), we have:*

$$\liminf_{T \to \infty} \frac{\mathbf{E}^{(\pi_t), M}[\mathrm{Reg}(T; M)]}{\log(T)} \geq K(M; \mathcal{M}) \tag{I.16}$$

where $K(M; \mathcal{M})$ is the solution to the optimization problem:

$$\inf_{\mu \in \mathrm{Inv}(M)} \sum_{z \in \mathcal{Z}} \mu(z) \Delta^*(z) \quad s.t. \quad \inf_{M^\dagger \in \mathrm{Cnf}(M)} \sum_{z \in \mathcal{Z}} \mu(z) \mathrm{KL}_z(M || M^\dagger) \geq 1 \tag{I.17}$$

with $\mathrm{KL}_z(M||M^\dagger) := \mathrm{kl}(r(z)||r^\dagger(z)) + \mathrm{KL}(p(z)||p^\dagger(z))$.

This lower bound is the subject of Part III in which it is discussed in further details.

Chapter 3

Technical toolbox

In this technical chapter, we provide a non-exhaustive list of general tools that are standard in the reinforcement learning literature and that we will use throughout the manuscript.

3.1 Changes of measures

Changes of measures will be used to derive lower bounds. The formulation of the inequality we make use of can first be found in Kaufmann et al. (2016) and was later generalized to Markov decision processes by the works of Marjani et al. (2021); Marjani and Proutiere (2021). Below, we write $M(z) := r(z) \otimes p(z)$, so that $\mathrm{KL}(M(z)||M^\dagger(z)) = \mathrm{KL}(r(z)||r^\dagger(z)) + \mathrm{KL}(p(z)||p^\dagger(z))$. We write $M \ll M^\dagger$ if the absolute continuity property $r(z) \otimes p(z) \ll r^\dagger(z) \otimes p^\dagger(z)$ holds for all $z \in \mathscr{Z}$. Recall that $\mathrm{kl}(-,-)$ is the Kullback-Leibler divergence between Bernoulli distributions, that is, $\mathrm{kl}(p,q) = p\log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q})$.

Lemma I.18 (Marjani and Proutiere (2021)). *Let $\mathscr{M}$ a space of Markov decision processes. For all models $M \ll M^\dagger \in \mathscr{M}$, planner (π_t) and $\sigma(O_t)$-measurable function $f : \mathscr{O}_T \to [0,1]$, we have:*

$$\sum_{z \in \mathscr{Z}} \mathbf{E}^{(\pi_t),M}[N_T(z)]\mathrm{KL}(M(z)||M^\dagger(z)) \geq \mathrm{kl}\Big(\mathbf{E}^{(\pi_t),M}[f(O_T)], \mathbf{E}^{(\pi_t),M^(\pi_t)}[f(O_T)]\Big) \tag{I.1}$$

This inequality is used to establish lower bounds both in the model dependent and model independent settings.

3.2 Standard concentration inequalities

In this section, we list various concentration results that will be used repeatedly in the sequel. Most of them are classic, or already known. For a few of them, I couldn't find a proof in a existing reference hence I provide one.

Lemma I.19 (Azuma's inequality, Azuma (1967)). *Let $(U_t)_{t\geq 0}$ a martingale difference sequence such that $\mathrm{sp}(U_t) \leq c$ a.s., i.e., there exists $a_t \in \mathbf{R}$ such that $a_t \leq U_t \leq a_t + c$ a.s. Then, for all $\delta > 0$,*

$$\mathbf{P}\left(\sum_{t=0}^{T-1} U_t \geq c\sqrt{\tfrac{1}{2}T\log\left(\tfrac{1}{\delta}\right)}\right) \leq \delta.$$

Lemma I.20 (Freedman's inequality, Freedman (1975)). *Let $(U_t)_{t\geq 0}$ a martingale difference sequence such that $|U_t| \leq c$ a.s., and denote its conditional variance $V_t := \mathbf{E}[U_t^2|\mathcal{O}_{t-1}]$. Then, for all $\delta > 0$,*

$$\mathbf{P}\left(\exists n \geq 1, \sum_{k=0}^{n-1} U_k \geq a \text{ and } \sum_{k=0}^{n-1} V_k \leq b\right) \leq \exp\left(-\frac{a^2}{2(ac+b)}\right).$$

Lemma I.21 (Freedman's inequality, Additive version Zhang et al. (2020)). *Let $(U_t)_{t\geq 0}$ a martingale difference sequence such that $|U_t| \leq c$ a.s., and denote its conditional variance $V_t := \mathbf{E}[U_t^2|\mathcal{O}_{t-1}]$. Then, for all $\delta > 0$,*

$$\mathbf{P}\left(\exists T' \leq T : \sum_{t=0}^{T'-1} U_t \geq \sqrt{2\sum_{t=0}^{T'-1} V_t \log\left(\frac{T}{\delta}\right)} + 4c\log\left(\frac{T}{\delta}\right)\right) \leq \delta.$$

Lemma I.22 (Time-uniform Azuma, Bourel et al. (2020)). *Let (U_t) a martingale difference sequence such that, for all $\lambda \in \mathbf{R}$, $\mathbf{E}[\exp(\lambda U_t)|U_1, \dots, U_{t-1}] \leq \exp(\frac{\lambda^2\sigma^2}{2})$. Then:*

$$\forall \delta > 0, \quad \mathbf{P}\left(\exists n \geq 1, \quad \left(\sum_{k=1}^{n} U_k\right)^2 \geq n\sigma^2\left(1+\frac{1}{n}\right)\log\left(\frac{\sqrt{1+n}}{\delta}\right)\right) \leq \delta.$$

Lemma I.23 (Time-uniform Weissman). *Let q a distribution over $\{1, \dots, d\}$. Let (U_t) a sequence of i.i.d. random variables of distribution q. Then:*

$$\forall \delta > 0, \quad \mathbf{P}\left(\exists n \geq 1, \left\|\sum_{i=1}^{n}(e_{U_i} - q)\right\|_1^2 \geq nd\log\left(\frac{2\sqrt{1+n}}{\delta}\right)\right) \leq \delta.$$

Proof. Remark that $\left\|\sum_{k=1}^{n}(e_{U_k} - q)\right\|_1 = \max_{v\in\{-1,1\}^d} \sum_{k=1}^{n} \langle e_{U_k} - q, v\rangle$. Let $W_k^v := \langle e_{U_k} - q, v\rangle$. Remark that for each $v \in \{-1,1\}^d$, (W_k^v) is a family of i.i.d. random variables with $-\langle q, v\rangle \leq W_k^v \leq 1 - \langle q, v\rangle$, so $\mathbf{E}[\exp(\lambda W_k^v)] \leq \exp(\frac{\lambda^2}{8})$ by Hoeffding's Lemma. By Lemma I.22, we have:

$$\begin{aligned}
\mathbf{P}\left(\exists n \geq 1, \left\|\sum_{k=1}^{n}(e_{U_k} - q)\right\|_1 \geq \sqrt{nd\log\left(\frac{2\sqrt{1+n}}{\delta}\right)}\right) &= \mathbf{P}\left(\exists v \in \{-1,1\}^d, \exists n, \sum_{k=1}^{n} W_k^v \geq \sqrt{nd\log\left(\frac{2\sqrt{1+n}}{\delta}\right)}\right) \\
&\leq \sum_{v\in\{-1,1\}^d} \mathbf{P}\left(\exists n, \sum_{k=1}^{n} W_k^v \geq \sqrt{nd\log\left(\frac{2\sqrt{1+n}}{\delta}\right)}\right) \\
&\leq \sum_{v\in\{-1,1\}^d} \mathbf{P}\left(\exists n, \sum_{k=1}^{n} W_k^v \geq \sqrt{\frac{1}{2}n\left(1+\frac{1}{n}\right)\log\left(\frac{\sqrt{1+n}}{2^{-d}\delta}\right)}\right) \\
&\leq 2^d \cdot 2^d\delta = \delta.
\end{aligned}$$

This concludes the proof. □

Lemma I.24 (Time-uniform Empirical Bernstein). *Let $(U_k)_{k\geq 1}$ a martingale difference sequence such that $\mathrm{sp}(U_n) \leq c$ a.s., let $\hat{U}_n := \frac{1}{n}\sum_{k=1}^{n} U_k$ the empirical mean and $\hat{V}_n := \frac{1}{n}\sum_{k=1}^{n}(U_k - \hat{U}_n)^2$ the population variance. Then,*

$$\forall \delta > 0, \forall T > 0, \quad \mathbf{P}\left(\exists t \leq T, \sum_{i=1}^{t} U_i \geq \sqrt{2t\hat{V}_t\log\left(\frac{3T}{\delta}\right)} + 3c\log\left(\frac{3T}{\delta}\right)\right) \leq \delta.$$

Proof. This is obtained with a union bound on the values of $n \leq T$, then applying Lemma I.26. □

Lemma I.25 (Time-uniform Empirical Likelihoods, Jonsson et al. (2020)). *Let q a distribution on $\{1,\dots,d\}$. Let (U_t) a sequence of i.i.d. random variables of distribution q. Then:*

$$\forall \delta > 0, \quad \mathbf{P}\Big(\exists n \geq 1, n\mathrm{KL}(\hat{q}_n||q) > \log\big(\tfrac{1}{\delta}\big) + (d-1)\log\Big(e\big(1+\tfrac{n}{d-1}\big)\Big)\Big) \leq \delta.$$

Lemma I.26 (Empirical Bernstein inequality, Audibert et al. (2009)). *Let $(U_k)_{k\geq 1}$ a martingale difference sequence such that $\mathrm{sp}(U_n) \leq c$ a.s., let $\hat{U}_n := \frac{1}{n}\sum_{k=1}^n U_k$ the empirical mean and $\hat{V}_n := \frac{1}{n}\sum_{k=1}^n (U_k - \hat{U}_n)^2$ the population variance. Then,*

$$\forall \delta > 0, \forall n \geq 1, \quad \mathbf{P}\left(\sum_{k=1}^n U_k \geq \sqrt{2n\hat{V}_n \log\big(\tfrac{3}{\delta}\big)} + 3c\log\big(\tfrac{3}{\delta}\big)\right) \leq \delta.$$

Lemma I.27 (Bennett's inequality, Audibert et al. (2009)). *Let $(U_t)_{t\geq 0}$ a martingale difference sequence such that $|U_t| \leq c$ a.s., and denote its conditional variance $V_t := \mathbf{E}[U_t^2|\mathscr{O}_{t-1}]$. Then,*

$$\forall \delta > 0, \forall n \geq 1, \quad \mathbf{P}\left(\exists k \leq n, \sum_{i=1}^k U_i \geq \sqrt{2\sum_{i=1}^n V_i \log\big(\tfrac{1}{\delta}\big)} + \tfrac{1}{3}c\log\big(\tfrac{1}{\delta}\big)\right) \leq \delta.$$

Lemma I.28 (Lemma 3 of Zhang and Xie (2023)). *Let (U_t) be a sequence of random variables such that $0 \leq U_t \leq c$ a.s., and let $\mathscr{O}_t := \sigma(U_0, U_1, \dots, U_{t-1})$. Then:*

$$\forall \delta > 0, \quad \mathbf{P}\left(\exists T \geq 0, \sum_{t=0}^{T-1} U_t \geq 3\sum_{t=0}^{T-1} \mathbf{E}[U_t|\mathscr{O}_{t-1}] + c\log\big(\tfrac{1}{\delta}\big)\right) \leq \delta;$$

$$\forall \delta > 0, \quad \mathbf{P}\left(\exists T \geq 0, \sum_{t=0}^{T-1} \mathbf{E}[U_t|\mathscr{O}_{t-1}] \geq 3\sum_{t=0}^{T-1} U_t + c\log\big(\tfrac{1}{\delta}\big)\right) \leq \delta.$$

Lemma I.29 (Tails of Geometric Random Variables). *Let (X_i) a sequence of i.i.d. random variable of distribution $\mathrm{G}(p)$, and let $S_n := X_1 + \ldots + X_n$ their sum. Then, for all $c \geq 2$ and $t \geq 0$,*

$$\mathbf{P}(S_n \geq c(\mathbf{E}[S_n] + t)) \leq (1-p)^t \exp\Big(-\tfrac{(2c-3)n}{4}\Big).$$

Proof. This proof is standard and was found on `math.stackexchange.com`[1]. We rely on Chernoff's method as usual by using the Laplace transform $\mathbf{E}[e^{sX_i}]$. We have:

$$\mathbf{P}(S_n \geq c(\mathbf{E}[S_n] + t)) \leq e^{-sct} e^{-scn/p} \prod_{i=1}^n \mathbf{E}[e^{sX_i}].$$

We compute the Laplace transform of X_i: $\mathbf{E}[e^{sX_i}] = (1 - \frac{1-e^s}{p})^{-1}$. Setting $s = -\frac{1}{c}\log(1-p)$, we have $\exp(-sc) = 1-p$. In the above formula, we readily obtain:

$$\mathbf{P}(S_n \geq c(\mathbf{E}[S_n] + t)) \leq (1-p)^t \exp\Big(n\Big(\tfrac{a}{p} - \log\Big(1 - \tfrac{b}{p}\Big)\Big)\Big)$$

where $a := \log(1-p)$ and $b = 1-(1-p)^{1/c}$. We want that exponential to decrease quickly to 0 with n, i.e., we want $a/p - \log(1-b/p) < 0$. By Bernoulli's inequality, we have $b = 1-(1-p)^{1/c} \leq p/c \leq p/2$, hence $b/p \leq \frac{1}{2}$. Moreover, for $z \in (0, \frac{1}{2}]$, $\log(1-z) \geq -z - z^2$, hence

$$\tfrac{a}{p} - \log\Big(1 - \tfrac{b}{p}\Big) \leq \tfrac{a}{p}\tfrac{b}{p} - \log(\tfrac{1}{2}) \leq \tfrac{1}{2}\big(\tfrac{a}{b} + \tfrac{3}{2}\big).$$

Finally, since $a = \log(1-p)$ and $b \leq p/c$, it follows that $\frac{a}{b} \leq \frac{c\log(1-p)}{p} \leq -c$, so we get $\frac{a}{b} - \log(1-\frac{b}{p}) \leq \frac{1}{2}(-c + \frac{3}{2})$. This concludes the proof. □

[1] https://math.stackexchange.com/questions/110691/tail-bound-on-the-sum-of-independent-non-identical-geometric-random-variables

Part II

Minimax Optimal Regret in Average Reward MDPs

This part is dedicated to the minimax setting (see Section 2.3). We start by discussing the minimax lower bound of the regret in Chapter 4, and show that it scales with the span of the bias function rather than the diameter, addressing an open question of Fruit et al. (2018). The remaining chapters explain how this lower bound can be achieved with a tractable algorithm. Chapter 5 studies by how much the gain of a policy or of a Markov decision process is subject to change under modification of the reward function and transition kernel. It is a high level discussion on the technique that such an algorithm should rely on and can be considered as a friendly introduction to the heavier regret analysis of `PMEVI` in Chapter 7. Chapter 6 is an introduction to the famous design principle known as optimism-in-face-of-the-uncertainty, suited to the design of robust algorithms. We pinpoint a few holes in existing works and explain the challenges faced by optimistic algorithms regarding regret, motivating the main components of `PMEVI`, the solution provided in Chapter 7.

The main target of this part is to explain the algorithm `PMEVI` in Chapter 7 from my paper Boone and Zhang (2024), written in collaboration with Zihan Zhang, which is a general solution to the minimax regret problem. Its core component is an improved version of the Extended Value Iteration (`EVI`, Auer et al. (2009)) subroutine that we call Projected Mitigated `EVI`, that can replace `EVI` in every existing algorithms in the literature that relies on it, and make it minimax optimal.

This part is an extended version of a paper written in collaboration with Zihan Zhang.

> Boone, V. and Zhang, Z. (2024). Achieving Tractable Minimax Optimal Regret in Average Reward MDPs. _eprint: 2406.01234

Chapter 7 is directly adapted from my paper Boone and Zhang (2024). Chapters 4 to 6 are an extended introduction to the ideas and the literature behind the algorithm presented in Chapter 7. Chapter 4 is independent of the three other chapters. Chapter 5 presents a technique and may be skipped. However, reading Chapter 6 is important to understand Chapter 7 fully.

Chapter 4

Minimax lower bounds

The story begins with the number of samples required to determine the value of a function correctly. Say that the system is characterized by an unknown parameter $\theta \in \Theta \subseteq \mathbf{R}^d$ that can only be observed up to noise by collecting independent samples $X_1, X_2, \dots$ drawn from a fixed distribution $\nu \in \mathscr{P}(\mathbf{R})$ of first moment θ. In addition to that unknown parameter, we are given a family of functions $f_\ell : \Theta \to \mathbf{R}$ indexed by labels ℓ living in a label space $\mathscr{L}$. The very value of θ is actually of little interest, because the true objective is to find an **optimal label** ℓ^* achieving:

$$f_{\ell^*}(\theta) = \max_{\ell \in \mathscr{L}} f_\ell(\theta). \tag{II.1}$$

Let us just assume that such a label exists and is unique. If ν has a second-order moment, a natural approach is to estimate θ by $\frac{1}{n}(X_1 + \ldots + X_n)$ and maximize $f_\ell(\frac{1}{n}(X_1 + \ldots + X_n))$ for $\ell \in \mathscr{L}$. If, by any chance, the functions f_ℓ are L-Lipschitz, then by the Central Limit Theorem, the proxy $f_\ell(\frac{1}{n}(X_1 + \ldots + X_n))$ approximates $f_\ell(\theta)$ with an error of order $\sigma L/\sqrt{n}$ where $\sigma^2 := \mathbf{E}[(X_i - \theta)^2]$ is the variance on sampled values. A remarkable consequence of this qualitative result is that if one label ℓ achieves $f_{\ell^*}(\theta) \leq f_\ell(\theta) + 1/\sqrt{n}$, there are high chances for ℓ to be mistaken for the optimal label if we have collected less that $\sigma^2 L^2$ samples of ν. So, the required amount of samples increases with L (the sharpness of deviations) and σ^2 (the noise on observations).

This construction is reminiscent of minimax lower bounds. Minimax lower bounds are achieved at instances where the function to maximize is subject to **sharp deviations** and where there are suboptimal labels with near-optimal values, i.e., suboptimality and optimality are **hard to distinguish**. Regarding Markov decision processes, θ will be the true model $M \equiv (\mathscr{X}, p, r)$, the labels are deterministic policies Π, and the function to maximize is the gain g^π. Then, a small difficult problem can serve as a basic brick to construct artificially harder problems, simply by merging several independent copies of the same basic bricks into a single one. Solving the larger problem can only be solved by solving the inner sub-problems in parallel. Typically, such problems are constructed by putting K copies of the same problem next to each other, where in each, the optimal label is difficult to distinguish from the sub-optimal ones, but one of the K copies holds an optimal label that is slightly better than all of its copies'. Henceforth, this copying construction results in an instance with **many symmetries**, in which finding the optimal label is like looking for a needle in a haystack.

The three emphasized terms are the three ingredients to make a minimax lower bound: find a model with many symmetries, where sub-optimal policies are actually close to being optimal, and where the gain is very sensible to kernel and reward perturbations.

4.1 The variations of the gain function

The main difficulty of minimax lower bounds for Markov decision processes is to estimate how the gain of a policy is subject to vary. This is usually done in an *ad hoc* way, i.e., a very specific model is described and an explicit formula for the deviations of the gain is derived for this model specifically. In this manuscript, we provide a general result to more generically design lower bounds. Theorem II.1 below shows that the gain varies according to $\mathrm{sp}(h^\pi)$.

Theorem II.1. *Let $M \equiv (\mathscr{X}, p, r)$ and $\hat{M} \equiv (\mathscr{X}, \hat{p}, \hat{r})$ two Markov decision processes and fix $\pi \in \Pi^{\mathrm{SR}}$ a randomized policy. If $\mathrm{sp}(g^\pi(M)) = 0$, then*

$$\|g^\pi(\hat{M}) - g^\pi(M)\|_\infty \le \|\hat{r}^\pi - r^\pi\|_\infty + \tfrac{1}{2}\mathrm{sp}(h^\pi(M))\|\hat{p}^\pi - p^\pi\|_1. \tag{II.2}$$

Proof. Let $T \ge 1$ and $s \in \mathscr{S}$ an initial state. For short, let $\epsilon_r^\pi := \|\hat{r}^\pi - r^\pi\|_\infty$ and $\epsilon_p^\pi := \|\hat{p}^\pi - p^\pi\|_1$.

$$\begin{aligned}
(*) &:= \mathbf{E}_s^{\pi,\hat{M}}\left[\sum_{t=0}^{T-1} R_t\right] \\
&= \mathbf{E}_s^{\pi,\hat{M}}\left[\sum_{t=0}^{T-1} \hat{r}^\pi(S_t)\right] \\
&\le \mathbf{E}_s^{\pi,\hat{M}}\left[\sum_{t=0}^{T-1} r^\pi(S_t)\right] + T\epsilon_r^\pi \\
&\overset{(\dagger)}{=} \mathbf{E}_s^{\pi,\hat{M}}\left[\sum_{t=0}^{T-1} \Big(g^\pi(S_t) + \big(e_{S_t} - p(S_t,A_t)\big)h^\pi\Big)\right] + T\epsilon_r^\pi \\
&\overset{(\ddagger)}{\le} T g^\pi(s) + \mathbf{E}_s^{\pi,\hat{M}}\left[\sum_{t=0}^{T-1} \Big(\big(e_{S_{t+1}} - \hat{p}(S_t,A_t)\big)h^\pi + (\hat{p}(S_t,A_t) - p(S_t,A_t))h^\pi\Big)\right] + \mathrm{sp}(h^\pi) + T\epsilon_r^\pi \\
&\overset{(\S)}{=} T\Big(g^\pi(s) + \epsilon_r^\pi + \tfrac{1}{2}\mathrm{sp}(h^\pi)\epsilon_p^\pi\Big) + \mathrm{sp}(h^\pi)
\end{aligned}$$

where (†) invokes the Poisson equation $g^\pi(S_t) + h^\pi(S_t) = r^\pi(S_t) + p^\pi(S_t)h^\pi$, (‡) uses that $g^\pi(S_t) = g^\pi(s)$ for all $t \ge 0$ and (§) that, if $p, p' \in \mathscr{P}(\mathscr{S})$ and $u \in \mathbf{R}^{\mathscr{S}}$ then $|(p'-p)u| \le \frac{1}{2}\mathrm{sp}(u)\|p'-p\|_1$. Dividing by T and letting it go to infinity, we obtain the desired upper-bound. The lower bound is obtained similarly. □

The bound of (II.2) is **tight** in general, for instance when p^π is of full support, i.e., when $p^\pi(s'|s) > 0$ for all $s, s' \in \mathscr{S}$. For simplicity, assume that π is deterministic, let $s_{\min}$ and $s_{\max}$ two states respectively minimizing and maximizing $h^\pi(s)$. Let $\alpha \in \{-1, +1\}$. Given $\epsilon_r, \epsilon_p > 0$, set

$$\hat{r}^\pi(s) := r^\pi(s) + \alpha\epsilon_r \quad \text{and} \quad \hat{p}^\pi(-|s) := p^\pi(-|s) + \alpha\tfrac{1}{2}\epsilon_p\big(e_{s_{\max}} - e_{s_{\min}}\big). \tag{II.3}$$

Because p^π is of full support, $\hat{p}$ is indeed a probability distribution provided that ϵ_p is small enough. By following the computations of the proof Theorem II.1, we see that every inequality becomes an equality, leading to $\hat{g}^\pi(s) = g^\pi(s) + \alpha(\epsilon_r + \frac{1}{2}\mathrm{sp}(h^\pi)\epsilon_p)$.

Therefore, for all reward function r and all kernel p (up to uniform infinitesimal perturbation to make if fully supported), the bound of (II.2) is the best possible bound for the ℓ_1-norm. It can (and it will) be improved by using other distances or divergences. To design a tight minimax lower bound however, the bound of (II.2) will be enough.

4.2 Diameter, mixing time or bias span?

The historical minimax lower bound of Auer et al. (2009), $\sqrt{DSAT}$, depends on the diameter, while our gain deviation result (Theorem II.1) indicates that the gain varies with the span of the bias function, hence suggesting that the dependency may be off. It is known, since (Bartlett and Tewari, 2009, Theorem 4) at least, that a dependency in $\mathrm{sp}(h^*)$ is more accurate than a dependency in D because $\mathrm{sp}(h^*) \leq D$, see Proposition II.2 below. Moreover, by looking at how Auer et al. (2009) prove their lower bound, Fruit et al. (2018) makes the observation that the "hard instance" that they provide satisfies $\mathrm{sp}(h^*) = D$, still leaving the possibility to replace D by $\mathrm{sp}(h^*)$. After all, a few algorithms have regret bounds depending on $\mathrm{sp}(h^*)$ rather than D; but either these have a priori information on $\mathrm{sp}(h^*)$ or their regret depends on extra parameters, or have sub-optimal dependency in $\mathrm{sp}(h^*)$. Also, the diameter and the bias span are not the only contenders at minimax regret lower bounds and many works provide bounds based on the mixing time.

So, should it be the diameter, the bias span or the mixing time?

In Proposition II.2, we show that the bias function is the best possible reference we can hope for, because it is always smaller than the diameter and the mixing time.

We recall the definition of the mixing time below.

Definition II.1 (Levin and Peres (2017))**.** *The **mixing time** t_{mix} of an ergodic Markov chain with kernel P and invariant measure μ is*

$$t_{\mathrm{mix}}(\alpha) := \inf\left\{t \geq 0 : \forall s \in \mathscr{S}, \|e_s \cdot P^t - \mu\|_1 \leq \tfrac{1}{2}\alpha\right\} \tag{II.4}$$

for $\alpha = \frac{1}{4}$. The mixing time of a Markov decision process is the largest mixing time $t_{\mathrm{mix}}(\pi)$ over all deterministic policies $\pi \in \Pi$.

Remark that it is finite only when all policies are aperiodic. The following result is a combination of different sources Bartlett and Tewari (2009); Wang et al. (2022).

Proposition II.2 (Bartlett and Tewari (2009); Wang et al. (2022))**.** *The span of the optimal bias function is upper-bounded by the diameter and the mixing time. More specifically:*

1. *If $M \equiv (\mathscr{X}, p, r)$ is communicating, then $\mathrm{sp}(h^*(M)) \leq \mathrm{sp}(r)D(M)$;*
2. *If $M \equiv (\mathscr{X}, p, r)$ is ergodic, then $\mathrm{sp}(h^*(M)) \leq 2\|h^*(M)\|_\infty \leq \frac{2\mathrm{sp}(r)}{1-\alpha} t_{\mathrm{mix}}(\alpha; \pi^*)$ for all bias optimal policy π^*.*

Proof. The proof of (1) below is simplified from the original source Bartlett and Tewari (2009). Fix two states $s_1, s_2 \in \mathscr{S}$ and let π such that $\mathbf{E}^\pi_{s_1}[\tau_{s_2}] < \infty$. We have:

$$\begin{aligned} 0 \leq \mathbf{E}^\pi_{s_1}\left[\sum_{t=0}^{\tau_{s_2}-1} \Delta^*(Z_t)\right] &\overset{(\dagger)}{=} \mathbf{E}^\pi_{s_1}\left[\sum_{t=0}^{\tau_{s_2}-1} \left(g^*(S_t) - r(Z_t) + \left(e_{S_t} - p(Z_t)\right)h^*\right)\right] \\ &\overset{(\ddagger)}{\leq} \mathrm{sp}(r)\mathbf{E}^\pi_{s_1}[\tau_{s_2}] + h^*(s_1) - h^*(s_2) \end{aligned}$$

where (†) follows from the Bellman equation and (‡) from $\mathrm{sp}(g^* - r) = \mathrm{sp}(r)$. Accordingly, $h^*(s_2) - h^*(s_1) \leq \mathrm{sp}(r)\mathbf{E}^\pi_{s_1}[\tau_{s_2}]$. Conclude by picking the optimal policy to travel from s_1 to s_2.

For (2), refer to Wang et al. (2022). □

Wang et al. (2022) also points out that the bias span is strictly smaller than the diameter and the mixing time in general, by remaining bounded while the other two explode to infinity, see Figure 4.1.

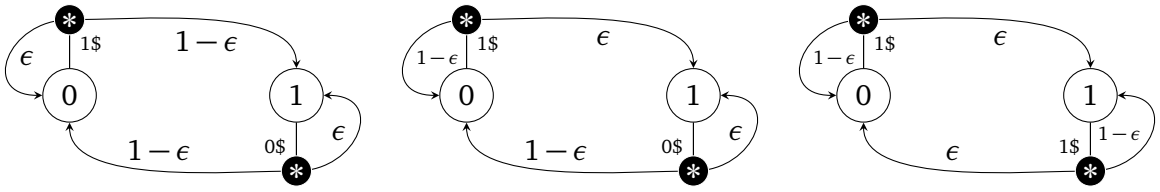


Figure 4.1: Markov reward processes with various relationships between $\mathrm{sp}(h^*)$, D and t_{mix}.

Following Proposition II.2, a bound depending on the bias span is far superior to a bound depending on the diameter or the mixing time. But then, isn't the $\sqrt{DSAT}$ lower bound of Auer et al. (2009) already showing that the lower bound is $\sqrt{\mathrm{sp}(h^*)SAT}$? **Yes**, it does. Yet, this answer is missing a big part of the story that can be summarized with the questions below.

> *Is the minimax lower bound $\sqrt{\mathrm{sp}(h^*)SAT}$ because it is $\sqrt{DSAT}$ for some instance where it happens that the diameter and the bias span are of the same order? Or, is the minimax lower bound $\sqrt{\mathrm{sp}(h^*)SAT}$ independently of the diameter? Is the diameter of a Markov decision process really what makes it hard to learn?*

This nuance is of utmost importance because the purpose of a lower bound is to describe what makes a problem difficult. Claiming that the lower bound depends on the diameter, *hence* that the lower bound depends on the bias, would mean what makes the learning problem difficult is the diameter, which is *by the way* lower bounded by the bias span. What we show downstream is the opposite. What truly makes the learning task difficult is the bias, and *by the way*, a few models happen to have the diameter of the same order than the bias span, hence the regret bounds for those depends on the diameter as well. A tangible example of such a model is the hard instance provided in Auer et al. (2009).

It is worth pointing out that in the parallel world of probably approximately correct reinforcement learning (PAC settings), where the goal is to collect observations to output an approximately optimal policy as fast as possible, the understanding of the dependency of the lower bounds with respect to the bias function has seen recent enlightening advances. In PAC learning, an algorithm is said **(ϵ, δ)-PAC** if it outputs an ϵ-gain optimal policy with probability at least $1-\delta$, and its **sample complexity** is the expected number of samples that the algorithm requires. It is known since Wang et al. (2022) that the minimax lower bound on the complexity is $D|\mathscr{Z}|\epsilon^{-2}\log(\frac{1}{\delta})$, and this sample complexity is achieved by Tuynman et al. (2024). An algorithm with sample complexity $t_{\mathrm{mix}}|\mathscr{Z}|\epsilon^{-2}\log(\frac{1}{\delta})$ is given by Wang et al. (2024). Wang et al. (2024) also provides the lower bound $\mathrm{sp}(h^*)|\mathscr{Z}|\epsilon^{-2}\log(\frac{1}{\delta})$ on the sample complexity and this lower bound is achieved by Zurek and Chen (2024) but there is a twist: $\mathrm{sp}(h^*)$ must be fed as input to the algorithm. Interestingly, Tuynman et al. (2024) shows that the lower bound $\mathrm{sp}(h^*)|\mathscr{Z}|\epsilon^{-2}\log(\frac{1}{\delta})$ cannot be achieved unless $\mathrm{sp}(h^*)$ is known by the planner.

Even more interestingly, in opposition to PAC learning, no knowledge on the span of the bias function is required to achieve minimax optimal regret.

These parallel works show that the optimal bias function and the diameter are objects of a different nature and that the simple relation $\mathrm{sp}(h^*) \leq \mathrm{sp}(r)D$ is, if useful, actually misleading. The bias function is more than a refinement of the diameter: It is a quantity that **cannot be estimated** under noise (see Tuynman et al. (2024)). And yet, the true achievable minimax lower bound depends on the bias function rather than the diameter.

4.3 The bias span minimax lower bound

The most precise, "non-asymptotic", minimax lower bound that provided in this manuscript is the following.

Theorem II.3. *Let $S \in 3\mathbf{N}$, $A \geq 3$ and let $\mathscr{X}$ be the induced tabular structure. Let $2 \leq c \leq d$ and let $T \geq \frac{400S(A-1)}{9c}(2d + c + \frac{1}{3}S - 1)^2 + \frac{4}{9}cS(A-1)$. For all every planner $(\pi_t) \in \Pi^{\mathrm{HR}}(\mathscr{X})$, there exists a model $M^\dagger \in \mathscr{M}(\mathscr{X})$ with $\mathrm{sp}(h^*(M^\dagger)) \leq c$ and $D(M^\dagger) \geq d$ such that:*

$$\mathbf{E}^{(\pi_t),M^\dagger}[\mathrm{Reg}(T)] \geq \tfrac{1}{144}\sqrt{3cS(A-1)T} - c. \tag{II.5}$$

In particular, if $\mathscr{M}(c,d)$ denotes the set of models such that $\mathrm{sp}(h^*) \leq c$ and $D \leq d$, we have $K(\mathscr{M}(c,\infty) \setminus \mathscr{M}(c,d)) = \Omega(\sqrt{c})$ for all $d > 0$.

This makes the minimax regret $\Omega(\sqrt{cSAT})$ independently of the diameter. If we refines the notation $\mathscr{M}(c,d)$ a little bit, Theorem II.3 shows that the minimax complexity (Definition I.20) of $\mathscr{M}([0,c],[d,+\infty])$, the space of tabular models with bias span $\mathrm{sp}(h^*) \in [0,c]$ and diameter $D \in [d,+\infty]$ is

$$K(\mathscr{M}([0,c],[d,+\infty])) = \Omega(\sqrt{cSA}) \tag{II.6}$$

In comparison, Auer et al. (2009) shows that $K(\mathscr{M}([0,d],[0,d])) = \Omega(\sqrt{dSA})$. Our bound (II.6) strictly generalizes their because $\mathscr{M}([0,d],[0,d]) \subseteq \mathscr{M}([0,d],[0,+\infty])$ and establishes the claims upstream: The minimax complexity scales with the bias regardless of the diameter.

This whole section is dedicated to a proof of Theorem II.3.

4.3.1 Construction of a hard instance

The critical model $M^\dagger$ of Theorem II.3 is found near a "hard instance" M. This "hard instance" is obtained as a model where all pairs are weakly optimal, so that the first order regret is null whatever the planner does. This model is full of symmetries and is designed so that the gain deviation bound (II.2) is tight. We consider an adequate state-action pair z that the learner visits more rarely than others in expectation, then consider a small modification of M denoted M^z_ϵ where z is slightly improved. Because M and M^z_ϵ are hard to distinguish, the planner will behave similarly on both model with significant probability, hence has significant chances to miss the optimal nature of z in M^z_ϵ, hence have high regret in M^z_ϵ. The proof consists in making all this precise.

We begin by describing the hard instance.

This hard instance is obtained by merging several "core" hard instances together. The core is a three state Markov reward process described in Figure 4.2 which diameter and bias span are of the right order of magnitude.

The perturbed core model introduced by Figure 4.3 is such that $g_\epsilon \sim \frac{1}{2} + \frac{1}{2}\epsilon c = \frac{1}{2} + \mathrm{sp}(h)\epsilon$ when $\epsilon \to 0$, making the deviation gain bound of Theorem II.1 tight as motivated. To construct the hard instance, the actions of the core instance of Figure 4.2 are first duplicated to form the **cluster instance**, by multiplying the action $(*)$ $A-1$ times from every state, resulting in a Markov decision process M' with 3 states and $A-1$ actions per state. From the cluster instance, the **hard instance** is obtained by duplicating the cluster into $\frac{1}{3}S$ individual copies $M'^{(0)}, \dots, M'^{(S/3-1)}$, later arranged in a torus using the reserved action $(\dagger)$, as depicted in Figure 4.3. Formally, we write $s^{(i)}$ the state $s \in \{0,1,2\}$ of the cluster $M'^{(i)}$. This last action $(\dagger)$ is a copy of $(*)$ that, instead of leaving the cluster $M'^{(i)}$ stable, goes to the next cluster $M'^{(i+1)}$. Visually,

$$p(-|s^{(i)},\dagger) := p(-|s^{(i+1)},*). \tag{II.7}$$

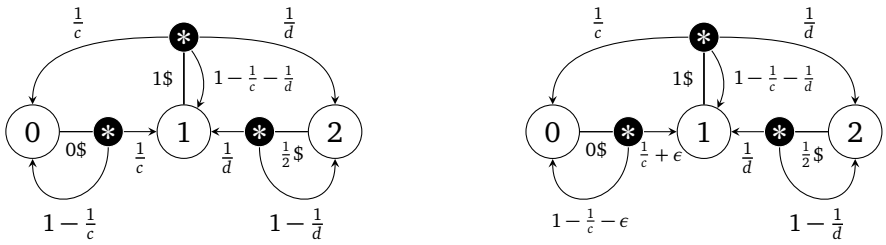


Figure 4.2: (To the left) The core of the hard instance parameterized by $0 < c \le d$. It is a three state Markov reward process with diameter $2d + c$ and bias span $\frac{1}{2}c$. Its gain is $g = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$, bias $h = (-\frac{1}{2}c, 0, 0) + \lambda e$ and invariant measure is $\mu = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. (To the right) The ϵ**-perturbed** core of the hard instance. Its gain is $g_\epsilon = (\frac{3}{2}\frac{1+\epsilon c}{3+2\epsilon c}, \frac{3}{2}\frac{1+\epsilon c}{3+2\epsilon c}, \frac{3}{2}\frac{1+\epsilon c}{3+2\epsilon c})$ and bias $h_\epsilon = (-\frac{3c}{6+4\epsilon c}, 0, -\frac{\epsilon c d}{6+4\epsilon c}) + \lambda e$.

On Figure 4.3, we see that playing (†) from state 1 in cluster i behaves similarly to playing $(*)$ from state 1 in cluster $i+1$. The resulting model has bias and diameter of the desired orders.

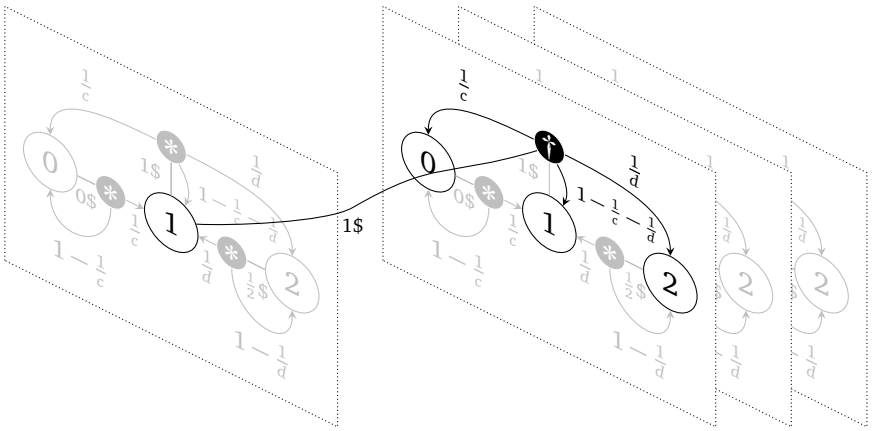


Figure 4.3: The hard instance is obtained by arranging multiple cluster instances (Figure 4.2) in a torus. Its per-state optimal gain is $g = \frac{1}{2}$. Whatever the cluster i, the optimal bias vector is $h_i = (-\frac{1}{2}c, 0, 0) + \lambda e$, where λ doesn't depend on i.

4.3.2 A few properties of the hard instance

The constructed hard instance has bias span and diameter of the required orders.

Lemma II.4. *Consider the hard instance M as given by Figure 4.3 with parameters $c, d > 0$. Then:*

(1) *Every policy is bias optimal on M;*
(2) $\mathrm{sp}(h^*(M)) = \frac{1}{2}c$;
(3) $2d + c \le D(M) \le 2d + c + \frac{1}{3}S - 1$.

These assertions follow immediately from the construction of M and the numerical properties of the core instance as described in Figure 4.2. The minimax lower bound is established by looking at the regret of planners on small perturbations of the hard instance M. Given a perturbation $\epsilon > 0$ and a pair $z \equiv (0^{(i)}, a) \in \mathscr{Z}$ with $a \neq (\dagger)$, the **(ϵ, z)-perturbed** model is the modification M_ϵ^z of M obtained by changing $p(0^{(i)}, a)$ to the ϵ-perturbed version of the core instance as described by Figure 4.2, i.e., $p_\epsilon^z(-|0^{(i)}, a) := p(-|0^{(i)}, a) + \epsilon(e_{1^{(i)}} - e_{0^{(i)}})$. The perturbed model has the following properties.

Lemma II.5. *Given a perturbation $\epsilon > 0$ and a pair $z \equiv (0^{(i)}, a) \in \mathcal{Z}$ with $a \neq (\dagger)$, the (ϵ, z)-perturbed hard model satisfies:*

(1) $\mathrm{sp}(h^*(M_\epsilon^z)) \leq \mathrm{sp}(h^*(M)) \leq 10(2d + c + \frac{1}{3}S - 1)c \cdot \epsilon$ *provided that* $\epsilon < (2d + c + \frac{1}{3}S - 1)^{-1}$;
(2) $D(M_\epsilon^z) \geq d$.

Proof. Fix π a bias optimal policy of M_ϵ^z. We see that π can be chosen unichain, playing z from $0^{(i)}$, playing any $(*)$ action from $1^{(i)}, 2^{(i)}$ and $(\dagger)$ from every other. In M, this policy is bias optimal, with bias $h^*(M)$. Invoking Lemma III.43, we have:

$$\|h^*(M_\epsilon^z) - h^*(M)\|_\infty \leq \big(2D^\pi(M_\epsilon^z)\mathrm{sp}(h^\pi(M)) + \tfrac{1}{2}\mathrm{sp}(h_1^\pi(M))\big)\|p_\epsilon^z - p\|_1 \tag{II.8}$$

where D^π is the **policy diameter** of π (see Definition III.8) and h_1^π is the first order bias (see Definition I.10). By definition, $\|p_\epsilon^z - p\|_1 = 2\epsilon$. Using Lemma III.43, we have $D^\pi(M_\epsilon^z) \leq 2D^\pi(M)$ provided that $2\epsilon < \frac{1}{D^\pi(M)}$. Using Lemma III.38, we have $\mathrm{sp}(h_1^\pi(M)) \leq 2\mathrm{sp}(h^\pi(M))D^\pi(M)$. Moreover, expanding the definition of $D^\pi(M)$, we find $D^\pi(M) \leq 2d + c + \frac{1}{3}S - 1$. All together, we get

$$\|h^*(M_\epsilon^z) - h^*(M)\|_\infty \leq 5(2d + c + \tfrac{1}{3}S - 1)c \cdot \epsilon \tag{II.9}$$

provided that $\epsilon < (2d + c + \frac{1}{3}S - 1)^{-1}$. This proves (1). Assertion (2) is immediate. □

In particular, the diameter of M_ϵ^z is of the right order and its bias span is $h^*(M) + \mathrm{O}(\epsilon) = \frac{1}{2}c + \mathrm{O}(\epsilon)$ when $\epsilon \to 0$, hence is of the desired order as well. For $\epsilon < \frac{1}{4} \cdot (10(2d + c\frac{1}{3}S - 1))^{-1}$, we obtain $\mathrm{sp}(h^*(M_\epsilon^z)) \leq \frac{3}{4}c$.

4.3.3 Proving the minimax lower bound: Proof of Theorem II.3

Consider an arbitrary planner (π_t) and consider the model M described in Figure 4.3. By construction, M is a concatenation of $\frac{1}{3}S$ cluster instances $M'^{(0)}, \dots, M'^{(S/3-1)}$ and the states of M can be split into three categories $\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2$ with

$$\mathcal{S}_i := \big\{i^{(j)} : j \in \{0, 1, \dots, \tfrac{1}{3}S\}\big\}. \tag{II.10}$$

That is, $\mathcal{S}_0$ is the collection of states 0 of the clusters, etc. Given $\mathcal{S}' \subseteq \mathcal{S}$, we denote $N_T(\mathcal{S}') := \sum_{s \in \mathcal{S}'} N_T(s)$.

The **idea of the proof** is to improve a pair that the planner visits poorly. By symmetry, there must be a cluster i and an action $a \in \mathcal{A} \setminus \{\dagger\}$ such that $\mathbf{E}^{(\pi_t), M}[N_T(0^{(i)}, a)] = \mathrm{O}(\frac{1}{SA})$, see (**STEP 1**) with (II.11). This poorly visited pair is denoted z and we consider all (ϵ, z)-perturbed models for some $\epsilon > 0$ to be tuned later on, where z is made better than every over pair in the model. So, in M_ϵ^z, a good planner must play z as often as possible. Indeed, with (**STEP 2**) and (II.13), the regret of the planner is shown to be directly linked to the number of times z is pulled. However, if ϵ is small enough then M_ϵ^z and M are pretty much indistinguishable, and the planner will behave very similarly under M and M_ϵ^z. This is formalized with a change of measure argument in (**STEP 3**) with (II.15), relating $\mathbf{E}^{(\pi_t), M_\epsilon^z}[N_T(z)]$ to $\mathbf{E}^{(\pi_t), M}[N_T(z)]$. By choice of z, we argue that for $\epsilon = \Theta(\sqrt{SA/(cT)})$, we have $\mathbf{E}^{(\pi_t), M}[N_T(z)] \leq \frac{2}{9}T$ that, combined with the regret lower bound of (**STEP 2**), is enough to conclude that $\mathbf{E}^{(\pi_t), M_\epsilon^z}[\mathrm{Reg}(T)] = \Omega(\sqrt{cSAT})$ for this choice of ϵ, see (II.18).

(**STEP 1**) *We have* $\mathbf{E}^{(\pi_t), M}[N_T(\mathcal{S}_0)] = \frac{1}{3}T + \mathrm{O}(1)$ *and there exists* $s \in \mathcal{S}_0$ *as well as* $a \in \mathcal{A}(s) \setminus \{\dagger\}$ *such that:*

$$\mathbf{E}^{(\pi_t), M}[N_T(s, a)] \leq \frac{T}{S(A-1)} + \frac{3D(M)}{S(A-1)}. \tag{II.11}$$

Proof. By construction, the invariant measure of every policy satisfies $\mu(\mathscr{S}_0) = \mu(\mathscr{S}_1) = \mu(\mathscr{S}_2)$. Consider the reward function $f(s,a) := \mathbf{1}(s \in \mathscr{S}_0)$ tracking the visits of $\mathscr{S}_0$. Let g^f, h^f, Δ^f the optimal gain, bias and gap functions of the Markov decision process obtained by changing the reward function to f; The associated optimal policy π^f is maximizing its number of visit counts on $\mathscr{S}_0$. We see that $g^f = \frac{1}{3}e$ and $\Delta^f \geq 0$.[1] Let $C := \mathrm{sp}(h^f)$, that satisfies $C \leq D(M)$ from Proposition II.2. We have:

$$\begin{aligned}\mathbf{E}^{(\pi_t),M}[N_T(\mathscr{S}_0)] &= \mathbf{E}^{(\pi_t),M}\left[\sum_{t=0}^{T-1} f(Z_t)\right] \\ &= \mathbf{E}^{(\pi_t),M}\left[\sum_{t=0}^{T-1}\big(g^f(S_t) + \big(e_{S_t} - p(Z_t)\big)h^f - \Delta^f(Z_t)\big)\right] \\ &\leq \tfrac{1}{3}T + C - \mathbf{E}\left[\sum_{t=0}^{T-1}\Delta^f(Z_t)\right] \leq \tfrac{1}{3}T + C. \end{aligned} \tag{II.12}$$

Similarly, by considering the Markov decision process obtained by changing the reward function to $-f$ and mimicking the above rationale, we show that $\mathbf{E}^{(\pi_t),M}[N_T(\mathscr{S}_0)] \geq \frac{1}{3}T - C'$ for some $C' < \infty$, that we can assume smaller than C up to increasing C. This proves the first half of the claim. For the second half, we continue from (II.12). Because $|\mathscr{S}_0| = \frac{1}{3}S$ and $\mathbf{E}^{(\pi_t),M}[N_T(\mathscr{S}_0)] = \sum_{s\in\mathscr{S}_0}\mathbf{E}^{(\pi_t),M}[N_T(s)]$, we deduce that the state $s \in \mathscr{S}_0$ which is the least visited in expectation is such that $\mathbf{E}^{(\pi_t),M}[N_T(s)] \leq \frac{T}{S} + \frac{3C}{S}$. From this state $s = 0^{(i)}$, there are $A-1$ copies of of action $(*)$ and since $\mathbf{E}^{(\pi_t),M}[N_T(s)] = \sum_{a\in\mathscr{A}(s)}\mathbf{E}^{(\pi_t),M}[N_T(s,a)]$, the least played $(*)$ action from s satisfies the claim. □

(**STEP 2**) *Denote z the pair (s,a) introduced by (**STEP 1**), see (II.11). Modify M to M_ϵ^z by changing $p(0^{(i)},a)$ to the ϵ-**perturbed** version as represented in Figure 4.2, i.e., $p_\epsilon^\dagger(-|0^{(i)},a) := p(-|0^{(i)},a) + \epsilon(e_{1^{(i)}} - e_{0^{(i)}})$. We have:*

$$\mathbf{E}^{(\pi_t),M_\epsilon^z}[\mathrm{Reg}(T)] \geq \frac{\epsilon c}{4\epsilon c + 3}\big(T - (3+2\epsilon c)\mathbf{E}^{(\pi_t),M_\epsilon^z}[N_T(z)]\big) - c. \tag{II.13}$$

Proof. This modification introduced in M_ϵ^z improves the pair $z \equiv (0^{(i)},a)$, making action a strictly better than any other from $0^{(i)}$. We see that $g^*(M_\epsilon^z) = \frac{3}{2}\frac{1+\epsilon c}{3+2\epsilon c}$. Let $\pi^\dagger$ the policy picking the action $\dagger$ from every state. Since M and M_ϵ^z are identical on $\mathscr{Z}\setminus\{z\}$, we find $g^{\pi^\dagger}(M_\epsilon^z) = g^{\pi^\dagger}(M)$ and $h^{\pi^\dagger}(M_\epsilon^z) = h^{\pi^\dagger}(M)$. Moreover, we know that $\pi^\dagger$ has null gaps on M, hence it has null gaps on M_ϵ^z excepted at z, where a straight forward computation shows that

$$\Delta^{\pi^\dagger}(z; M_\epsilon^z) = -\tfrac{1}{2}\epsilon c. \tag{II.14}$$

Following this, we find

$$\begin{aligned}(-) &:= \mathbf{E}^{(\pi_t),M_\epsilon^z}[\mathrm{Reg}(T)] \\ &:= \mathbf{E}^{(\pi_t),M_\epsilon^z}\left[T g^*(M_\epsilon^z) - \sum_{t=0}^{T-1} R_t\right] \\ &= \mathbf{E}^{(\pi_t),M_\epsilon^z}\left[T g^*(M_\epsilon^z) - \sum_{t=0}^{T-1}\Big(g^{\pi^\dagger}(S_t,M) + \big(e_{S_t} - p_\epsilon^z(Z_t)\big)h^{\pi^\dagger}(M) - \Delta^{\pi^\dagger}(Z_t; M_\epsilon^z)\Big)\right] \\ &\overset{(\S)}{\geq} \frac{T\cdot\epsilon c}{4\epsilon c + 6} - \mathbf{E}^{(\pi_t),M_\epsilon^z}[N_T(z)]\cdot\tfrac{1}{2}\epsilon c - \mathrm{sp}(h^{\pi^\dagger}(M))\end{aligned}$$

[1] Actually, we can show that $\Delta^f = 0$ by using the many symmetries of M.

where (§) unfolds the definition of both gain values and invokes (II.14) . Rearrange terms and conclude using that $\mathrm{sp}(h^{\pi^\dagger}(M)) = \mathrm{sp}(h^*(M)) = \frac{1}{2}c \le c$. □

(**STEP 3**) *For all $\epsilon > 0$ and $T \ge 0$, we have:*

$$\mathbf{E}^{M^z_\epsilon}[N_T(z)] \le \mathbf{E}^M[N_T(z)] + 2T \cdot \mathbf{E}^M[N_T(z)]\mathrm{kl}\Big(\tfrac{1}{c}, \tfrac{1}{c}+\epsilon\Big) + \mathbf{E}^M[N_T(z)]\sqrt{2T\mathrm{kl}\Big(\tfrac{1}{c}, \tfrac{1}{c}+\epsilon\Big)} \quad \text{(II.15)}$$

where we have removed the dependency on the planner (π_t) to lighten up notations.

Proof. By Lemma I.18, we have:

$$\begin{aligned}\mathrm{kl}\Big(\tfrac{1}{T}\mathbf{E}^{M^z_\epsilon}[N_T(z)], \tfrac{1}{T}\mathbf{E}^M[N_T(z)]\Big) &\le \sum_{(s,a)\in\mathscr{Z}} \mathbf{E}^M[N_T(s,a)]\ \mathrm{KL}(M(s,a)||M^z_\epsilon(s,a)) \\ &= \mathbf{E}^M[N_T(z)]\ \mathrm{KL}(p(z)||p^z_\epsilon(z)) \\ &= \mathbf{E}^M[N_T(z)]\ \mathrm{kl}\Big(\tfrac{1}{c}, \tfrac{1}{c}+\epsilon\Big).\end{aligned}$$

Using the numerical inequality $\mathrm{kl}(p, p+\epsilon) \ge \frac{\epsilon^2}{2(p+\epsilon)}$ that holds for $p \ge 0$ and $\epsilon \ge 0$, from the above follows that if $\mathbf{E}^{M^z_\epsilon}[N_T(z)] \ge \mathbf{E}^M[N_T(z)]$, then

$$\Big(\tfrac{1}{T}\mathbf{E}^{M^z_\epsilon}[N_T(z)] - \tfrac{1}{T}\mathbf{E}^M[N_T(z)]\Big)^2 \le \tfrac{2}{T}\mathbf{E}^{M^z_\epsilon}[N_T(z)] \cdot \mathbf{E}^M[N_T(z)]\mathrm{kl}\Big(\tfrac{1}{c}, \tfrac{1}{c}+\epsilon\Big).$$

This is an equation of the form $(x-\alpha)^2 \le \beta x$, that solves as $x \le \alpha + \beta + \sqrt{\alpha\beta}$. □

We have everything we need in order to conclude. In view of (II.13), the regret on M^z_ϵ is large if $T - (3+2\epsilon c)\mathbf{E}^{(\pi_t), M^z_\epsilon}[N_T(z)]$ is large. Given $\lambda > 0$, we look for a sufficient condition on ϵ (possibly asymptotic with respect to T) such that $\mathbf{E}^{(\pi_t), M^z_\epsilon}[N_T(z)] \le (1-\lambda)T$. Invoking (II.15), a sufficient condition is:

$$(3+2\epsilon c)\Big(\mathbf{E}^M[N_T(z)] + 2T \cdot \mathbf{E}^M[N_T(z)]\mathrm{kl}\Big(\tfrac{1}{c}, \tfrac{1}{c}+\epsilon\Big) + \mathbf{E}^M[N_T(z)]\sqrt{2T\mathrm{kl}\Big(\tfrac{1}{c}, \tfrac{1}{c}+\epsilon\Big)}\Big) < (1-\lambda)T.$$

The informational terms $\mathrm{kl}(\frac{1}{c}, \frac{1}{c}+\epsilon)$ are simplified to polynomial terms by using that $\forall \epsilon < \frac{1}{3}, \mathrm{kl}(\frac{1}{c}, \frac{1}{c}+\epsilon) \le 3c\epsilon^2$.[2] Invoking (II.11) to further upper bound $\mathbf{E}^M[N_T(z)] \le \frac{T}{S(A-1)} + \frac{3D}{S(A-1)}$ and changing $3+2\epsilon c$ to 4 (holding for $\epsilon < \frac{1}{2}c$, we obtain the sufficient condition

$$24T(T+3D)c \cdot \epsilon^2 + 4\sqrt{6}(T+3D)\sqrt{Tc} \cdot \epsilon < S(A-1)(1-\lambda)T \quad \text{(II.16)}$$

which is quadratic in $\epsilon \in (0, \frac{1}{3} \wedge \frac{1}{2}c)$. With a bit of algebra that is not very interesting, and using $S(A-1) \ge 6$, for $T \ge 3D$, it is natural to choose $\lambda = \frac{1}{3}$, and the condition (II.16) is simplified to the sufficient condition:

$$\epsilon < \frac{1}{6}\sqrt{\frac{S(A-1)}{cT}}. \quad \text{(II.17)}$$

Following (II.17), we set $\epsilon := \frac{1}{6}\sqrt{\frac{S(A-1)}{cT}}$, and we have $T - (3+2\epsilon c)\mathbf{E}^{(\pi_t), M^z_\epsilon}[N_T(z)] \ge \frac{1}{3}T$. Injecting in (II.13), we obtain:

$$\mathbf{E}^{(\pi_t), M^z_\epsilon}[\mathrm{Reg}(T)] \ge \frac{\epsilon c}{4\epsilon c + 3}\frac{1}{3}T - c \overset{(\S)}{\ge} \frac{1}{72}\sqrt{cS(A-1)T} - c \quad \text{(II.18)}$$

where (§) holds for $T > \frac{4}{9}cS(A-1)$. We finally change the dependency on c by a dependency on $\mathrm{sp}(h^*(M^z_\epsilon))$. By Lemma II.5, if $\epsilon \le \frac{1}{4} \cdot (10(2d + c + \frac{1}{3}S - 1))^{-1}$, then $\mathrm{sp}(h^*)(M^z_\epsilon) \le \frac{3}{4}c$. This condition on ϵ corresponds to $T \ge \frac{400S(A-1)}{9c}(2d + c + \frac{1}{3}S - 1)^2$, concluding the proof. ■

[2] It is shown using the integral formula $\mathrm{kl}(p, p+\epsilon) = \int_0^\epsilon \frac{x^2}{(p+x)(1-p-x)}dx$.

Chapter 5

Interlude: A story about the deviations of the gain

Is the minimax complexity truly $\sqrt{\mathrm{sp}(h^*)SAT}$? This manuscript is unfortunately short of an answer. What will be shown nonetheless, is that a regret of order $\sqrt{\mathrm{sp}(h^*)SAT\log(T)}$ can be achieved. Whether the logarithmic factor is mild or not remains an open question that has actually never been approached to begin with in the literature on Markov decision processes.

In this chapter, **we display a technique** and explain why this technique is mandatory.

This technique is usually referred to as the "**variance reduction method**" in the literature Azar et al. (2017); Kakade et al. (2020); Lattimore and Hutter (2012); Munos and Moore (1999), first introduced in the discounted reward setting where the objective function is $\mathbf{E}[\sum_{t=0}^{\infty}\gamma^t R_t]$ instead. It has become an inevitable tools since in broader settings, including our own. Its adaptation to undiscounted infinite reinforcement learning goes back to Fruit et al. (2020); Maillard et al. (2014); Talebi and Maillard (2018) and the manuscript of Fruit (2019) provides interesting insight on the technique. The nearly minimax optimal planner that we describe in Chapter 7, `PMEVI`, is inspired from `EBF` of Zhang and Ji (2019), that also relies on the variance reduction method to obtain nearly optimal regret bounds. In essence, the technique is used to control by how much an additive function of a Markov chain drifts away from its expectation, and is applicable to our reinforcement learning setting, but also in PAC learning for undiscounted infinite horizon problems Zhang and Xie (2023); Zurek and Chen (2024), finite horizon minimax regret problems Azar et al. (2017); Efroni et al. (2019); Li et al. (2020); Zanette and Brunskill (2019); Zhang et al. (2021, 2020), discounted PAC learning Li et al. (2021) and more.

5.1 Deviations of the gain of a fixed policy

If an algorithm has regret guarantees of order $\sqrt{\mathrm{sp}(h^*)SAT}$, the algorithm must gather enough information to estimate optimal policies on the fly. Although regret minimization do not require the algorithm to certify that the deployed policy is optimal, the algorithm cannot successfully deploy optimal policies without having some hidden driving mechanism that guarantees that these policies have high gain. In a dreamed world, we may assume that the algorithm has T samples for every state-action pair. Given this unreasonable amount of information, to which precision is the algorithm capable of recovering the gain of a policy?

For simplicity, assume that the reward function is known and that only the kernel are to be learned. If the algorithm has n samples for every state-action pair, then by denoting $\hat{p}_n$ the empirically observed kernel,[1] Weissman's inequality Weissman et al. (2003) (see Lemma I.23)

[1] i.e., $\hat{p}_n(s'|s,a)$ is proportional to the number of times the transition (s,a,s') has been observed.

states that $\|\hat{p}_n(z) - p(z)\|_1 \lesssim \sqrt{S/n}$. So, following the gain deviation inequality in ℓ_1-norm (Theorem II.1), we have:

$$\|g^\pi(\hat{M}_n) - g^\pi(M)\|_\infty \lesssim \sqrt{\frac{\mathrm{sp}(h^\pi)^2 S}{n}} \tag{II.1}$$

and the bound is optimal. There are at least two problems with this bound. First, the dependency in $\mathrm{sp}(h^\pi)$ is off. Second, there is an extra S. The take-home conclusion of (II.1) is that if one estimates the error of the gain with a ℓ_1-norm bound, it shouldn't be possible to obtain better regret than $\sqrt{\mathrm{sp}(h^*)^2 S^2 AT}$; This is the regret guarantees of `UCRL2` Auer et al. (2009) and `REGAL` Bartlett and Tewari (2009) that indeed relies on Weissman's inequality to estimate the error on the estimated kernel.

From this example, we see that we can trace back the gain deviation inequality (of the type of (II.1)) used by a method by looking at its regret bound, with the following principle:

> *If a method achieves a regret of order $\sqrt{f(M)SAT}$, then it must indirectly prove an inequality of the form $\|g^*(\hat{M}_n) - g^*(M)\|_\infty \lesssim \sqrt{f(M)/n}$.*

This also works in the other direction, and an inequality of the form $\|g^*(\hat{M}_n) - g^*(M)\|_\infty \lesssim \sqrt{f(M)/n}$ will definitely help to show a regret bound of order $\sqrt{f(M)SAT}$. To start with, we focus on the estimation of the gain of a single policy, or, equivalently, that the Markov decision process is a Markov reward process. In this section, we will first focus on how to get optimal dependency on $\mathrm{sp}(h^*)$. The dependency in S is another matter which is related to the construction of confidence regions rather than to the deviation of the gain function.

> **Important remark.** In the remaining of the paragraph, we consider Markov reward processes rather than Markov decision processes. The reward vector r is fixed and we focus on the sensibility of the gain function g on the kernel p. Therefore, we write $g(p)$ rather than $g(r,p)$.

5.1.1 The Azuma-Hoeffding bound

Denote p the kernel of the policy and assume that $\hat{p}_n$ is the empirical estimate of p obtained by collecting n independent samples of every $p(s)$, $s \in \mathscr{S}$. The gain under $p, \hat{p}_n$ for the same reward vector r are denoted $g(p)$ and $g(\hat{p}_n)$ respectively. To relate $g(\hat{p}_n)$ to $g(p)$, the inequality that powers methods like `UCRL2` Auer et al. (2009), `REGAL` Bartlett and Tewari (2009) or `KLUCRL` Filippi et al. (2010) in the background is **Azuma-Hoeffding's inequality** Azuma (1967) (Lemma I.19), stating that, for all vector $u \in \mathbf{R}^{\mathscr{S}}$, we have:

$$\forall s \in \mathscr{S}, \forall n \geq 1, \forall x > 0, \quad \mathbf{P}((\hat{p}_n(s) - p(s))u > x) \leq \exp\left(-\frac{2nx^2}{\mathrm{sp}(u)^2}\right). \tag{II.2}$$

In other words, Azuma-Hoeffding's inequality states that $(\hat{p_n}(s) - p(s))u$ is $\frac{1}{n}\mathrm{sp}(u)^2$-sub-Gaussian (see Boucheron et al. (2013)). Written in terms of error range, we find that $(\hat{p}_n(s) - p(s))u \leq \mathrm{sp}(u)\sqrt{\frac{1}{2n}\log(\frac{1}{\delta})}$ with probability at least $1-\delta$. Motivated by the Azuma-Hoeffding's style inequality (II.2), we have the following proposition.

> **Proposition II.6.** *Consider two policy kernels $\hat{p}$ (empirical estimate) and p (true kernel).*

Assuming that $(\hat{p}(s)-p(s))h$ satisfies the single-sided Azuma-Hoeffding's style inequality:

$$(\hat{p}(s)-p(s))h \leq \sqrt{\frac{\mathrm{sp}(h)^2\ell}{2n}} \tag{II.3}$$

where $n, \ell > 0$ are constants and $h \equiv h(p)$. If $g(p) \in \mathbf{R}e$ then the gain under dynamics $\hat{p}$ is upper-bounding as:

$$g(\hat{p}) \leq g(p) + \sqrt{\frac{\mathrm{sp}(h)^2\ell}{2n}} \tag{II.4}$$

In particular, if $n \geq \frac{\mathrm{sp}(h)^2\ell}{2\epsilon^2}$, then $g(\hat{p}) \leq g(p)+\epsilon$.

Proof. The expectations under $p, \hat{p}$ are respectively denoted $\mathbf{E}[-]$ and $\hat{\mathbf{E}}[-]$. We expand the regard with $h(s) = r(s) - g(s) + p(s)h$, where g and h are the gain and the bias vectors under the dynamics p. We have:

$$\begin{aligned}
\hat{\mathbf{E}}\left[\sum_{t=0}^{T-1} r(S_t)\right] &= \hat{\mathbf{E}}\left[\sum_{t=0}^{T-1}(g + h(S_t) - p(S_t)h)\right] \\
&= Tg + \hat{\mathbf{E}}\left[\sum_{t=0}^{T-1}(e_{S_{t+1}} - p(S_t))h\right] \\
&= Tg + \hat{\mathbf{E}}[h(S_0) - h(S_T)] + \hat{\mathbf{E}}\left[\sum_{t=0}^{T-1}(\hat{p}(S_t) - p(S_t))h\right] \\
&\overset{(*)}{\leq} Tg + \hat{\mathbf{E}}[h(S_0) - h(S_T)] + T\sqrt{\frac{\mathrm{sp}(h)^2\ell}{2n}}
\end{aligned}$$

where $(*)$ is obtained by (II.3). Divide by T and let it go to infinity. □

We immediately see that the dependency in $\mathrm{sp}(h^*)$ is sub-optimal. This is because the Azuma-Hoeffding bound of $(\hat{p}_n - p)u$ looses information on higher moments of $(\hat{p}_n - p)u$ that help to carry information from a state to another, that appear crucial in this setting. We need something that is more refined. This is where the variance reduction method kicks in.

5.1.2 The variance reduction method and the Berstein bound

The insufficient (II.2) is replaced by a variance-aware inequality. Freedman's inequality Freedman (1975) (Lemma I.20) states that, for all vector $u \in \mathbf{R}^{\mathscr{S}}$, we have

$$\forall s \in \mathscr{S}, \forall n \geq 1, \forall x > 0, \quad \mathbf{P}((\hat{p}_n(s) - p(s))u > x) \leq \exp\left(-\frac{x^2}{2n(x\,\mathrm{sp}(u) + \mathbf{V}(p,u))}\right). \tag{II.5}$$

This means that the tails of $(\hat{p}_n(s) - p(s))u$ are sub-exponential for large x and $\frac{2}{n}\mathbf{V}(p,u)$-sub-Gaussian for x in neighborhood of 0. This bound can equivalently be written as $(\hat{p}_n - p)u \leq \sqrt{\frac{2}{n}\mathbf{V}(p,u)\log(\frac{1}{\delta})} + \frac{4}{n}\mathrm{sp}(u)\log(\frac{1}{\delta})$ with probability at least $1-\delta$, and is more commonly referred to as **Bernstein's inequality** Bernstein (1924). When n is large in front of $\log(\frac{1}{\delta})$, this bound is much sharper than Azuma-Hoeffding's inequality.[2] By using this inequality, Proposition II.6 is improved.

[2] This is only true if $n \gg \log(\frac{1}{\delta})$, otherwise Azuma-Hoeffding's inequality is preferable.

Proposition II.7. *Consider two policy kernels $\hat{p}$ (empirical estimate) and p (true kernel). Assuming that $(\hat{p}(s)-p(s))h$ satisfies the single-sided Bernstein-style inequality:*

$$(\hat{p}(z)-p(z))h \le \sqrt{\frac{\mathbf{V}(p,h)\ell}{n}} + \frac{\mathrm{sp}(h)\ell}{n}, \tag{II.6}$$

where $n,\ell > 0$ are constants and $h \equiv h(p)$. If $g(p) \in \mathbf{R}e$ then the optimal gain under $\hat{p}$ is upper-bounded as:

$$g(\hat{p}) \le g(p) + \frac{\mathrm{sp}(h)\ell}{n} + \sqrt{\frac{2\mathrm{sp}(h)\ell}{n}}. \tag{II.7}$$

In particular, if $n \ge (\frac{2}{\epsilon^2} + \frac{1}{\epsilon})\mathrm{sp}(h^*)\ell$, then $g(\hat{p}) \le g(p) + 2\epsilon$.

Proof. The strategy starts similarly as in Proposition II.6. The expectations under $p, \hat{p}$ are respectively denoted $\mathbf{E}[-]$ and $\hat{\mathbf{E}}[-]$. We expand the aggregate rewards with $h(s) = r(s) - g(s) + p(s)h$, where g and h are the gain and the bias vectors under the dynamics p. We have:

$$\begin{aligned}
\hat{\mathbf{E}}\left[\sum_{t=0}^{T-1} r(S_t)\right] &= \hat{\mathbf{E}}\left[\sum_{t=0}^{T-1}(g + h(S_t) - p(S_t)h)\right] \\
&= Tg + \hat{\mathbf{E}}\left[\sum_{t=0}^{T-1}(e_{S_{t+1}} - p(S_t))h\right] \\
&= Tg + \hat{\mathbf{E}}[h(S_0) - h(S_T)] + \hat{\mathbf{E}}\left[\sum_{t=0}^{T-1}(\hat{p}(S_t) - p(S_t))h\right] \\
&\overset{(*)}{\le} Tg + \hat{\mathbf{E}}[h(S_0) - h(S_T)] + \hat{\mathbf{E}}\left[\sum_{t=0}^{T-1}\sqrt{\frac{\mathbf{V}(p(S_t),h)\ell}{n}}\right] + \frac{T\mathrm{sp}(h)\ell}{n} \\
&\overset{(\dagger)}{\le} Tg + \hat{\mathbf{E}}[h(S_0) - h(S_T)] + \hat{\mathbf{E}}\left[\sqrt{\frac{T\ell}{n}\cdot\sum_{t=0}^{T-1}\mathbf{V}(p(S_t),h)}\right] + \frac{T\mathrm{sp}(h)\ell}{n} \\
&\overset{(\ddagger)}{\le} Tg + \hat{\mathbf{E}}[h(S_0) - h(S_T)] + \sqrt{\frac{T\ell}{n}\cdot\hat{\mathbf{E}}\left[\sum_{t=0}^{T-1}\mathbf{V}(p(S_t),h)\right]} + \frac{T\mathrm{sp}(h)\ell}{n}
\end{aligned}$$

where $(*)$ follows by (II.6), $(\dagger)$ invokes Cauchy-Schwartz' inequality, and $(\ddagger)$ invokes Jensen's inequality. We now have to deal with the expected sum of variances $\hat{\mathbf{E}}[\sum_{t=0}^{T-1}\mathbf{V}(p(S_t),h)]$. Using Bellman's equation again: $\mathbf{V}(p(s),h) = p(s)h^2 - h^2(s) + 2h(s)(r(s)-g(s))$. Using that $\mathrm{sp}(h^2) \le \mathrm{sp}(h)^2$ and $\mathrm{sp}(r-g) \le 1$, we get:

$$\hat{\mathbf{E}}\left[\sum_{t=0}^{T-1}\mathbf{V}(p(S_t),h)\right] \le \mathrm{sp}(h)^2 + 2T\mathrm{sp}(h).$$

Plug it in the previous equation, divide by T and let it go to infinity. □

We obtain an optimal dependency in $\mathrm{sp}(h)$. If the support of $\hat{p}$ is moreover the same as p, or more generally if the recurrent states under $\hat{p}$ are a subset of those of under p, we change $\mathrm{sp}(h)$ for the span of the bias truncated to the recurrent states of the policy, leading to improved bounds for unichain policies. Proposition II.6 and Proposition II.7 are both given as upper bounds, but lower bounds can similarly be established. In the end, the bound based on Bersntein's inequality (Proposition II.7) provides deviations of the gain of a policy of the range of $\sqrt{\mathrm{sp}(h^\pi)/n}$.

5.2 Deviations of the optimal gain of a Markov decision process

If the bound of Proposition II.7 is tight and the gain of a policy truly varies with $\sqrt{\mathrm{sp}(h^\pi)}$ then we are in trouble, because the gain of the policy with the highest bias span is then difficult to estimate. Then minimax lower bound requires to make everything depend on $\mathrm{sp}(h^*)$, hence the variations of the gain of every policy must inevitably be controlled with respect to $\sqrt{\mathrm{sp}(h^*)}$ rather than $\sqrt{\mathrm{sp}(h^\pi)}$. Thankfully, there is the following remarkable result.

Proposition II.8. *Consider a communicating Markov decision process $M \equiv (\mathscr{X}, p, r)$ and let $\hat{p}$ another kernel. Under the assumption that $(\hat{p}(z) - p(z))h^*$ satisfies the single-sided Bernstein-style inequality:*

$$(\hat{p}(z) - p(z))h^*(M) \le \sqrt{\frac{\mathbf{V}(p(z), h^*)\ell}{n}} + \frac{\mathrm{sp}(h^*)\ell}{n} \tag{II.8}$$

where $n, \ell > 0$ are constants, then the optimal gain under $\hat{p}$ is upper-bounded as:

$$g^*(r, \hat{p}) \le g^*(M) + \frac{3\mathrm{sp}(h^*(M))}{2n} + \sqrt{\frac{2\mathrm{sp}(h^*(M))\ell}{n}}. \tag{II.9}$$

In particular, if $n \ge (\frac{2}{\epsilon^2} + \frac{3}{2\epsilon})\mathrm{sp}(h^*)\ell$, then $g^*(r, \hat{p}) \le g^*(M) + 2\epsilon$.

I have claimed earlier that a method achieving a regret of order $\sqrt{\mathrm{sp}(h^*)SAT}$ must indirectly prove an inequality of the form $\|g^*(\hat{M}_n) - g^*(M)\|_\infty \lesssim \sqrt{\mathrm{sp}(h^*)/n}$. The proof of Proposition II.8 was actually extracted from the regret analysis of EBF of Zhang and Ji (2019), which was then the only algorithm achieving minimax optimal regret, and can be thought as a heavily simplified regret analysis. Also, this proof is a good entry to the ideas behind the regret analysis of PMEVI.

Proof. We expand the reward again via the Poisson equation $h^*(s) = r(s,a) - g^*(s) + p(s,a)h^* + \Delta(s,a)$. Using this, we obtain:

$$\begin{aligned}
\hat{\mathbf{E}}^{\hat{\pi}}\left[\sum_{t=0}^{T-1} r(Z_t)\right] &= \hat{\mathbf{E}}^{\hat{\pi}}\left[\sum_{t=0}^{T-1}(g^*(S_t) + h^*(S_t) - p(Z_t)h^* - \Delta^*(Z_t))\right] \\
&= Tg^* + \hat{\mathbf{E}}^{\hat{\pi}}\left[\sum_{t=0}^{T-1}(h^*(S_t) - \hat{p}(Z_t)h^*)\right] \\
&\qquad + \hat{\mathbf{E}}^{\hat{\pi}}\left[\sum_{t=0}^{T-1}(\hat{p}(Z_t) - p(Z_t))h^*\right] - \hat{\mathbf{E}}^{\hat{\pi}}\left[\sum_{t=0}^{T-1}\Delta^*(Z_t)\right] \\
&\le Tg^* + \mathrm{sp}(h^*) + \hat{\mathbf{E}}^{\hat{\pi}}\left[\sum_{t=0}^{T-1}(\hat{p}(Z_t) - p(Z_t))h^*\right] - \hat{\mathbf{E}}^{\hat{\pi}}\left[\sum_{t=0}^{T-1}\Delta^*(Z_t)\right].
\end{aligned}$$

This is quite important not to throw again the negative term $\hat{\mathbf{E}}^{\hat{\pi}}[\sum_{t=0}^{T-1}\Delta^*(Z_t)]$ that will cancel out an important term shortly. We expand $(\hat{p}(Z_t) - p(Z_t))h^*$ using the assumed Bernstein-style inequality $(\hat{p}(Z_t) - p(Z_t))h^* \le \sqrt{\mathbf{V}(p(Z_t), h^*)\ell/n} + \mathrm{sp}(h^*)\ell/n$, leading to:

$$\hat{\mathbf{E}}^{\hat{\pi}}\left[\sum_{t=0}^{T-1}(\hat{p}(Z_t) - p(Z_t))h^*\right] = \frac{\mathrm{sp}(h^*)T\ell}{n} + \hat{\mathbf{E}}^{\hat{\pi}}\left[\sum_{t=0}^{T-1}\sqrt{\frac{\mathbf{V}(p(Z_t), h^*)\ell}{n}}\right]$$

$$\overset{(*)}{\leq} \frac{\mathrm{sp}(h^*)T\ell}{n} + \hat{\mathbf{E}}^{\hat{\pi}}\left[\sqrt{T\cdot\sum_{t=0}^{T-1}\frac{\mathbf{V}(p(Z_t),h^*)\ell}{n}}\right]$$
$$\overset{(\dagger)}{\leq} \frac{\mathrm{sp}(h^*)T\ell}{n} + \sqrt{\frac{T\ell}{n}\cdot\hat{\mathbf{E}}^{\hat{\pi}}\left[\sum_{t=0}^{T-1}\mathbf{V}(p(Z_t),h^*)\right]}$$

where $(*)$ follows from Cauchy-Schwartz' inequality and $(\dagger)$ from Jensen's inequality. We again have to control the term $\hat{\mathbf{E}}^{\hat{\pi}}[\sum_{t=0}^{T-1}\mathbf{V}(p(Z_t),h^*)]$. Using the Poisson equation again, we have:

$$\mathbf{V}(p(s,a),h^*) = p(s,a)h^{*2} - h^*(s)^2 + 2h^*(s)(\Delta^*(s,a) + r(s,a) - g^*(s)).$$

Therefore, and using $\mathrm{sp}(h^{*2}) \leq \mathrm{sp}(h^*)^2$, we get:

$$\hat{\mathbf{E}}^{\hat{\pi}}\left[\sum_{t=0}^{T-1}\mathbf{V}(p(Z_t),h^*)\right] \leq \mathrm{sp}(h^*)^2 + 2T\,\mathrm{sp}(h^*) + 2\mathrm{sp}(h^*)\hat{\mathbf{E}}^{\hat{\pi}}\left[\sum_{t=0}^{T-1}\Delta^*(Z_t)\right].$$

All together, we have:

$$\hat{\mathbf{E}}^{\hat{\pi}}\left[\sum_{t=0}^{T-1} r(Z_t)\right] \leq \left\{\begin{array}{c} Tg^* + \frac{\mathrm{sp}(h^*)T\ell}{n} + T\sqrt{\frac{2\mathrm{sp}(h^*)\ell}{n}} \\ + \sqrt{T\cdot\frac{2\mathrm{sp}(h^*)\ell}{n}\hat{\mathbf{E}}^{\hat{\pi}}\left[\sum_{t=0}^{T-1}\Delta^*(Z_t)\right]} - \hat{\mathbf{E}}^{\hat{\pi}}\left[\sum_{t=0}^{T-1}\Delta^*(Z_t)\right] \\ + \sqrt{T\cdot\frac{\mathrm{sp}(h^*)^2\ell}{n}} + \mathrm{sp}(h^*) \end{array}\right\} \tag{II.10}$$

Denoting $x := \hat{\mathbf{E}}^{\hat{\pi}}[\sum_{t=0}^{T-1}\Delta^*(Z_t)]$, we get an equation of the form $\hat{\mathbf{E}}^{\hat{\pi}}[\sum_{t=0}^{T-1} r(Z_t)] \leq T\alpha(T) + \beta(T)\sqrt{Tx} - x$ where the values of $\alpha(T)$ and $\beta(T)$ should be readily obvious. Straight forward analysis shows that $\beta(T)\sqrt{Tx} - x \leq \frac{1}{4}\beta^2(T)T$, therefore, by dividing by T in the above and letting it go to infinity, we get:

$$\hat{g}^{\hat{\pi}} \leq g^* + \frac{3\mathrm{sp}(h^*)\ell}{2n} + \sqrt{\frac{2\mathrm{sp}(h^*)\ell}{n}}.$$

To conclude the proof, observe that $3\mathrm{sp}(h^*)\ell/(2n) \leq \epsilon$ is equivalent to $n \geq \frac{3}{2\epsilon}\mathrm{sp}(h^*)\ell$, while $\sqrt{2\mathrm{sp}(h^*)\ell/n} \leq \epsilon$ is equivalent to $n \geq \frac{2}{\epsilon^2}\mathrm{sp}(h^*)\ell$. So, if $n \geq (\frac{2}{\epsilon^2} + \frac{3}{2\epsilon})\mathrm{sp}(h^*)\ell$, both are satisfied. This concludes the proof. □

This result is remarkable for the following reason: Whatever the policy, under kernel perturbation, its gain cannot move to much beyond the optimal gain with respect to $\sqrt{\mathrm{sp}(h^*)}$ rather than $\sqrt{\mathrm{sp}(h^\pi)}$. In opposition to the single policy results Propositions II.6 and II.7, the proof technique of Proposition II.8 cannot be used to show that $g^\pi(\hat{p}) \geq g^\pi(p) - \mathrm{O}(\sqrt{\mathrm{sp}(h^*)/n})$ and this is actually wrong. However, if policy is a bias optimal policy and if $(\hat{p}(z) - p(z))h^* \geq -\sqrt{\mathbf{V}(p(z),h^*)\ell/n} - \mathrm{sp}(h^*)\ell/n$, Proposition II.7 can be adapted to show that $g^\pi(\hat{p}) \geq g^\pi(p) - \mathrm{O}(\sqrt{\mathrm{sp}(h^*)/n})$. In other words, all bias optimal policies of M are nearly optimal in (r,p).

The technique used in the proof of Proposition II.2 is, to some extent, a prototype version of the regret analysis of `PMEVI`. In (II.10), we observe the presence of $x = \sum_{t=0}^{T-1}\Delta^*(Z_t)$ with is nothing less than the first order regret induced by iterating that policy, and we already see appear an equation mixing x and $\sqrt{x}$ where we show that the term $\sqrt{x}$ is negligible and can be ignored. This kind of argument will come back in the analysis of `PMEVI`.

Corollary II.9. *Consider a communicating Markov decision process $M \equiv (\mathscr{X}, p, r)$ and let $\hat{p}$ another kernel. Under the assumption that $(\hat{p}(z) - p(z))h^*$ satisfies the two-sided Bernstein-style inequality:*

$$|(\hat{p}(z) - p(z))h^*(M)| \leq \sqrt{\frac{\mathbf{V}(p(z),h^*)\ell}{n}} + \frac{\mathrm{sp}(h^*)\ell}{n} \tag{II.11}$$

where $n, \ell > 0$ are constants, then the optimal gain under $\hat{p}$ satisfies

$$\|g^*(r,\hat{p}) - g^*(M)\|_\infty \leq \frac{6\mathrm{sp}(h^*(M))}{2n} + 2\sqrt{\frac{2\mathrm{sp}(h^*(M))\ell}{n}}. \tag{II.12}$$

Also, if $n \geq (\frac{2}{\epsilon^2} + \frac{3}{2\epsilon})\mathrm{sp}(h^*)\ell$, then $\|g^*(r,\hat{p}) - g^*(M)\|_\infty \leq 4\epsilon$ and all bias optimal policies of M are 4ϵ-gain optimal in $\hat{M} \equiv (r,\hat{p})$.

5.3 A few comments on the optimality of these bounds

Corollary II.9 convey the main message and illustrate the technique behind the heavier proof of `PMEVI` in Chapter 7. However, much more could be said about gain deviation bounds. For instance, $\mathrm{sp}(h^*)$ could actually be changed to the span of the bias function truncated to the recurrent states of the optimal policy. Actually, the dependence of all these bounds in $\sqrt{\mathrm{sp}(h)T}$ can be improved to $\sqrt{T\sum_s \mu(s)\mathbf{V}(p(s),h)}$ by avoiding to expand the sum of variances. The displayed technique can also be adapted to bound the deviations of $\sum_{t=0}^{T-1} R_t$ in high probability and leads to similar bounds. In this direction, Freedman's inequality (Lemma I.20) together from ideas of the original works of Freedman (1975) lead to an iterated logarithm law for aggregate rewards and central limit theorems. Concerning the way $h^*(M)$ may be used to bound the gain deviations of all policies, it is likely that a few interesting points are still to be understood. However, the fixed policy setting is morally dealt with, because it is about the asymptotic behavior of additive functionals on Markov chains and this problem has been intensively investigated. As an entry point to this literature, one can go through the works of Maxwell and Woodroofe (2000); Peligrad (2020) about central limit theorems on $\sum_{t=0}^{T-1} R_t$ and explore the neighborhood literature.

Chapter 6

Optimism in the face of uncertainty

The nearly minimax optimal algorithm presented in this chapter belongs to a large family of algorithms that follow the **optimism-in-the-face-of-uncertainty** principle (OFU), stating that whenever one lacks clear evidence to determine the value of something, this value should be estimated as high as statistically plausible; In other words, be optimistic about the so-called value. The term *value* is purposely vague as there are multiple way to implement optimism, the kind of objects that are attached a value, or the very nature of this value. To give an idea of why optimism is a natural approach to regret minimization, let me provide a variant of the motivation of Lattimore and Szepesvári (2020) that people that know me will recognize.

Say that you wander in a city, looking for the best place to drink a coffee. Although Google Maps provides prior information on the coffee that you will get, it is only by going at a coffee shop that you may know if the place suits you. And yet, your experience is subject to stochasticity, as the music can be trapped in the worst part of the playlist, the current coffee brew may not be to your taste, today's sweets may not be to your liking or a bunch of children may have decided to investigate the place. Nonetheless, some coffee shops are better than others, depending on how much the baristas are coffee nerds, the skills of the cake baker, the location or whatnot. By being optimistic, you will try most of the places downtown several times, and progressively have a better and better idea of the place that suits the best, so that you can progressively be less and less optimistic and spend most of your time in your favorite coffee shops. By being pessimistic, you will try a few places until you find one that meets your basic requirements, and are very likely to completely miss the place that would have been the best. You may even flag as bad a place that is actually great, just because the only time you tried it, you had a worse experience than what the place usually offers.

How much should you be optimistic depends on how much you are sensitive to the coffee's quality (i.e., the cost of a suboptimal action) but also on how much you dislike changing places, especially during summer (i.e., the cost of switching strategy). The second point is already a difference with the book of Lattimore and Szepesvári (2020), because it means that we absolutely want to avoid changing strategies too often.

> **Important remark.** Optimism is not the only way. The main challenger to optimism is **posterior sampling**, where in face of several possible strategies, you pick one proportionally to how much you believe this strategy is likely to be optimal. This line originates from repeatedly rediscovered work of Thompson (1933) which is originally a learning algorithm for multi-armed bandits. A few variants exist for Markov decision processes, including `PSRL` of Osband and Roy (2017); Osband et al. (2013), `TSDE` of Ouyang et al. (2017) and `Optimistic-PSRL` Agrawal and Jia (2023). The ideas of posterior sampling are very promising and do seem, to my taste, to be largely under explored in undiscounted

infinite horizon reinforcement learning.

6.1 Confidence regions and policy-wise optimism

Perhaps the first algorithm to implement the idea of optimism in the multi-armed bandit setting is the seminal paper of Lai and Robbins (1985) and optimism is already present in the paper of Burnetas and Katehakis (1997) on ergodic Markov decision processes. The algorithm PMEVI presented downstream relies on a **policy-wise optimism** that can be given credit to UCRL2 of Auer et al. (2009), at least in its current form. The idea is that at a given time, the planner will choose which policy to play based on the maximal plausible gain of that policy. This maximal plausible gain is will referred to as the **optimistic gain** of the policy and depends on how the planner shapes their uncertainty; Namely, on the confidence region built for the hidden model M. This confidence region $\mathscr{M}_t$ is built out of the current observations $O_t := (S_0, A_0, R_0, \dots, S_t)$. The optimistic gain of policy π is computed as the largest achievable gain $g^\pi(\tilde{M})$ for $\tilde{M} \in \mathscr{M}_t$.

Definition II.2. *For $\pi \in \Pi^{\mathrm{SR}}$, the **optimistic gain of π** on $\mathscr{M}_t$ is the vector $g^\pi(\mathscr{M}_t) \in \mathbf{R}^{\mathcal{S}}$ given by:*

$$g^\pi(s; \mathscr{M}_t) := \sup_{\tilde{M} \in \mathscr{M}_t} g^\pi(s; \tilde{M}). \tag{II.1}$$

*The **optimistic gain** is $\max_\pi g^\pi(\mathscr{M}_t)$.*

These optimistic gains will be later rewritten as optimal gains of well chosen models.

The general architecture of policy-wise optimistic algorithms is given with Algorithm II.1. Over time, the planner maintain a **current policy** π^k that is used to pick actions until it is decided obsolete and is renewed. The time segment $\{t_k, \dots, t_{k+1} - 1\}$ on which this policy is used is referred to as an **episode**. The test (on line 3) deciding whether a policy should be changed or not is the **episode rule**. When the policy is changed (on line 4), it is picked as a policy maximizing the optimistic gain from the current state; We say that the planner picks an **optimistic policy**. This choice may not be unique, especially when information is lacking in the early learning phases, and ties are broke according to the **tie breaking rule**.

Algorithm II.1 The architecture of policy-wise optimistic algorithms.

```
1:  k ← 0, initialize π^0;
2:  for t = 0, 1, ... do
3:      if current policy π^k is obsolete then
4:          π^k ← arg max_π g^π(S_t; M_t);
5:          k ← k + 1;
6:          t_k ← t;
7:      end if
8:      π_t ← π^k;
9:      Iterate π_t, i.e., play A_t ~ π_t(−|S_t), observe reward R_t and transition S_{t+1};
10: end for
```

As discussed in Bourel et al. (2020); Fruit (2019); Fruit et al. (2018), breaking ties by randomizing the policy is usually a good choice because ties usually happens in the early phase where exploration must be prioritized. Apart from tie breaking, policy-wise optimistic algorithms have two essential elements of design.

(1) (**Confidence region**) The choice of confidence region $\mathscr{M}_t$; and

(2) (**Episode rule**) The way policies are terminated, i.e., episodes ended.

The choice of confidence region has a direct impact on the regret. If too loose, the algorithm will be over-optimistic and play suboptimal policies for longer times than required; If too narrow, the algorithm will be under-optimistic and has non-negligible chances to utterly miss the optimal policy and commit instead to playing a suboptimal policy. The choice of episode rule is also important, but do not need careful tuning to achieve minimax optimal regret. It plays a different role that will be discussed in Part IV.

So, how should be the confidence region be chosen?

6.2 Extended MDPs and Extended Value Iteration (EVI)

Behind the choice of the confidence region hides a significant difficulty. How should the optimistic policy be computed (at line 4)? A few works, such as Bartlett and Tewari (2009); Zhang and Ji (2019) avoid the question and make use of oracles to extract optimistic policies from their confidence region. If the confidence region has a specific form however, optimistic policies can be computed with a variant of Value Iteration (Algorithm I.1) since Nilim and El Ghaoui (2005).

Definition II.3 (Nilim and El Ghaoui (2005))**.** *A confidence region $\mathcal{M}_t$ is said **rectangular** or in **product form** if it can be written as $\mathcal{M}_t \equiv \prod_{z\in\mathcal{Z}}(\mathcal{R}_t(z)\times\mathcal{P}_t(z))$, where $\mathcal{R}_t(z)\subseteq\mathbf{R}$ and $\mathcal{P}_t(z)\subseteq\mathcal{P}(\mathcal{S})$ are the respective confidence regions for $r(z)$ and $p(z)$.*

In other words, a confidence region is in product form if it is made of a collection of independent confidence region: one for every reward and kernel. Under this assumption, $\mathcal{M}_t$ can be seen a Markov decision process with compact action space that we call an **extended** Markov decision process. The proper construction is due to Auer et al. (2009).

Definition II.4 (Auer et al. (2009))**.** *Given a compact confidence region $\mathcal{M}_t$ in product form, the **extended formulation** of $\mathcal{M}_t$ is the Markov decision process similarly denoted $\mathcal{M}_t \equiv (\widetilde{\mathcal{Z}}_t, \tilde{p}, \tilde{r})$ with state space $\mathcal{S}$ and action space:*

$$\widetilde{\mathcal{A}}_t(s) := \bigcup_{a\in\mathcal{A}(s)} \{a\}\times\mathcal{R}_t(s,a)\times\mathcal{P}_t(s,a) \tag{II.2}$$

*called **extended actions**. It extended pair space is denoted $\widetilde{\mathcal{Z}}_t$. When choosing an extended action $\tilde{a}=(a,r'(s,a),p'(s,a))$, we have $\tilde{r}(s,\tilde{a}) := r'(s,a)$ and $\tilde{p}(-|s,\tilde{a}) := p'(-|s,a)$.*

Remark that policies of $\mathcal{M}_t$ are **extended policies** and consist in tuples $\tilde{\pi}\equiv(\pi,p',r')$ consisting in a standard policy π as well as a Markov reward process (r',p') modeling the dynamics under this policy. By seeing $\mathcal{M}_t$ as a Markov decision process, we can import all the standard machinery of Markov decision process introduced in Part I: the gain, the bias, Poisson equations, Bellman equations and algorithms that compute optimal policies such as Value Iteration (Algorithm I.1).

Yet, there is an obvious issue.

6.2.1 The Pitfall: Compact action spaces and Bellman equations

The extended formulation of $\mathcal{M}_t$ has compact action space space, and in Part I only gave a treatment of finite state and action spaces models. As a matter of fact, Bellman equations are not always guaranteed to have solutions if the action space is infinite (e.g., continuous and compact). This disproportionate detail is not mild although it is often overlooked in the reinforcement

learning literature. To my knowledge, only Fruit (2019) mentions the issue. Some works spot the issue from afar to miraculously dodge it, because $\mathscr{M}_t$ can sometimes be reduced to a finite action space, in particular when $\mathscr{M}_t$ is a polyhedron, for example in Auer et al. (2009); Bourel et al. (2020); Fruit et al. (2020, 2018). Some other works fall short of handling the issue and this is the case of KLUCRL of Filippi et al. (2010); Maillard (2019).

The existence of solutions to the Bellman equations for (weakly communicating) infinite action spaces Markov decision processes is due to Schweitzer (1985) and requires non trivial assumptions. It the action spaces are compact however, Schweitzer (1987) provides a simpler proof under the communicating assumption with a proof based on Brouwer's fixpoint theorem; The communicating assumption is someway mandatory because beyond the communicating setting, optimal policies of compact action spaces models are not guaranteed to be time-independent anymore. Before diving in, randomized policies of models with compact action spaces need a few words, because the action space may become infinite. When the action space is compact, a randomized policy π is map $\pi : s \in \mathcal{S} \mapsto \pi(-|s) \in \mathcal{P}(\mathcal{A}(s))$ where $\mathcal{P}(\mathcal{A}(s))$ is the space of Borel probability measures on $\mathcal{A}(s)$. By continuity of the reward and kernel functions, a randomized policy has a well-defined reward vector r^π and kernel p^π. Therefore, even though the compact action space may be infinite and the policy π randomized, its gain and bias functions g^π and h^π are still well-defined by Definition I.5 and Definition I.7. The space of deterministic and randomized policies are respectively denoted Π^{SD} and Π^{SR}.

Proposition II.10 (Schweitzer (1985))**.** *Let $M \equiv (\mathcal{X}, p, r)$ be a weakly-communicating Markov decision process with finite state space $\mathcal{S}$ and compact action spaces $\mathcal{A}(s)$.*[a] *Let $g^*(s) := \sup_{\pi\in\Pi^{\mathrm{SR}}} g^\pi(s)$ the optimal gain vector and denote $\Pi^* := \{\pi \in \Pi^{\mathrm{SR}} : g^\pi = g^*\}$. If the condition*

$$\Pi^* \cap \Pi^{\mathrm{SD}} \neq \varnothing \quad \textit{and} \quad \sup_{\pi\in\Pi^*\cap\Pi^{\mathrm{SD}}} \max(h^\pi) < \infty \tag{II.3}$$

is met, then $g^* \in \mathbf{R}e$ and there exists $h^* \in \mathbf{R}^{\mathcal{S}}$ satisfying the Bellman equation $g^* + h^* = \max_{\pi\in\Pi}(r^\pi + P^\pi h^*)$. Moreover, any greedy response to h^* achieves optimal gain g^*.

[a] We assume that $r(s,-)$ and $p(s,-)$ are continuous functions of $a \in \mathcal{A}(s)$.

The first condition "$\Pi^* \cap \Pi^{\mathrm{SD}} \neq \varnothing$" states that gain optimal stationary policies must exist. The second condition "$\sup_{\pi\in\Pi^*\cap\Pi^{\mathrm{SD}}} \max(h^\pi)$" means that no gain optimal policies can make a state arbitrarily good, because $h^\pi(s)$ measures what is scored in addition to the gain from state, hence the higher $h^\pi(s)$, the greater the advantage to initialize the dynamics at s under π. Roughly speaking, the higher is $h^\pi(s)$, the better is s under the policy π. The result of Schweitzer (1985) is actually an equivalence and the pair of conditions (II.3) is also necessary for the existence of a solution to the Bellman equations. If failing, the Bellman equations have no solution.

Remarkably, if the action space is compact, the communicating assumption is sufficient.

Proposition II.11 (Schweitzer (1987))**.** *Let $M = (\mathcal{X}, p, r)$ a communicating Markov decision process with finite state space s with compact action space.*[a] *Then:*

1. *Its Bellman operator $L : u \in \mathbf{R}^{\mathcal{S}} \mapsto \max_{\pi\in\Pi^{\mathrm{SD}}}(r^\pi + P^\pi u)$ admits a span-fixpoint, i.e., $\exists u \in \mathbf{R}^{\mathcal{S}}, Lu - u \in \mathbf{R}e$;*
2. *If $p(s|s,a) > 0$ for all $(s,a) \in \mathcal{X}$, then the iterates of the Bellman operator converge to a span fixpoint with geometric speed, i.e., there is $\gamma < 1$ such that whatever the initialization $u \in \mathbf{R}^{\mathcal{S}}$, $\mathrm{sp}(L^{n+1}u - L^n u) = \mathrm{o}(\gamma^n)$ when $n \to \infty$.*

[a] We assume that $r(s,-)$ and $p(s,-)$ are continuous functions of $a \in \mathcal{A}(s)$.

The existence of a span fixpoint of the Bellman operator is equivalent to the solution to the Bellman equations. Indeed, if h^* is the span fixpoint, then letting $g^* := L \in \mathbf{R}e$, we see that $g^*(s) + h^*(s) \geq r(s,a) + p(s,a)h^*$ with equality for at least one action $a \in \mathscr{A}(s)$, showing that Bellman optimal policies exist (see Definition I.9). Such greedy responses to u (see Definition I.12) are gain optimal (see Proposition I.4) and Proposition I.8 is straightforwardly generalized to models with compact action spaces: Greedy responses to ϵ-fixpoints of the Bellman operator are ϵ-gain optimal policies. Moreover, the convergence of the iterates of L is fast provided that $p(s|s,a) > 0$ that can always be assumed up to an aperiodic transform, hence guaranteeing the convergence of Lazy Value Iteration (Algorithm I.2).

The communicating assumption is not necessary in Proposition II.11, but one enters dangerous territories by dropping it as gain optimal policies may not be stationary anymore, see Leizarowitz (2002). Thankfully, all the extended Markov decision processes that we will be considering correspond to communicating compact action space models, hence Proposition II.11 will be enough to guarantee the well-definition of their gain and computational tractability.

6.2.2 Optimistic models and Extended Value Iteration

With Proposition II.11, we can provide generic properties guaranteeing the well-definition of the optimistic gain. We can now provide a more complete description of the optimistic gain and introduce the notion of **optimistic model**.

Corollary II.12. *Let $\mathscr{M}_t$ a rectangular confidence region and let $\mathscr{L}_t$ the Bellman operator of its extended formulation, called the **extended Bellman operator**. Assume that, for all $z \in \mathscr{Z}$, there exists $p'(z) \in \mathscr{P}_t(z)$ of full support. Then:*

(1) *$\mathscr{L}_t$ has a span fix-point and the optimistic gain satisfies $g^*(\mathscr{M}_t) = \max_\pi g^\pi(\mathscr{M}_t)$;*
(2) *Given $\pi \in \Pi(\mathscr{Z})$, define $\mathscr{M}_t^\pi := \prod_{s\in\mathscr{S}} \sum_{a\in\mathscr{A}(s)} \pi(a|s)(\mathscr{R}_t(s,a) \times \mathscr{P}_t(s,a))$ the confidence region for π, and let $\mathscr{L}_t^\pi$ the Bellman operator of its extended formulation. The optimistic gain of π satisfies $g^\pi(\mathscr{M}_t) = g^*(\mathscr{M}_t^\pi)$;*
(3) *Running Lazy Value Iteration (Algorithm I.1) with $\mathscr{L}_t$ or $\mathscr{L}_t^\pi$ converges geometrically fast to a near span fixpoint of the respective operators.*

*Every extended greedy response to $\mathscr{L}_t$ is a triple (π, r', p') of a choice of actions, of rewards and of kernels of these actions. The couple (r', p') is called an **optimistic model of π**.*

The condition "*for all $z \in \mathscr{Z}$, there exists $p'(z) \in \mathscr{P}_t(z)$ of full support*" is always met, but is worth pointing out. As explained earlier, if the extended $\mathscr{M}_t$ is not communicating (e.g., multi-chain), then none of the above is guaranteed to hold.

From a computational perspective, the third assertions states that Lazy Value Iteration is guaranteed to converge. In the literature, most algorithms rely on Value Iteration (Algorithm I.1) to compute optimistic gains and policies that converges without requiring an aperiodicity transform, because the extended model $\mathscr{M}_t$ is usually aperiodic already. The algorithm consists in iterating the extended Bellman operator:

$$\begin{aligned}\mathscr{L}_t u(s) &:= \max_{a\in\mathscr{A}(s)} \max_{r'(s,a)\in\mathscr{R}_t(s,a)} \max_{p'(s,a)\in\mathscr{P}_t(s,a)} \left\{r'(s,a) + p'(s,a)u\right\} \\ &= \max_{a\in\mathscr{A}(s)} \left\{\max(\mathscr{R}_t(s,a)) + \max_{p'(s,a)\in\mathscr{P}_t(s,a)} p'(s,a)u\right\}.\end{aligned} \tag{II.4}$$

We observe that the computation of the extended Bellman operator (II.4) can be decoupled as a maximization problem other rewards ($\max(\mathscr{R}_t(s,a))$) which is generally trivial, and as the maximization of a linear functional ($\max_{p'(s,a)\in\mathscr{P}_t(s,a)} p'(s,a)u$) which is solved with confidence

region specific algorithms and oppose no real difficulty in general. Because the extended Bellman operator is iterated in place of the Bellman operator, this algorithm is referred to as **Extended Value Iteration** (EVI), see Auer et al. (2009). Once it has converged to an ϵ_t-span fixpoint, the algorithm extracts a greedy policy to deploy as the current policy.

Algorithm II.2 EVI-based optimistic method.

$k \leftarrow 0$, initialize π^0;
for $t = 0, 1, \dots$ **do**
  if current policy π^k is obsolete **then**
    $u^k \leftarrow \mathtt{EVI}(\mathscr{M}_t, \epsilon_t, 0^{\mathscr{S}})$;
    $\pi^k \leftarrow$ any π s.t. $\mathscr{L}_t(u^k) = \mathscr{L}_t^{\pi}(u^k)$;
    $k \leftarrow k+1$; $t_k \leftarrow t$.
  end if
  Set $\pi_t \leftarrow \pi^k$ and iterate π_t.
end for

Algorithm II.3 Extended Value Iteration (EVI).

Parameters: A region $\mathscr{M}_t$, a precision $\epsilon > 0$, an (optional) $u_0 \in \mathbf{R}^{\mathscr{S}}$;

if u_0 is not initialized **then** $u_0 \leftarrow 0 \cdot e$;
for $n = 1, 2, \dots,$ **do**
  $u_n \leftarrow \mathscr{L}_t u_{n-1}$; ▷ *Extended B.O.*
  if $\mathrm{sp}(u_n - u_{n-1}) < \epsilon$ **then break;**
end for
return u_n.

6.3 EVI-based algorithms in the literature

Many existing algorithms are specific instances of Algorithm II.2 with different confidence regions, depending on the norm used to measure errors, and variants of the same episode rule.

6.3.1 The eminent doubling trick

Most methods use the **doubling trick** (DT) or variants thereof to manage episodes. Simply stated, the doubling trick says that the episode must be terminated once a playable pair has doubled its visit counts since the beginning of the episode. Recall that the **visit counts** of a pair is given by $N_T(z) := \sum_{t=0}^{T-1} \mathbf{1}(Z_t = z)$. The doubling trick is formally given by:

$$t_{k+1} = \inf\big\{t > t_k : \exists a \in \mathscr{A}(S_t), N_t(S_t, a) \geq 1 \vee 2N_{t_k}(S_t, a)\big\}. \tag{DT}$$

When π_{t-1} is deterministic, this condition is equivalent to $\pi_{t-1}(S_t)) \geq 2N_{t_k}(S_t, \pi_{t-1}(S_t))$, i.e., the pair that is about to be played has doubled its number of visits since the beginning of the episode. This simple rule is easy to implement and will be enough to achieve minimax optimal regret.

Remarkably, it guarantees that the number of episodes $K(T)$ is logarithmic $K(T) = \mathrm{O}(\log(T))$, see Auer et al. (2009).

6.3.2 Choosing the right confidence region

The minimax lower bound points out that the tedious part of the learning task comes from learning kernels; Also most of the literature focuses on finding the type of kernel confidence region that provides the better regret guarantees. Among them, three are dominant.

- **ℓ_1-confidence regions**. The errors on kernels are quantified using the ℓ_1-norm. Such confidence regions are constructed out of Weissman's inequality (Lemma I.23) and take the form of:

$$\mathscr{P}_t(z) := \{p'(z) \in \mathscr{P}(\mathscr{S}) : \|p'(z) - \hat{p}_t(z)\|_1^2 \leq Sx\}. \tag{II.5}$$

- **KL-confidence regions**. The error on kernels are quantified using the Kullback-Leibler divergence. Such confidence regions are constructed using inequalities on the empirical likelihood of observation (Lemma I.25) and take the form of:

$$\mathscr{P}_t(z) := \{p'(z) \in \mathscr{P}(\mathscr{S}) : \mathrm{KL}(p'(z)||\hat{p}_t(z)) \leq Sx\}. \tag{II.6}$$

- **Bernstein confidence regions**. The error on kernels are quantified using an empirical Bernstein inequality (Lemma I.26). Such confidence regions take the form of:

$$\mathscr{P}_t(z) := \left\{p'(z) \in \mathscr{P}(\mathscr{S}) : \forall s \in \mathscr{S}, |p'(s|z) - \hat{p}_t(s|z)| \leq \sqrt{\hat{p}_t(s|z)(1-\hat{p}_t(s|z))x} + x\right\}. \tag{II.7}$$

In (II.5), (II.6) and (II.7), $\hat{p}_t(z)$ is the empirically observed transition kernel at z at time t and $x > 0$ measures the confidence level of the confidence region and is typically of order $x \asymp \log(t/\delta)/N_t(z)$ where $N_t(z)$ is the number of visits of the pair z prior to time t and $\delta > 0$ is the desired probability of error. For instance, ℓ_1-confidence regions include `UCRL` Auer and Ortner (2006), `UCRL2` Auer et al. (2009); KL-confidence regions are specific to `KLUCRL` Filippi et al. (2010); Talebi and Maillard (2018); Bernstein-confidence regions include `UCRL2-B` Fruit et al. (2020), `UCRL-V` Tossou et al. (2019) and arguably `UCRL3` Bourel et al. (2020).[1]

The shape of these confidence regions are displayed on Figure 6.1, that shows clearly that no confidence region is overwhelmingly better than the others. In practice, when the confidence level $x > 0$ is bounded away from 0, they all have pro and cons. The ℓ_1 and Bernstein confidence regions are simple polyhedral shapes that are algorithmically convenient. The ℓ_1-confidence region is completely symmetric while Bernstein's inequality isn't, by taking into account transition specific variances. However, especially for $x \gg 1$, Bernstein's inequality is often worse than Weissman's inequality meaning that for rarely visited pairs, Weissman's inequality is preferable to Bernstein's. The KL-confidence region is smoother, but is more computationally expensive to track. However, it leads to much better regret than the two others in experiments. Also, although the regions provided by (II.5), (II.6) and (II.7) are qualitative, tuning $x > 0$ is important in practice; Every constant matters as it will impact the behavior of the algorithm in the early steps and influence the tendency of the algorithm to over-explore.

These kernel confidence regions are easily generalized to provide reward confidence regions as well, although many works don't bother much about rewards and choose a ℓ_1-confidence region based on Azuma-Hoeffind's inequality (Lemma I.19).

6.4 Regret analysis and encountered challenges

The regret of Algorithm II.1 is decomposed episodically, as the actually collected rewards are compared to what the planner is expecting to obtain. At episode k, the current policy is $\pi^k = \pi_{t_k}$, and we denote $(\tilde{r}^k, \tilde{p}^k)$ the optimistic model of π^k under $\mathscr{M}_{t_k}$ (Corollary II.12). To simplify the computation, we assume that π^k is deterministic. The optimistic gain and bias vectors of π^k are denoted $\tilde{g}^k := g(\tilde{r}^k, \tilde{p}^k)$ and $\tilde{h}^k := h(\tilde{r}^k, \tilde{p}^k)$. The regret decomposition follows these steps:

[1] `UCRL3` is heavily hand-tuned, making it difficult to fit this simple classification. It relies on finely tuned confidence bounds, based on Azuma-Hoeffding's inequality with tuned variance parameters, making them **variance-aware** and close in spirit to Bernstein's inequality. Also, `UCRL3` relies on a variant of `EVI` called `EVI-NOSS` that produces near-optimistic policies and tries to stick to optimistic kernels $p'(z) \in \mathscr{P}_t(z)$ with the same support as the empirical kernel $\hat{p}_t(z)$.

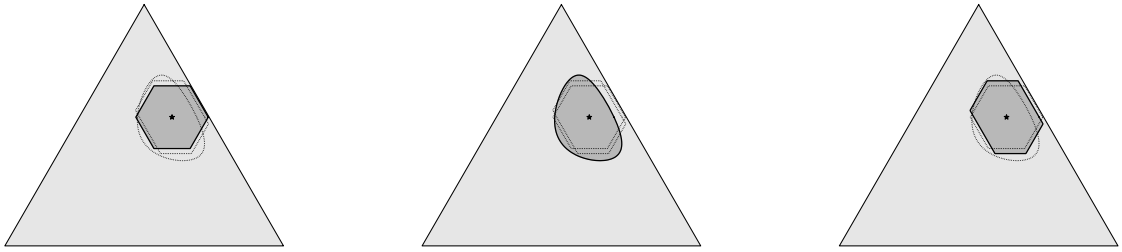


Figure 6.1: The big three of confidence regions in reinforcement learning. (From left to right) ℓ_1-confidence region, KL-confidence region and Bernstein confidence region for a three-dimensional kernel $p = \frac{1}{15}(8,2,5)$ and the same confidence level. On every plot, the main confidence region is displayed with a filled gray region, and the two others are displayed with a dotted outline for easier comparison.

Main terms in the regret decomposition of Auer et al. (2009):

$$Tg^* - \sum_{t=0}^{T-1} R_t \;\to\; \sum_{t=0}^{T-1}(g^* - r(Z_t)) \;\to\; \sum_k \sum_{t=t_k}^{t_{k+1}-1} (\tilde{g}^k - \tilde{r}^k(S_t)) \;\to\; \sum_k \sum_{t=t_k}^{t_{k+1}-1} (e_{S_t} - \tilde{p}^k(S_t))\tilde{h}^k \;\to\; \sum_k \sum_{t=t_k}^{t_{k+1}-1} (p^{\pi^k}(S_t) - \tilde{p}^k(S_t))\tilde{h}^k$$

Definition of $\mathrm{Reg}(T)$ | Remove reward noise | Episodic decomposition + optimism | Optimistic Poisson equation of π^k | Focus on optimistic dynamical error

This decomposition is due to Auer et al. (2009). At every step spawns an error term that is negligible in front of the **optimistic dynamical error** $\sum_k \sum_{t=t_k}^{t_{k+1}-1} (p^{\pi^k(S_t)} - \tilde{p}^k(S_t))\tilde{h}^k$ which is usually the focus of the analysis. To simplify notations a little bit, we write p^k for the true kernel of π^k. The initial technique of Auer et al. (2009) is to bound $(p^k(S_t) - \tilde{p}^k(S_t))\tilde{h}^k(S_t)$ with the improved Hölder's inequality $|(p^k(S_t) - \tilde{p}^k(S_t)|\tilde{h}^k \le \frac{1}{2}\|p^k(S_t) - \tilde{p}^k(S_t)\|_1 \mathrm{sp}(\tilde{h}^k)$. If the confidence region $\mathscr{M}_{t_k}$ doesn't fail, it contains M hence $D(\mathscr{M}_{t_k}) \le D(M)$, so by Proposition II.2 the optimistic span is bounded by the true diameter $\mathrm{sp}(\tilde{h}^k) \le D(M)$. Intuitively, this means that optimism doesn't make some states unreasonably better than others. Hence, provided that $M \in \mathscr{M}_{t_k}$ for all $k \ge 1$, the optimistic dynamical error is bounded as:

$$\sum_k \sum_{t=t_k}^{t_{k+1}-1} \big(p^k(S_t) - \tilde{p}(S_t)\big)\tilde{h}^k \lesssim \frac{1}{2} D(M) \sum_k \sum_{t=t_k}^{t_{k+1}-1} \big\|p^k(S_t) - \tilde{p}(S_t)\big\|_1. \tag{II.8}$$

For UCRL2, that uses ℓ_1-confidence region for kernels, this norm is $\mathrm{O}(\sqrt{S\log(T)/N_{t_k}(S_t, \pi^k(S_t))})$. The doubling trick makes sure that $N_t(Z_t) \le 2N_{t_k}(Z_t)$, hence we obtain:

$$\begin{aligned}
\sum_k \sum_{t=t_k}^{t_{k+1}-1} \big(p^k(S_t) - \tilde{p}(S_t)\big)\tilde{h}^k &\approx \mathrm{O}\left(D(M) \sum_k \sum_{t=t_k}^{t_{k+1}-1} \sqrt{\frac{S\log(T)}{N_t(Z_t)}} \right) \\
&= \mathrm{O}\left(D(M) \sum_{z\in\mathscr{Z}} \sum_{n=1}^{N_T(z)} \sqrt{\frac{S\log(T)}{n}} \right) \\
&= \mathrm{O}\left(D(M) \sum_{z\in\mathscr{Z}} \sqrt{SN_T(z)\log(T)} \right) \overset{(*)}{=} \mathrm{O}\Big(D(M)\sqrt{S\cdot SAT\cdot\log(T)} \Big)
\end{aligned}$$

where $(*)$ uses that $\sum_{z\in\mathscr{Z}} N_T(z) = T$ to conclude with Jensen's inequality. We obtain a regret bound of order $DS\sqrt{AT\log(T)}$. Depending on the kernel confidence region, the inequality (II.8)

is changed for another. Past (II.8), the computation is tight, i.e., the computations that follow don't lose information. Prior to (II.8), one can check that the error terms are $\mathrm{O}(\sqrt{SAT\log(T)})$, hence are negligible in front of $\sqrt{DSAT\log(T)}$. The extra $\sqrt{DS}$ is therefore due to the bound (II.8). Improving the bound on (II.8) is the main focus of all the literature on optimistic methods.

6.4.1 Shaving one $\sqrt{D}$ with variance aware confidence regions

In Chapter 5, we have argued that the Hölder bound on $(\hat{p}_t - p)u$ is incapable of leading to tight bounds on the deviation of the gain function; This is no surprise that the computation following (II.8) in Section 6.4 is short of getting the right dependency on the diameter. This is because Hölder's inequality does not take the variance into account. By using variance-aware concentration inequalities and variance-aware kernel confidence regions (such as the KL-confidence or the Bernstein confidence region), the right dependency on the diameter can be obtained. Examples are UCRL2-B Fruit et al. (2020) and UCRL3 Bourel et al. (2020) (Bernstein confidence regions) and KLUCRL Talebi and Maillard (2018) (KL-confidence region). Despite addressing the diameter dependency correctly, these methods still suffer from an extra $\sqrt{S}$ in their regret guarantees.

This can be explained as follows. At the end of the day, we want to bound $(\tilde{p}(z)-p(z))\tilde{h}$ and this is done by quantifying the maximum likely error on $(p(z)-\hat{p}_n(z))u$, where $\hat{p}_n(z)$ is the empirically observed transition kernel at z, for $u \in \mathbf{R}^{\mathscr{S}}$ with $\|u\|_1 = 1$ up to normalizing. With a bit of information theory, with n samples, the best bound possible is of order:

$$|p(z)-\hat{p}_n(z)|u \leq \sqrt{\frac{2\mathbf{V}(p(z),u)\log\left(\frac{1}{\delta}\right)}{n}} + \mathrm{O}\left(\frac{\log\left(\frac{1}{\delta}\right)}{n}\right) \tag{II.9}$$

where the $\mathrm{O}(\frac{1}{n})$ term is due to the higher order moments of $(p(z)-\hat{p}_n(z))u$. This is by the way the idea behind Bernstein's inequality. However, (II.9) is achievable only if u is fixed. Here, u models $\tilde{h}/\|\tilde{h}\|_1$ which is unknown in advance, so a union bound is performed for all possible values of u, and the question becomes:

$$\textit{What is the best possible bound for } U_n := \max_{u:\|u\|_1=1}(p(z)-\hat{p}_n(z))u\textit{?}$$

When we pick $p(z) = (\frac{1}{S},\dots,\frac{1}{S})$ as the uniform distribution, we have $U_n = \frac{1}{2}\sum_{s\in\mathscr{S}}|\hat{p}_n(s|z)-\frac{1}{S}|$ so $\mathbf{E}[U_n] = \frac{S}{2}\mathbf{E}|\hat{p}_n(s|z)-\frac{1}{S}|$ and it is an amusing exercise to show that $\mathbf{E}[U_n] \gtrsim \frac{1}{2}\sqrt{\pi n S}$ when $n \to \infty$. Therefore, one expects that an upper bound in the style of (II.9) that holds simultaneously for all $u \in \mathbf{R}^{\mathscr{S}}$ satisfying $\|u\|_1 = 1$ to grow as $\Theta(\sqrt{S\log(1/\delta)/n})$. This is the case for Weissman's inequality Weissman et al. (2003) for instance.

The take-home idea is the following.

> **Important idea.** If one wants to bound the error on the co-vector $\hat{p}_n(z)-p(z)$ in every direction, the bound should scale as $\sqrt{S\log(1/\delta)/n}$. Because in finite dimension, co-vectors and vectors are naturally identifiable, it means that any kernel confidence region must scale with $\sqrt{S}$.[a] And indeed, all the sub-optimal methods ranging from UCRL2 to UCRL3 in Table 2.1 suffer from this extra $\sqrt{S}$: Because if the kernel confidence region does not contain $\sqrt{S}$ by design, then it is too narrow.
>
> [a] This is the mistake made in the design of UCRL-V Tossou et al. (2019) that tries to escape the $\sqrt{S}$-factor by using sub-modularity; There kernel confidence region ends up being too narrow.

6.4.2 Shaving one $\sqrt{S}$ by moving beyond EVI

The previous paragraph motivates the idea that EVI-based algorithm (Algorithm II.2) cannot get rid of the $\sqrt{S}$-factor unless their kernel confidence region is over-aggressive. This means that something new is needed. Thankfully, there already exists an algorithm achieving minimax optimality: EBF Zhang and Ji (2019). In this work, the authors identify a collection of properties that, if always satisfies by the optimistic policy, guarantee minimax optimal regret. Although this work is short of providing a way to find such a policy in the ocean of all existing policies, or even of providing a way to verify that a policy satisfies all the requirements, they initiate an important idea: Only the deviations of $\hat{p}_n(z) - p(z)$ at h^* matter, and h^* can be estimated with an external subroutine. So, if h^* is estimated by some $u \in \mathbf{R}^{\mathcal{S}}$, we morally want to stick to optimistic kernels $\tilde{p}(z)$ such that $(\tilde{p}(z) - \hat{p}_n(z))u$ satisfies an inequality in the style of (II.9) without that additional $\sqrt{S}$. This idea will lead to the **mitigation** operation of PMEVI.

This will be better discussed in Chapter 7.

6.4.3 Changing D to $\mathrm{sp}(h^*)$

A few works Bartlett and Tewari (2009); Fruit et al. (2018); Zhang and Ji (2019) manage to change D to $\mathrm{sp}(h^*)$ in their regret guarantees, but all these methods work with prior information on the bias function, e.g., of the form "$\mathrm{sp}(h^*) \leq c$". Also, only SCAL Fruit et al. (2018) provides a tractable way of using that prior information. The idea developed in this work is to make sure, at every step of EVI (Algorithm II.3), that the ending vector u_n satisfies $\mathrm{sp}(u_n) \leq c$ via an operation that they call truncation, but that we will rather refer to as **projection** because it corresponds to projecting u_n onto a convex set. Fruit et al. (2018) make clear that the projection operation Γ must satisfy a few algebraic properties: monotony, non span-expansiveness and linearity (see Proposition I.7). These properties guarantee that the result of the modified EVI algorithm corresponds to an optimistic policy with optimistic bias span at most c. Directly injecting into (II.8), this directly changes $D(M)$ and improve the regret bound.

In his manuscript, Fruit (2019) provides an improved version of SCAL, SCAL*, together with a precious intuition: What matters is not really to be optimistic about the value of policies, but rather that EVI works with an **optimistic Bellman operator** $\mathscr{L}$. With PMEVI, we generalize the projection operation of Fruit et al. (2018) and go further with the idea of optimistic Bellman operators.

Chapter 7

Projected Mitigated Extended Value Iteration (PMEVI)

In the previous chapter, we have argued that EVI (Algorithm II.3) and policy-wise optimism lack something to achieve minimax optimal regret, because the design of the kernel confidence region intrinsically produces an extra $\sqrt{S}$. With PMEVI, the sub-routine EVI is indeed improved via a combination of two operations, but this goes even further. It is rather surprising that, in opposition to this previous line of work, the theoretical analysis of PMEVI suggests that the choice of the confidence region $\mathscr{M}_t$ has little importance. In fact, EVI takes an arbitrary (well-behaved) confidence region in, infer bias information similarly to EBF Zhang and Ji (2019) and makes use of it to heavily refine the extended Bellman operator (II.4) associated to the input confidence region, and this confidence region can almost be arbitrary (provided that it is correct) **especially on kernels**: PMEVI achieves minimax optimal regret guarantees even with $\mathscr{P}_{\mathbb{z}}(t) = \mathscr{P}(\mathscr{S})$. The algorithm PMEVI can further take arbitrary prior information $\mathscr{H}_*$ (possibly none, i.e., $\mathscr{H}_* = \mathbf{R}^{\mathscr{S}}$) on the bias vector to tighten its bias confidence region. The pseudo-code given in Algorithm II.5 is the high level structure of the algorithm PMEVI-DT. In Section 7.1, we explain how EVI is refined using bias information and in Section 7.1.1, we explain how bias information is obtained.

Algorithm II.4 PMEVI-DT$(\mathscr{H}_*, T, t \mapsto \mathscr{M}_t)$

Parameters: Bias prior $\mathscr{H}_*$, horizon T, a system of confidence region $t \mapsto \mathscr{M}_t$

for $k = 1, 2, \dots$ **do**
  Set $t_k \leftarrow t$, update confidence region $\mathscr{M}_t$;
  $\mathscr{H}_t' \leftarrow$ BiasEstimation$(\mathscr{O}_t, \mathscr{M}_t, \delta)$:
  $\mathscr{H}_t \leftarrow \mathscr{H}_* \cap \{u : \mathrm{sp}(u) \le T^{1/5}\} \cap \mathscr{H}_t'$;
  $\Gamma_t \leftarrow$ BiasProjection$(\mathscr{H}_t, -)$;
  $\beta_t \leftarrow$ VarianceApprox$(\mathscr{H}_t', \mathscr{O}_t)$;
  $\mathfrak{h}_k \leftarrow$ PMEVI$(\mathscr{M}_t, \beta_t, \Gamma_t, \sqrt{\log(t)/t})$;
  $\mathfrak{g}_k \leftarrow \mathfrak{L}_t \mathfrak{h}_k - \mathfrak{h}_k$;
  Update policy $\pi_k \leftarrow$ Greedy$(\mathscr{M}_t, \mathfrak{h}_k, \beta_t)$;
  repeat
    Play $A_t \leftarrow \pi_k(S_t)$, observe R_t, S_{t+1};
    Increment $t \leftarrow t + 1$;
  until (DT) $N_t(S_t, \pi_k(S_t)) \ge 1 \vee 2N_{t_k}(Z_t)$.
end for

Algorithm II.5 PMEVI$(\mathscr{M}, \beta, \Gamma, \epsilon)$

Parameters: region $\mathscr{M}$, mitigation β, projection Γ, precision ϵ, initial vector v_0 (optional)

if v_0 not initialized **then** set $v_0 \leftarrow 0$;
$n \leftarrow 0$
$\mathscr{L} \leftarrow$ extended operator associated to $\mathscr{M}$;
repeat
  $v_{n+\frac{1}{2}} \leftarrow \mathscr{L}^{\beta} v_n$;
  $v_{n+1} \leftarrow \Gamma v_{n+\frac{1}{2}}$;
  $n \leftarrow n + 1$;
until $\mathrm{sp}(v_n - v_{n-1}) < \epsilon$
return v_n.

7.1 Projected mitigated extended value iteration (PMEVI)

Assume that an external mechanism provides a confidence region $\mathscr{H}_t$ for the bias function h^*. Provided that $\mathscr{M}_t$ is correct ($M \in \mathscr{M}_t$) and that $\mathscr{H}_t$ is correct ($h^* \in \mathscr{H}_t$), we want to find a pair of policy-model $(\pi, \tilde{M})$ that maximize the gain and such that $h^\pi(\tilde{M}) \in \mathscr{H}_t$. This is done with an improved version of (II.4) combining two ideas.

1. **Projection (Section 7.1.1).** Whenever it is correct, the bias confidence region $\mathscr{H}_t$ informs the learner that the search of an optimistic model can be constrained to those with bias within $\mathscr{H}_t$. This is done by projecting $\mathscr{L}_t^\beta$ (see **mitigation**) using an operator $\Gamma_t : \mathbf{R}^{\mathscr{S}} \to \mathscr{H}_t$, that has to satisfy a few non-trivial regularity conditions that are specified in Proposition II.13.

2. **Mitigation (Section 7.1.2).** When one is aware that $h^* \in \mathscr{H}_t$, the **dynamical bias update** $\tilde{p}(s,a)u_i$ in (II.4) can be better controlled, by trying to restrict (II.4) to some $\tilde{p}(s,a)$ such that $\tilde{p}(s,a)u_i \le \hat{p}_t(s,a)u_i + (p(s,a) - \hat{p}_t(s,a))u_i$ with the knowledge that $u_i \in \mathscr{H}_t$.

   For a fixed $u \in \mathbf{R}^{\mathscr{S}}$, the empirical Bernstein inequality (Lemma I.26) provides a variance bound of the form $(\hat{p}_t(s,a) - p(s,a))u \le \beta_t(s,a,u)$. By computing $\beta_t(s,a) := \max_{u \in \mathscr{H}_t} \beta_t(s,a,u)$, the search makes sure that $(\hat{p}_t(s,a) - p(s,a))h^* \le \beta_t(s,a)$ even though h^* is unknown. For $\beta \in \mathbf{R}_+^{\mathscr{Z}}$, we introduce the β**-mitigated** extended Bellman operator:

$$\mathscr{L}_t^\beta u(s) := \max_{a \in \mathscr{A}(s)} \sup_{\tilde{r}(s,a) \in \mathscr{R}_t(s,a)} \sup_{\tilde{p}(s,a) \in \mathscr{P}_t(s,a)} \left\{ \tilde{r}(s,a) + \min\{\tilde{p}(s,a)u_i, \hat{p}_t(s,a)u_i + \beta_t(s,a)\} \right\} \tag{II.1}$$

The proposition below shows how well-behaved the composition $\mathfrak{L}_t := \Gamma_t \circ \mathscr{L}_t^\beta$ is. Its proof requires to build a complete analysis of projected mitigated Bellman operators. This is deferred to the appendix.

Proposition II.13. *Fix $\beta \in \mathbf{R}_+^{\mathscr{Z}}$ and assume that there exists a projection operator $\Gamma_t : \mathbf{R}^{\mathscr{Z}} \to \mathscr{H}_t$ which is **(O1)** monotone: $u \le v \Rightarrow \Gamma u \le \Gamma v$; **(O2)** non span-expansive: $\mathrm{sp}(\Gamma u - \Gamma v) \le \mathrm{sp}(u - v)$; **(O3)** linear: $\Gamma(u + \lambda e) = \Gamma u + \lambda e$ and **(O4)** $\Gamma u \le u$. Then, the **projected mitigated extended Bellman operator** $\mathfrak{L}_t := \Gamma_t \circ \mathscr{L}_t^\beta$ has the following properties:*

(1) *There exists a unique $\mathfrak{g}_t \in \mathbf{R}e$ such that $\exists \mathfrak{h}_t \in \mathscr{H}_t, \mathfrak{L}_t \mathfrak{h}_t = \mathfrak{h}_t + \mathfrak{g}_t$;*
(2) *If $M \in \mathscr{M}_t$, $h^* \in \mathscr{H}_t$ and $(\hat{p}_t(s,a) - p(s,a))h^* \le \beta_t(s,a)$, then $\mathfrak{g}_t \ge g^*(M)$;*
(3) *If $\mathscr{M}_t$ is convex, then for all $u \in \mathbf{R}^{\mathscr{S}}$, the policy π picking the actions achieving $\mathscr{L}_t^\beta u$ satisfies $\mathfrak{L}_t u = \tilde{r}^\pi + \tilde{P}^\pi u$ for $\tilde{r}^\pi(s) \le \sup \mathscr{R}_t(s, \pi(s))$ and $\tilde{P}^\pi(s) \in \mathscr{P}_t(s, \pi(s))$;*
(4) *For all $u \in \mathbf{R}^{\mathscr{S}}$ and $n \ge 0$, $\mathrm{sp}((\mathfrak{L}_t)^{n+1}u - (\mathfrak{L}_t)^n u) \le \mathrm{sp}((\mathscr{L}_t)^{n+1}u - (\mathscr{L}_t)^n u)$.*

The property (1) guarantees that $\mathfrak{L}_t$ has a fix-point while (2) states that this fix-point corresponds to an optimistic gain $\mathfrak{g}_t$ if the model and the bias confidence region are correct and the mitigation isn't too aggressive. Combined with (3), the Poisson equation of a policy corresponds to this fix-point, i.e., $\tilde{r}^\pi + \tilde{P}^\pi \mathfrak{h}_t = \mathfrak{h}_t + \mathfrak{g}_t$, so that $\mathfrak{g}_t$ is the gain and $\mathfrak{h}_t \in \mathscr{H}_t$ is a legal bias for π under the model $(\tilde{r}^\pi, \tilde{P}^\pi)$. Lastly, the property (4) guarantees that the iterates $\mathfrak{L}_t^n u$ converge to a fix-point of $\mathfrak{L}$ at least as quickly as $\mathscr{L}_t^n u$ goes to a fix-point of $\mathscr{L}_t$; the convergence of $\mathscr{L}_t^n u$ is already guaranteed by existing studies and is discussed in the appendix.

Provided that the bias confidence region is constructed, Proposition II.13 foreshadows how powerful is the construction: The algorithm PMEVI, obtained by iterating $\mathfrak{L}_t$ instead of $\mathscr{L}_t$ in EVI, can replace the well-known EVI within any algorithm of the literature that relies on it (UCRL2 Auer et al. (2009), UCRL2B Fruit et al. (2020) or KL-UCRL Filippi et al. (2010)) for an immediate improvement of its theoretical guarantees.

7.1.1 Building the bias confidence region and its projection operator

The bias confidence region used by PMEVI-DT is obtained as a collection of constraints of the form:

$$\forall s \neq s', \quad \mathfrak{h}(s) - \mathfrak{h}(s') - c(s,s') \leq d(s,s'). \tag{II.2}$$

Such constraints include (1) prior bias constraints (if any) of the form of $\mathfrak{h}(s) - \mathfrak{h}(s') \leq c_*(s,s')$; (2) span constraints of the form $\mathfrak{h}(s) - \mathfrak{h}(s') \leq c_0 := T^{1/5}$ spawning the span semi-ball $\{u : \mathrm{sp}(u) \leq T^{1/5}\}$; and (3) pair-wise constraints obtained by estimating bias differences in the style of Zhang and Ji (2019); Zhang and Xie (2023) that we further improve. We start by defining a bias difference estimator.

Definition II.5 (Bias difference estimator)**.** *Given a pair of states $s \neq s'$, their sequence of **commute times** $(\tau_i^{s\leftrightarrow s'})_{i\geq 0}$ is defined by $\tau_{2i}^{s\leftrightarrow s'} := \inf\{t > \tau_{2i-1}^{s\leftrightarrow s'} : S_t = s\}$ and $\tau_{2i+1}^{s\leftrightarrow s'} := \inf\{t > \tau_{2i}^{s\leftrightarrow s'} : S_t = s'\}$ with the convention that $\tau_{-1}^{s\leftrightarrow s'} = -\infty$. The number of commutations up to time t is $N_t(s \leftrightarrow s') := \inf\{i : \tau_i^{s\leftrightarrow s'} \leq t\}$, and $\hat{g}(t) := \frac{1}{t}\sum_{i=0}^{t-1} R_i$ is the empirical gain. The **bias difference estimator** at time T is any quantity $c_T(s,s') \in \mathbf{R}$ such that:*

$$N_t(s \leftrightarrow s') c_T(s,s') = \sum_{t=0}^{N_T(s\leftrightarrow s')-1} (-1)^i \sum_{t=\tau_i^{s\leftrightarrow s'}}^{\tau_{i+1}^{s\leftrightarrow s'}-1} (\hat{g}(T) - R_t). \tag{II.3}$$

Lemma II.14. *With probability $1-2\delta$, for all $T' \leq T$ and all $\tilde{g} \geq g^*$, the quantity $(*) := N_{T'}(s \leftrightarrow s')|h^*(s) - h^*(s') - c_{T'}(s,s')|$ satisfies*

$$(*) \leq 3\mathrm{sp}(h^*) + (1 + \mathrm{sp}(h^*))\sqrt{8T\log(\tfrac{2}{\delta})} + 2\sum_{t=0}^{T'-1}(\tilde{g} - R_t). \tag{II.4}$$

Lemma II.14 says that the quality of the estimator $c_T(s,s')$ is directly linked to the number of observed commutes between s and s' as well as the regret. The idea is that if the algorithm makes many commutes between s and s' and if its regret is small, then the algorithm mostly takes optimal paths from s to s'. The bound provided by Lemma II.14 is not accessible to the learner however, because $\mathrm{sp}(h^*)$ is unknown in general. To overcome this issue, $\mathrm{sp}(h^*)$ is upper-bounded by $c_0 := T^{1/5}$. Overall, this leads to the design of the algorithm estimating the bias confidence region as specified in Algorithm II.6.

Algorithm II.6 `BiasEstimation`$(\mathscr{O}_t, \mathscr{M}_t, \delta)$

Parameters: History $\mathscr{O}_t$, model region $\mathscr{M}_t$, confidence $\delta > 0$

1: Estimate bias differences c_t via (II.3);
2: Estimate optimistic gain $\tilde{g} \leftarrow \min_{k<K(t)} \mathfrak{g}_k$;
3: Inner regret estimation $B_0 \leftarrow t\tilde{g} - \sum_{i=0}^{t-1} R_i$;
4: $\ell \leftarrow \sqrt{8T\log(\frac{2}{\delta})}$, $c_0 \leftarrow T^{\frac{1}{5}}$;
5: Estimate the bias difference errors as:

$$d_t(s,s') \equiv \mathrm{error}(c_t,s,s') := \frac{3c_0 + (1+c_0)(1+\ell) + 2B_0}{N_t(s \leftrightarrow s')}$$

6: **return** $(c_t, \mathrm{error}(c_t,-,-))$, (II.2) defines $\mathscr{H}'_t$.

Algorithm II.7 `BiasProjection`$(\mathscr{H}_t, u)$

Parameters: $\mathscr{H}_t$ a collection of linear constraints (II.2), $u \in \mathbf{R}^{\mathscr{S}}$ to project

1: $v \leftarrow 0^{\mathscr{S}}$;
2: **for** $s \in \mathscr{S}$ **do**
3: Using linear programming, compute:
4: $v(s) \leftarrow \sup\{w(s) : w \leq u \text{ and } w \in \mathscr{H}_t\}$;
5: **end for**
6: **return** v.

Coupled with prior information and span constraints, the obtained bias confidence region $\mathscr{H}_t$ is a polyhedron of the same kind as the one encountered in Zhang and Xie (2023) generated by constraints of the form (II.2), and similarly to their Proposition 3, one can project onto $\mathscr{H}_t$ in polynomial time with Algorithm II.7. Moreover, the resulting projection operator satisfies the prerequisites (**O1-4**) of Proposition II.13, making sure that PMEVI (Algorithm II.5) is well-behaved. This is proved in the appendix Section 7.B.2.

Lemma II.15. *Assume that $\mathscr{H}$ is a set of $\mathfrak{h} \in \mathbf{R}^{\mathscr{S}}$ satisfying a system of equations of the form of (II.2). If $\mathscr{H}$ is non empty, then the operator $\Gamma u :=$ BiasProjection$(\mathscr{H}, u)$ (see Algorithm II.7) is a projection on $\mathscr{H}$ and satisfies the properties (**O1-4**) defined in Proposition II.13.*

7.1.2 Mitigation using finer bias dynamical error

The fact that $h^* \in \mathscr{H}_t$ with high probability is used in PMEVI-DT to restrict the search of EVI by reducing the dynamical bias error. This reduction is based on a empirical Bernstein inequality (see Lemma I.26) applied to $(\hat{p}(s,a) - p(s,a))u$. Here, it gives that with probability $1-\delta$, we have:

$$(\hat{p}_t(s,a) - p(s,a))u \le \sqrt{\frac{2\mathbf{V}(\hat{p}_t(s,a),u)\log\left(\frac{3T}{\delta}\right)}{\max\{1,N_t(s,a)\}}} + \frac{3\mathrm{sp}(u)\log\left(\frac{3T}{\delta}\right)}{\max\{1,N_t(s,a)\}} =: \beta_t(s,a,u) \tag{II.5}$$

where $\mathbf{V}(\hat{p}_t(s,a),u)$ is the variance of u under the probability vector $\hat{p}_t(s,a)$. More specifically, if q is a probability on $\mathscr{S}$ and $q \in \mathbf{R}^{\mathscr{S}}$, we set $\mathbf{V}(q,u) := \sum_s q(s)(u(s) - q \cdot u)^2$. In (II.5), $u \in \mathbf{R}^{\mathscr{S}}$, $(s,a) \in \mathscr{Z}$ and $T \ge 1$ are fixed. One is tempted to use (II.5) directly to mitigate the extended Bellman operator, but the resulting operator is ill-behaved because it looses monotony. This issue is avoided by changing $\beta_t(s,a,u)$ to $\max_{u \in \mathscr{H}_t} \beta_t(s,a,u)$ in (II.3). We obtain a variance maximization problem, which is a **convex maximization problem** with linear constraints. Even in very simple settings, such optimization problems are NP-hard Pardalos and Schnitger (1988) hence computing $\max_{u \in \mathscr{H}_t} \beta_t(s,a,u)$ is not reasonable in general. Thankfully, this value can be upper-bounded by a tractable quantity that is enough to guarantee regret efficiency. The mitigation β_t used by PMEVI-DT is provided with Algorithm II.8.

Algorithm II.8 VarianceApproximation$(\mathscr{H}'_t, \mathscr{O}_t)$

Parameters: Bias region $\mathscr{H}'_t$, history $\mathscr{O}_t$

Extract constraints $(c, \mathrm{error}(c,-,-)) \leftarrow \mathscr{H}'_t$;
Set $c_0 \leftarrow T^{\frac{1}{5}}$;
Pick a reference point $h_0 \leftarrow$ BiasProjection$(\mathscr{H}_t, c(-,s_0))$;
for $(s,a) \in \mathscr{Z}$ **do**
  $\rho \leftarrow \log\left(\frac{SAT}{\delta}\right) / \max\{1, N_t(s,a)\}$;
  $\mathrm{var}(s,a) \leftarrow \mathbf{V}(\hat{p}_t(s,a), h_0) + 8c_0 \sum_{s' \in \mathscr{S}} \hat{p}_t(s'|s,a) c(s',s)$;
  $\beta_t(s,a) \leftarrow \sqrt{2\mathrm{var}(s,a)\rho} + 3c_0\rho$ or $+\infty$ if $N_t(s,a) = 0$;
end for
return β_t.

7.2 Elements of regret analysis of PMEVI

7.2.1 Regret guarantees of PMEVI

Theorem II.16 below shows that PMEVI has minimax optimal regret under regularity assumptions on the used confidence region $\mathscr{M}_t$. Assumption 1 asserts that the confidence region holds uniformly with high probability. Assumption 2 asserts that the reward confidence region is sub-Weissman (see Lemma I.23) and Assumption 3 assumes that the model confidence region makes sure that EVI (II.4) converges in the first place. Assumption 4 asserts that the prior bias region is correct.

Assumption 1. *With probability $1-\delta$, we have $M \in \bigcap_{k=1}^{K(T)} \mathscr{M}_{t_k}$.*

Assumption 2. *There exists a constant $C > 0$ such that for all $(s,a) \in \mathscr{S}$, for all $t \le T$, we have:*

$$\mathscr{R}_t(s,a) \subseteq \{\tilde{r}(s,a) \in \mathscr{R}(s,a) : N_t(s,a)\|\hat{r}_t(s,a) - \tilde{r}(s,a)\|_1^2 \le C \log\Big(\tfrac{2SA(1+N_t(s,a))}{\delta}\Big)\}.$$

Assumption 3. *For $t \ge 0$, $\mathscr{M}_t$ is a convex region in product form and $\mathscr{L}_t^n u$ converges a fix-point.*

Assumption 4. *The prior bias region $\mathscr{H}_*$ contains $h^*(M)$ and is generated by constraints of the form:*

$$\forall s \neq s', \quad \mathfrak{h}(s) - \mathfrak{h}(s') \le c_*(s,s')$$

with $c_*(s,s') \in [-\infty, \infty]$ (possibly infinite).

Refer to Section 7.A.2 for the feasibility of Assumption 1, Section 7.A.2.3 for Assumption 2, and Section 7.A.3 for Assumption 3.

Theorem II.16 (Main result)**.** *Let $c > 0$. Assume that* PMEVI-DT *runs with a confidence region system $t \mapsto \mathscr{M}_t$ that guarantees Assumptions 1-3. If $T \ge c^5$, then for every weakly communicating model with $\mathrm{sp}(h^*) \le c$ and such that Assumption 4 is satisfied ($h^* \in \mathscr{H}_*$),* PMEVI-DT *achieves regret:*

$$\mathrm{O}\Big(\sqrt{cSAT \log\big(\tfrac{SAT}{\delta}\big)}\Big) + \mathrm{O}\Big(cS^{\frac{5}{2}}A^{\frac{3}{2}}T^{\frac{9}{20}} \log^2\big(\tfrac{SAT}{\delta}\big)\Big)$$

with probability $1-26\delta$, and in expectation if $\delta < \sqrt{1/T}$. Moreover, if PMEVI-DT *runs with the same confidence regions that* UCRL2 *Auer et al. (2009), then it enjoys a time complexity $\mathrm{O}(DS^3AT)$.*

To have a completely prior-less algorithm, pick $\mathscr{H}_* = \mathbf{R}^{\mathscr{S}}$. The proof of Theorem II.16 is tedious and is deferred to the appendix. We will focus here on the main ideas.

7.2.2 Main line of the regret analysis of PMEVI

We start by recalling a few notations. At episode k, the played policy is denoted π_k. As a greedy response to $\mathfrak{h}_k$, by Proposition II.13 (3), there exists $\tilde{r}_k(s) \le \sup \mathscr{R}_{t_k}(s, \pi_k(s))$ and $\tilde{P}_k(s) \in \mathscr{P}_{t_k}(s, \pi(x))$ such that $\mathfrak{h}_k + \mathfrak{g}_k = \tilde{r}_k + \tilde{P}_k \mathfrak{h}_k$. The reward-kernel pair $\tilde{M}_k = (\tilde{r}_k, \tilde{P}_k)$ is referred to as the **optimistic model** of π_k. We write $P_k := P_{\pi_k}(M)$ the true kernel and $\hat{P}_k := P_{\pi_k}(\hat{M}_{t_k})$ the empirical kernel. Likewise, we define the reward functions r_k and $\hat{r}_k$. The optimistic gain and bias satisfy $\mathfrak{g}_k = g(\pi_k, \widetilde{M}_k)$ and $\mathfrak{h}_k = h(\pi_k, \widetilde{M}_k)$. We further denote $c_0 = T^{\frac{1}{5}}$.

The regret is first decomposed episodically with $\mathrm{Reg}(T) = \sum_k \sum_{t=t_k}^{t_{k+1}-1} (g^* - R_t)$. The first step goes back to the analysis of UCRL2 Auer et al. (2009), and consists in upper-bounding the regret over episode k with optimistic quantities that are exclusive to that episode.

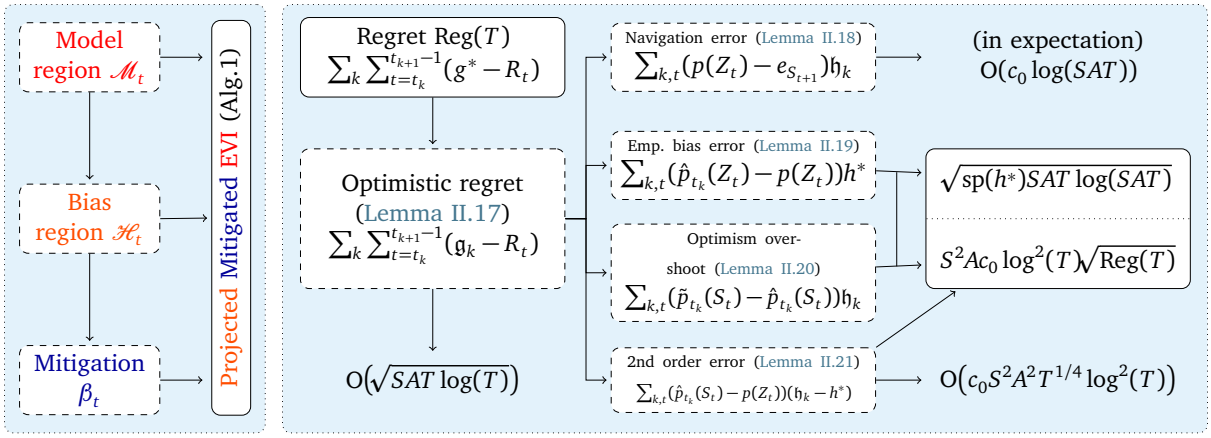


Figure 7.1: An overview of PMEVI-DT and its regret analysis. In the above, $\mathfrak{g}_k$ and $\mathfrak{h}_k$ are the optimistic gain and bias functions produced by PMEVI (see Algorithm 2) at episode k, and $\hat{p}_{t_k}$ and $\tilde{p}_{t_k}$ are respectively the empirical and optimistic kernel models at episode k.

Lemma II.17 (Reward optimism). *With probability $1-6\delta$, we have:*

$$\mathrm{Reg}(T) \le \sum_k \sum_{t=t_k}^{t_{k+1}-1} (\mathfrak{g}_k - R_t) \le \sum_k \sum_{t=t_k}^{t_{k+1}-1} (\mathfrak{g}_k - \tilde{r}_k(Z_t)) + O\Big(\sqrt{SAT\log\big(\tfrac{T}{\delta}\big)}\Big). \tag{II.6}$$

We introduce the optimistic regrets $B(T) := \sum_k \sum_{t=t_k}^{t_{k+1}-1}(\mathfrak{g}_k - R_t)$ and $\tilde{B}(T) := \sum_k \sum_{t=t_k}^{t_{k+1}-1}(\mathfrak{g}_k - \tilde{r}_k(Z_t))$. Rewriting the summand $\mathfrak{g}_k - \tilde{r}_k(Z_t)$ using the Poisson equation $\mathfrak{h}_k + \mathfrak{g}_k = \tilde{r}_k + \tilde{P}_k\mathfrak{h}_k$, we get:

$$\tilde{B}(T) = \sum_k \sum_{t=t_k}^{t_{k+1}-1} \big(\tilde{p}_k(S_t) - e_{S_t}\big)\mathfrak{h}_k.$$

The analysis proceed by decomposing the above expression of $\tilde{B}(T)$ in the style of Zhang and Ji (2019). We write $\sum_{t=t_k}^{t_{k+1}-1}(\tilde{p}_k(S_t) - e_{S_t})\mathfrak{h}_k$ as:

$$\sum_{t=t_k}^{t_{k+1}-1} \Bigg(\underbrace{\big(p_k(S_t) - e_{S_t}\big)\mathfrak{h}_k}_{\text{navigation error } (1k)} + \underbrace{(\hat{p}_k(S_t) - p_k(S_t))h^*}_{\text{empirical bias error } (2k)} + \underbrace{(\tilde{p}_k(S_t) - \hat{p}_k(S_t))\mathfrak{h}_k}_{\text{optimistic overshoot } (3k)} + \underbrace{(\hat{p}_k(S_t) - p_k(S_t))(\mathfrak{h}_k - h^*)}_{\text{second order error } (4k)} \Bigg)$$

Each error term is bounded separately. Below, we denote $\mathbf{V}(q,u) := \sum_s q(s)(u(s) - q\cdot u)^2$.

Lemma II.18 (Navigation error). *With probability $1-7\delta$, the navigation error is bounded by:*

$$\sum_k \sum_{t=t_k}^{t_{k+1}-1} (p_k(S_t) - e_{S_t})\mathfrak{h}_k \le \sqrt{2\sum_{t=0}^{T-1} \mathbf{V}(p(Z_t), h^*)\log\big(\tfrac{T}{\delta}\big)} + 2SA^{\frac{1}{2}}\sqrt{3B(T)}\log\big(\tfrac{T}{\delta}\big) + \widetilde{O}\big(T^{\frac{7}{20}}\big).$$

Lemma II.19 (Empirical bias error). *With probability $1-\delta$, the empirical bias error is bounded by:*

$$\sum_k \sum_{t=t_k}^{t_{k+1}-1} (\hat{p}_k(S_t) - p_k(S_t))h^* \le 4\sqrt{SA\sum_{t=0}^{T-1} \mathbf{V}(p(Z_t), h^*)\log\big(\tfrac{SAT}{\delta}\big)} + O\big(\log^2(T)\big).$$

Lemma II.20 (Optimism overshoot). *With probability* $1-6\delta$*, the optimism overshoot is bounded by:*

$$\sum_k \sum_{t=t_k}^{t_{k+1}-1} (\tilde{p}_k(S_t)-\hat{p}_k(S_t))\mathfrak{h}_k \leq \left\{ \begin{matrix} 4\sqrt{2SA\sum_{t=0}^{T-1}\mathbf{V}(p(Z_t),h^*)\log\left(\frac{SAT}{\delta}\right)} \\ +8(1+c_0)S^{\frac{3}{2}}A\log^{\frac{3}{2}}\left(\frac{SAT}{\delta}\right)\sqrt{B(T)}+\widetilde{\mathrm{O}}\left(T^{\frac{1}{4}}\right) \end{matrix} \right\}.$$

Lemma II.21 (Second order error). *With probability* $1-6\delta$*, the second order error is bounded by:*

$$\sum_k \sum_{t=t_k}^{t_{k+1}-1} (\hat{p}_k(S_t)-p_k(S_t))(\mathfrak{h}_k-h^*) \leq 16S^2A(1+c_0)\log^{\frac{1}{2}}\left(\frac{S^2AT}{\delta}\right)\sqrt{2B(T)}+\widetilde{\mathrm{O}}\left(T^{\frac{1}{4}}\right).$$

We see that the empirical bias error (Lemma II.19) and the optimism overshoot (Lemma II.20) both involve the sum of variances $\sum_{t=0}^{T-1}\mathbf{V}(p(Z_t),h^*)$, which is shown in Lemma II.40 to be of order $\mathrm{sp}(h^*)\mathrm{sp}(r)T+\sum_{t=0}^{T-1}\Delta^*(Z_t)$. The pseudo-regret term $\sum_{t=0}^{T-1}\Delta^*(Z_t)$ is bounded with the regret using Corollary II.42, then by $B(T)$. With high probability, we obtain an equation of the form:

$$B(T) \leq C\sqrt{(1+\mathrm{sp}(h^*))SAT\log\left(\tfrac{T}{\delta}\right)}+CS^2A(1+c_0)\log^2(T)\sqrt{B(T)}+\tilde{\mathrm{O}}\left(T^{\frac{1}{4}}\right)$$

where C is a constant. Setting $\alpha := CS^2A(1+c_0)\log^2(T)$ and $\beta := C\sqrt{(1+\mathrm{sp}(h^*))SAT\log(T/\delta)}+\tilde{\mathrm{O}}(T^{1/4})$, the above equation is of the form $B(T)\leq\beta+\alpha\sqrt{B(T)}$. Solving in $B(T)$, we find $B(T)\leq\beta+2\sqrt{\alpha\beta}+\alpha^2$. The dominant term is β, hence we readily obtain:

$$B(T) \leq C\sqrt{(1+\mathrm{sp}(h^*))\mathrm{sp}(r)SAT\log\left(\tfrac{T}{\delta}\right)}+\widetilde{\mathrm{O}}\left(\mathrm{sp}(h^*)\mathrm{sp}(r)S^{\frac{5}{2}}A^{\frac{3}{2}}(1+c_0)T^{\frac{1}{4}}\right). \tag{II.7}$$

Since $c_0=\mathrm{o}(T^{\frac{1}{4}})$, we conclude that $B(T)=\mathrm{O}\left(\sqrt{\mathrm{sp}(h^*)SAT\log(T/\delta)}\right)$, ending the proof.

> **Important remark.** The result of Theorem II.16 is in probability. A result in expectation could be obtained by setting a time-adaptive confidence level $\delta\equiv\delta(t):=\frac{1}{t}$.

7.3 Experimental illustrations

To get a grasp of how PMEVI-DT behaves in practice, we provide in Figure 7.2 of few illustrative experiments. In both experiments, the environment is a river-swim which is a model known to be hard to learn despite its size, with high diameter and bias span. Its description is found in Bourel et al. (2020) and is reported in the appendix for self-containedness.

We observe on the first experiment that PMEVI behaves almost identically to its EVI counterparts when no prior on the bias region is given. This is because most of the regret is due to the earlier learning phase, when bias information is impossible to get (the regret is still growing linearly and the bias estimator is off). Accordingly, the bias confidence region is too large and all projections onto it are trivial during the iterations of PMEVI. Thankfully, this also makes the calls to PMEVI not substantially heavier than calls to EVI from a computational perspective. On the second experiment, we measure the influence of prior bias information on the behavior of PMEVI-DT. We observe that PMEVI-DT is very efficient at using adequate bias prior information to strikingly reduce the burn-in cost of the learning process on this 3-state riverswim.

7.4 Future directions

Figure 7.2 is two-sided. On the one hand, the left-side experiments indicates that PMEVI does not improve the regret in practice. I only had the time to get a superficial understanding of

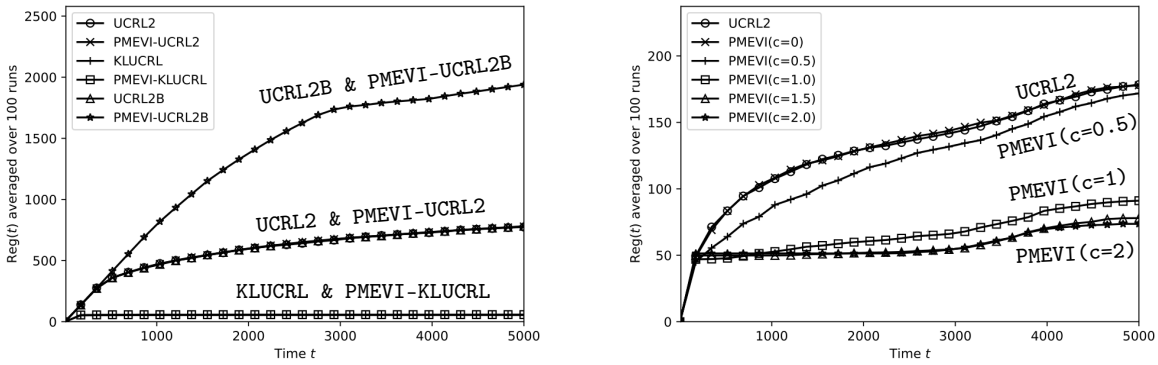


Figure 7.2: (**To the left**) Running a few algorithms of the literature on 5-state river-swim and comparing their average regret against their PMEVI variants, obtained by changing calls to the EVI sub-routine to calls to PMEVI. (**To the right**) Running UCRL2 and PMEVI-DT with the same confidence region that UCRL2 on a 3-state river-swim. PMEVI-DT is run with prior knowledge $h^*(s_1) \leq h^*(s_2) - c \leq h^*(s_3) - 2c$ for $c \in \{0, 0.5, 1, 1.5, 2\}$.

the phenomenon. On the left part of Figure 7.2, we indeed observe that there is hardly any difference between the expected regret of EVI and PMEVI-based algorithms, while PMEVI have better minimax regret guarantees. I don't think that the lack of difference is completely explained by the large second order term in the regret analysis (see Theorem II.16). Figure 7.2 displays **model-dependent** regret curves because the regret is averaged over multiple run on the same environment, while the regret analysis is **minimax**. To this end, Figure 7.2 should not be seen as any attempt at validating or invalidating the theoretical results of Theorem II.16 and to some extent, the two are perhaps incomparable. Figure 7.2 should be seen as experimental insights; nothing less, nothing more.

One may argue that a model-dependent regret analysis of PMEVI would be more appropriate – There I would like to put disclaimers. Regarding the lower bound of Part III, I believe that policy-based optimism is not suited to reach model-dependent optimal regret, because any such optimal algorithm should see exploration as a form of optimization problem under information constraints. This optimization aspect is absent from EVI and PMEVI-based algorithms alike; Such methods are suited to robust regret guarantees but not for asymptotically optimal regret dependent guarantees. This being said, it is still possible to do a model-dependent analysis of PMEVI-based algorithms. Auer et al. (2009); Filippi et al. (2010) provide $O(\log(T))$ bounds for their algorithms UCRL2 and KL-UCRL with an analysis that it is probably a good starter for PMEVI-based methods. Such an analysis would only be asymptotic and would fail to address the short-horizon concerns on the estimation of the bias that I have talked about. In fact, in Part IV, we argue that both EVI and PMEVI-based algorithms have $O(\log(T)\log\log(T))$ model dependent regret guarantees through a general argument. The analysis is too general to distinguish between the specificities between EVI and PMEVI however. I conjecture that PMEVI-based algorithms would improve on the bound of UCRL2 Auer et al. (2009) by changing a few diameter dependent quantities to bias span dependent ones, although the diameter would still be present as the switching cost induced by the logarithmic number of episodes.

This being said, PMEVI-based methods are tractable and can be analyzed experimentally. By looking at when projection and mitigation operations trigger, it seems that the bias confidence region is simply too large. I made my best to optimize Lemma II.14 by using as much data from the history as possible, yet I did not try to optimize the numerical constants. Moreover, by

taking into account the whole history of play, the bias estimator is therefore polluted by the early learning data where the algorithm's average quality of play is very poor. On the other hand, the right-side experiments of Figure 7.2 indicate that with the adequate bias information, PMEVI has much better performance than without. Hence, improving the bias estimation sub-routine is crucial to make PMEVI superior to EVI in experiments.

Tuning PMEVI foreshadows a intensive experimental campaign on EVI and PMEVI. From experiments I drove locally, it seems difficult to understand what makes $h^*(\mathcal{M}_t)$ and $h^*(M)$ are far away from each other. It seems pretty common for $h^*(\mathcal{M}_t)$, which is obtained without the projection-mitigation correction, to be close to $h^*(M)$ so that there is no point in using PMEVI. However, such experiments would definitively driven by a model-dependent view which is different from the model independent approach of the regret analysis that PMEVI has been tuned for. Taking a bit of distance from optimistic methods might be necessary. As said above, policy based optimism alone may not be able to reach the model dependent regret lower bound of Part III. The question of whether optimism can be reformulated to address that issue is an interesting research direction.

Appendix of Chapter 7

7.A Construction of PMEVI-DT

This section provides the technical details required to understand the design of PMEVI-DT in Chapter 7. We further discuss the assumptions 1-4 appearing in Theorem II.16 and provide sufficient conditions so that they are met.

7.A.1 Proof of Lemma II.14, estimation of the bias error

Fix $s, s' \in \mathscr{S}$. We denote $\alpha_T := N_T(s \leftrightarrow s')(h^*(s) - h^*(s') - c_T(s,s'))$. We will start by considering the better estimator $c'_T(s,s')$ that satisfies the same equation (II.3) than $c_T(s,s')$ but with $\hat{g}(T)$ changed to h^*, readily:

$$N_t(s \leftrightarrow s')c'_T(s,s') = \sum_{t=0}^{N_T(s\leftrightarrow s')-1} (-1)^i \sum_{t=\tau_i^{s\leftrightarrow s'}}^{\tau_{i+1}^{s\leftrightarrow s'}-1} (g^* - R_t).$$

To avoid a typographical clutter, we write τ_i instead of $\tau_i^{s\leftrightarrow s'}$ in the remaining of the proof and we write $\alpha'_T := N_T(s \leftrightarrow s')(h^*(s) - h^*(s') - c'_T(s,s')$.

(**STEP 1**) We start by relating the two estimators. Intuitively, $\hat{g}(T)$ is a good estimator for g^* when the regret is small. Recall that $\hat{g}(T) := \frac{1}{T}\sum_{t=0}^{T-1} R_t$, hence:

$$\sum_{t=0}^{T-1} |\hat{g}(T) - g^*| = \left| \sum_{t=0}^{T-1} (R_t - g^*) \right| = |\mathrm{Reg}(T)|.$$

Therefore,

$$|\alpha_T| \le \left|\alpha'_T\right| + \left|\alpha_T - \alpha'_T\right| \le \left|\alpha'_T\right| + \sum_{t=0}^{T-1} |\hat{g}(T) - g^*| \le \left|\alpha'_T\right| + |\mathrm{Reg}(T)|.$$

We are left with upper-bounding $\left|\alpha'_T\right|$.

(**STEP 2**) If i is even, then S_{τ_i} and $\S_{\tau_{i+1}} = s'$; otherwise $S_{\tau_i} = s'$ and $S_{\tau_{i+1}} = s$. In both cases, we have $h^*(S_{\tau_{i+1}}) - h^*(S_{\tau_i}) = (-1)^i(h^*(s') - h^*(s))$. Therefore, using Bellman's equation, the quantity $\mathrm{A} := \sum_{t=\tau_i}^{\tau_{i+1}-1}(g^* - R_t)$ satisfies:

$$\begin{aligned}
\mathrm{A} &= \sum_{t=\tau_i}^{\tau_{i+1}-1} \big(p(Z_t) - e_{S_t}\big)h^* + \sum_{t=\tau_i}^{\tau_{i+1}-1} (r(Z_t) - R_t) + \sum_{t=\tau_i}^{\tau_{i+1}-1} \Delta^*(Z_t) \\
&= \sum_{t=\tau_i}^{\tau_{i+1}-1} \big(e_{S_{t+1}} - e_{S_t}\big)h^* + \sum_{t=\tau_i}^{\tau_{i+1}-1} \big(p(Z_t) - e_{S_{t+1}}\big)h^* + \sum_{t=\tau_i}^{\tau_{i+1}-1} (r(Z_t) - R_t) + \sum_{t=\tau_i}^{\tau_{i+1}-1} \Delta^*(Z_t)
\end{aligned}$$

$$= (-1)^i (h^*(s') - h^*(s)) + \sum_{t=\tau_i}^{\tau_{i+1}-1} \big(p(Z_t) - e_{S_{t+1}}\big)h^* + \sum_{t=\tau_i}^{\tau_{i+1}-1} (r(Z_t) - R_t) + \sum_{t=\tau_i}^{\tau_{i+1}-1} \Delta^*(Z_t).$$

Multiplying by $(-1)^i$ and rearranging, $h^*(s') - h^*(s) + (-1)^{i+1} \sum_{t=\tau_i}^{\tau_{i+1}-1} (g^* - R_t)$ appears to be equal to:

$$(-1)^{i+1} \left(\sum_{t=\tau_i}^{\tau_{i+1}-1} \big(\big(p(Z_t) - e_{S_{t+1}}\big)h^* + r(Z_t) - R_t \big) + \sum_{t=\tau_i}^{\tau_{i+1}-1} \Delta^*(Z_t) \right).$$

Proceed by summing over i. By triangular inequality, we obtain:

$$\left|\alpha'_T\right| \leq \left| \sum_{i=0}^{N_T(s \leftrightarrow s')-1} \sum_{t=\tau_i}^{\tau_{i+1}-1} (-1)^{i+1} \big(\big(p(Z_t) - e_{S_{t+1}}\big)h^* + r(Z_t) - R_t \big) \right| + \sum_{i=0}^{N_T(s \leftrightarrow s')-1} \sum_{t=\tau_i}^{\tau_{i+1}-1} \Delta^*(Z_t).$$

Because all Bellman gaps Δ^* are non-negative, the second term is upper-bounded by the pseudo-regret $\sum_{t=0}^{T-1} \Delta^*(Z_t)$. The first term is a martingale, and the martingale difference sequence $(-1)^{i+1}((p(Z_t) - e_{S_{t+1}})h^* + r(Z_t) - R_t$ has span at most $\mathrm{sp}(h^*) + 1$ since rewards are supported in $[0,1]$. Although the number of involved is random, it is upper-bounded by T, hence by the maximal version of Azuma-Hoeffding's inequality (Lemma I.19), we have that with probability at least $1-\delta$, uniformly for $T' \leq T$,

$$\left| \sum_{i=0}^{N_{T'}(s \leftrightarrow s')-1} \sum_{t=\tau_i}^{\tau_{i+1}-1} (-1)^{i+1} \big(\big(p(Z_t) - e_{S_{t+1}}\big)h^* + r(Z_t) - R_t \big) \right| \leq (1 + \mathrm{sp}(h^*)) \sqrt{\tfrac{1}{2} T \log\big(\tfrac{2}{\delta}\big)}.$$

(**STEP 3**) We conclude that with probability $1-\delta$, for all $T' \leq T$,

$$\alpha_{T'} \leq (1 + \mathrm{sp}(h^*)) \sqrt{\tfrac{1}{2} T \log\big(\tfrac{2}{\delta}\big)} + \sum_{t=0}^{T'-1} \Delta^*(Z_t) + \big|\mathrm{Reg}(T')\big|.$$

We are left with relating both $\sum_{t=0}^{T'-1} \Delta^*(Z_t)$ and $|\mathrm{Reg}(T')|$ to $\sum_{t=0}^{T'-1} (\tilde{g} - R_t)$. Using the Bellman equation again, we find that:

$$\begin{aligned} \left| \sum_{t=0}^{T'-1} (g^* - R_t - \Delta^*(Z_t)) \right| &\leq |h^*(S_0) - h^*(S_{T'})| + \left| \sum_{t=0}^{T'-1} \big(\big(p(Z_t) - e_{S_{t+1}}\big)h^* + (r(Z_t) - R_t) \big) \right| \\ &\leq \mathrm{sp}(h^*) + (1 + \mathrm{sp}(h^*)) \sqrt{\tfrac{1}{2} T \log\big(\tfrac{2}{\delta}\big)} \end{aligned}$$

where the last inequality holds with probability $1-\delta$ uniformly over $T' \leq T$ by Azuma-Hoeffding's inequality again (Lemma I.19). Remark that if $y - z \leq x \leq y + z$, then $|x| \leq |y| + |z|$, hence we conclude that with probability $1-\delta$, for all $T' \leq T$:

$$\begin{aligned} \sum_{t=0}^{T'-1} \Delta^*(Z_t) + \big|\mathrm{Reg}(T')\big| &\leq 2 \sum_{t=0}^{T'-1} \Delta^*(Z_t) + (1 + \mathrm{sp}(h^*)) \sqrt{\tfrac{1}{2} T \log\big(\tfrac{2}{\delta}\big)} + \mathrm{sp}(h^*) \\ &\leq 2 \sum_{t=0}^{T'-1} (g^* - R_t) + 3(1 + \mathrm{sp}(h^*)) \sqrt{\tfrac{1}{2} T \log\big(\tfrac{2}{\delta}\big)} + 3\mathrm{sp}(h^*) \\ &\leq 2 \sum_{t=0}^{T'-1} (\tilde{g} - R_t) + 3(1 + \mathrm{sp}(h^*)) \sqrt{\tfrac{1}{2} T \log\big(\tfrac{2}{\delta}\big)} + 3\mathrm{sp}(h^*) \end{aligned}$$

where the last inequality invokes $\tilde{g} \geq g^*$. We conclude that, with probability $1-2\delta$, for all $T' \leq T$, we have:

$$N_{T'}(s \leftrightarrow s')(h^*(s) - h^*(s') - c_{T'}(s,s')) \leq 3\mathrm{sp}(h^*) + (1 + \mathrm{sp}(h^*)) \sqrt{8T \log\big(\tfrac{2}{\delta}\big)} + \sum_{t=0}^{T'-1} (\tilde{g} - R_t).$$

This concludes the proof. □

7.A.2 The confidence region of PMEVI-DT

The algorithm PMEVI-DT can be instantiated with a large panel of possibilities, depending on the type of confidence region one is willing to use for rewards and kernels. In this work, we allow for four types of confidence regions, described below. For conciseness, $q \in \{r, p\}$ is a symbolic letter that can be a reward or a kernel and denote $\mathcal{Q}_t(s,a)$ the confidence region for $q(s,a)$ at time t. If $q = r$, then $\dim(q) = 2$ (Bernoulli rewards) with $\mathcal{Q}(s,a) = [0,1]$; and if $q = p$, then $\dim(q) = S$ with $\mathcal{Q}(s,a) = \mathcal{P}(\mathcal{S})$.

(**C1**) *Azuma-Hoeffding* or *Weissman* type confidence regions, with $\mathcal{Q}_t(s,a)$ taken as:

$$\left\{\tilde{q}(s,a) \in \mathcal{Q}(s,a) : N_t(s,a)\|\hat{q}_t(s,a) - \tilde{q}(s,a)\|_1^2 \leq \dim(q)\log\left(\tfrac{2SA(1+N_t(s,a))}{\delta}\right)\right\}.$$

(**C2**) *Empirical Bernstein* type confidence regions, with $\mathcal{Q}_t(s,a)$ taken as:

$$\left\{\tilde{q}(s,a) \in \mathcal{Q}(s,a) : \forall i, |\hat{q}_t(i|s,a) - \tilde{q}(i|s,a)| \leq \sqrt{\frac{2\mathbf{V}(\hat{q}_t(i|s,a))\log\left(\frac{2\dim(q)SAT}{\delta}\right)}{N_t(s,a)}} + \frac{3\log\left(\frac{2\dim(q)SAT}{\delta}\right)}{N_t(s,a)}\right\}.$$

with the convention that $x/0 = +\infty$ for $x > 0$.

(**C3**) *Empirical likelihood* type confidence regions, with $\mathcal{Q}_t(s,a)$ taken as:

$$\left\{\tilde{q}(s,a) \in \mathcal{Q}(s,a) : N_t(s,a)\mathrm{KL}(\hat{q}_t(s,a)\|\tilde{q}(s,a)) \leq \log\left(\tfrac{2SA}{\delta}\right) + (\dim(q)-1)\log\left(e\left(1 + \tfrac{N_t(s,a)}{\dim q - 1}\right)\right)\right\}.$$

(**C4**) *Trivial* confidence region with $\mathcal{Q}_t(s,a) = \mathcal{Q}(s,a)$.

A few remarks are in order. When rewards are not Bernoulli, only the confidence regions (**C1**) and (**C4**) are elligible among the above. Then, Weissman's inequality must be changed to Azuma's inequality for σ-sub-Gaussian random variables, see Lemma I.22. Since rewards are supported in $[0,1]$, Hoeffding's Lemma guarantees that reward distributions are σ-sub-Gaussian with $\sigma = \frac{1}{2}$.

7.A.2.1 Correctness of the model confidence region $\mathcal{M}_t$ and Assumption 1

The confidence regions $\mathcal{Q}_t(s,a)$ described with (**C1-4**) are tuned so that the following result holds:

Lemma II.22. *Assume that, for all $q \in \{r,p\}$ and $(s,a) \in \mathcal{X}$, we choose $\mathcal{Q}_t(s,a)$ among (**C1-4**). Then Assumption 1 holds. More specifically, the region of models $\mathcal{M}_t := \prod_{s,a}(\mathcal{R}_t(s,a) \times \mathcal{P}_t(s,a))$ satisfies $\mathbf{P}(\exists t \leq T : M \notin \mathcal{M}_t) \leq \delta$.*

Proof. We show that, for all $q \in \{r,q\}$ and $(s,a) \in \mathcal{X}$, if $\mathcal{Q}_t(s,a)$ is chosen amoung (**C1-4**), then

$$\mathbf{P}(\exists t \leq T : q(s,a) \notin \mathcal{Q}_t(s,a)) \leq \delta.$$

If $\mathcal{Q}_t(s,a)$ is chosen with (**C1**), this is a direct application of Lemma I.23; with (**C2**), this is Lemma I.24; with (**C3**), this is Lemma I.25; and with (**C4**) this is by definition. □

7.A.2.2 Simultaneous correctness of bias confidence region $\mathscr{H}_t$, mitigation β_t and optimism

In this section, we show that if Assumption 1 holds, then the bias confidence region constructed by PMEVI-DT is correct with high probability, and that the mitigation is not too strong. Recall that $(\mathfrak{g}_k, \mathfrak{h}_k)$ are the optimistic gain and bias of the policy deployed in episode k (see Algorithm 1). In particular, we have $\mathfrak{g}_k = \mathfrak{L}_{t_k}\mathfrak{h}_k - \mathfrak{h}_k$ with $\mathfrak{h}_k \in \mathscr{H}_{t_k}$. We start by a result on the deviation of the variance, which is what the variance approximation Algorithm 5 is based on. Recall that the bias confidence region $\mathscr{H}_t$ is obtained as the collection of constraints:

(1) prior constraints (if any) $\mathfrak{h}(s) - \mathfrak{h}(s') \le c_*(s,s')$;

(2) span constraints $\mathfrak{h}(s) - \mathfrak{h}(s') \le c_0 := T^{1/5}$;

(3) dynamically infered constraints $|\mathfrak{h}(s) - \mathfrak{h}(s') - c_t(s',s)| \le \mathrm{error}(c_t, s', s)$ (see Algorithm 3).

We have the following result.

Lemma II.23. *Let $u, v \in \mathscr{H}_t$ and fix p a probability distribution on $\mathscr{S}$. Then for all $s \in \mathscr{S}$,*

$$\mathbf{V}(p,u) \le \mathbf{V}(p,v) + 8c_0 \sum_{s' \in \mathscr{S}} p(s')\ \mathrm{error}(c_t, s', s).$$

Proof. We start by establishing the following result: If p is a probability distribution on $\mathscr{S}$ and $u, v \in \mathbf{R}^{\mathscr{S}}$, we have:

$$\mathbf{V}(p,u) \le \mathbf{V}(p,v) + 2(p \cdot |u-v|)\max(u+v) \tag{II.8}$$

where $\cdot$ is the dot product, u^2 the Hadamard product uu and $|u|$ the vector whose entry s is $|u(s)|$. (II.8) is obtained with a straight forward computation:

$$\begin{aligned}
\mathbf{V}(p,u) - \mathbf{V}(p,v) &= p \cdot (u^2 - v^2) + (p \cdot v)^2 - (p \cdot u)^2 \\
&= p \cdot ((u-v)(u+v)) + (p \cdot (u-v))(p \cdot (u+v)) \\
&\le p \cdot (|u-v|(u+v)) + (p \cdot |u-v|)(p \cdot |u+v|) \\
&\le 2(p \cdot |u-v|)\max(u+v).
\end{aligned}$$

Observe that v can be changed to $v + \lambda e$, where e is the vector full of ones, without changing the result. The same goes for u. We now move to the proof of the main statement. First, translate u and v such that $u(s) = v(s) = 0$. Then, we have:

$$\begin{aligned}
p \cdot (u-v) &= \sum_{s' \in \mathscr{S}} p(s') \Big| u(s') - u(s) - c_t(s',s) + v(s) - v(s') + c_t(s',s) \Big| \\
&\le \sum_{s' \in \mathscr{S}} p(s') \Big(\big| u(s') - u(s) - c_t(s',s) \big| + \big| v(s') - v(s) - c_t(s',s) \big| \Big) \\
&\le 2 \sum_{s' \in \mathscr{S}} p(s')\ \mathrm{error}(c_t, s', s).
\end{aligned}$$

Conclude using that $\max(u+v) \le \max(u) + \max(v) + 2c_0$ for $u, v \in \mathscr{H}$ such that $u(s) = v(s) = 0$. □

Lemma II.24. *Assume that Assumption 1 holds and that $c_0 \ge \mathrm{sp}(h^*)$. Then, with probability $1-4\delta$, for all $k \le K(T)$, (1) $\mathfrak{g}_k \ge g^*$ and (2) $h^* \in \mathscr{H}_{t_k}$ and (3) for all (s,a), $(\hat{p}_{t_k}(s,a) - p(s,a))h^* \le \beta_{t_k}(s,a)$.*

Proof. Let E_1 the event $(\forall k \le K(T), M \in \mathscr{M}_{t_k})$. Let E_2 the event stating that, for all $T' \le T$,

$$N_{T'}(s \leftrightarrow s')\left|h^*(s) - h^*(s') - c_{T'}(s,s')\right| \le 3\mathrm{sp}(h^*) + (1+\mathrm{sp}(h^*))\sqrt{8T\log(\tfrac{2}{\delta})} + 2\sum_{t=0}^{T'-1}(\tilde{g} - R_t),$$

and let E_3 the event stating that, for all $T' \le T$ and for all $(s,a) \in \mathscr{Z}$, we have:

$$(\hat{p}_{T'}(s,a) - p(s,a))h^* \le \sqrt{\frac{2\mathbf{V}(\hat{p}_{T'}(s,a),h^*)\log\left(\frac{SAT}{\delta}\right)}{N_{T'}(s,a)}} + \frac{3\mathrm{sp}(h^*)\log\left(\frac{SAT}{\delta}\right)}{N_{T'}(s,a)}.$$

By Lemma II.14, we have $\mathbf{P}(E_2) \ge 1-2\delta$ and by Lemma I.24, we have $\mathbf{P}(E_3) \ge 1-\delta$, so $\mathbf{P}(E_1 \cap E_2 \cap E_3) \ge 1-4\delta$. We prove by induction on $k \le K(T)$ that, on $E_1 \cap E_2$, (1) $\mathfrak{g}_k \ge g^*$, (2) $h^* \in \mathscr{H}_{t_k}$ (3) and for all (s,a), $(\hat{p}_{t_k}(s,a) - p(s,a))h^* \le \beta_{t_k}(s,a)$, where $\mathfrak{g}_k$ is the optimistic gain of the policy deployed at episode k. For $k=0$, this is obvious. Indeed, $N_0(s \leftrightarrow s') = 0$ for all s,s' hence $c_0(s,s') = c_0 \ge \mathrm{sp}(h^*)$. Therefore,

$$\mathscr{H}_0 \supseteq \left\{\mathfrak{h} \in \mathbf{R}^{\mathscr{S}} : \mathrm{sp}(\mathfrak{h}) \le c_0\right\} \supseteq \left\{\mathfrak{h} \in \mathbf{R}^{\mathscr{S}} : \mathrm{sp}(\mathfrak{h}) \le \mathrm{sp}(h^*)\right\}$$

so contains h^*, proving (2). Moreover, since $N_0(s,a) = 0$, we have $\beta_0(s,a) = +\infty$, proving (3). Finally, since $M \in \mathscr{M}_0$ on E_1, by the statement (2) of Proposition II.13, we have $\mathfrak{g}_k \ge g^*$, hence proving (1).

Now assume that $k \ge 1$. By induction $\mathfrak{g}_\ell \ge g^*$ for all $\ell < k$, so on E_2 we have:

$$N_{t_k}(s \leftrightarrow s')\left|h^*(s) - h^*(s') - c_{t_k}(s,s')\right| \le 3\mathrm{sp}(h^*) + (1+\mathrm{sp}(h^*))\sqrt{8T\log(\tfrac{2}{\delta})} + 2\sum_{\ell=1}^{k-1}\sum_{t=t_\ell}^{t_{\ell+1}-1}(\mathfrak{g}_\ell - R_t).$$

By design of $\mathscr{H}_{t_k}$ (see Algorithm 3), we deduce that (2) $h^* \in \mathscr{H}_{t_k}$. Denote $h_0 \in \mathscr{H}_{t_k}$ the reference point used by Algorithm 5. We have, for all $(s,a) \in \mathscr{Z}$, on $E_1 \cap E_2 \cap E_3$, we have:

$$\begin{aligned}
\left(\hat{p}_{t_k}(s,a) - p(s,a)\right)h^* &\le \sqrt{\frac{2\mathbf{V}(\hat{p}_{t_k}(s,a),h^*)\log\left(\frac{SAT}{\delta}\right)}{N_{t_k}(s,a)}} + \frac{3\mathrm{sp}(h^*)\log\left(\frac{SAT}{\delta}\right)}{N_{t_k}(s,a)} \\
(h^* \in \mathscr{H}_{t_k} + \text{Lemma II.23}) &\le \sqrt{\frac{2\left(\mathbf{V}(\hat{p}_{t_k}(s,a),h_0)\log\left(\frac{SAT}{\delta}\right) + 8c_0\sum_{s'\in\mathscr{S}}\hat{p}_{t_k}(s'|s,a)\ \mathrm{error}(c_{t_k},s',s)\right)\log\left(\frac{SAT}{\delta}\right)}{N_{t_k}(s,a)}} + \frac{3c_0\log\left(\frac{SAT}{\delta}\right)}{N_{t_k}(s,a)} \\
&=: \beta_{t_k}(s,a)
\end{aligned}$$

by construction of Algorithm 5. Accordingly, (3) is satisfied. Finally, $M \in \mathscr{M}_{t_k}$ on E_1 so by Proposition II.13, we have (1) $\mathfrak{g}_k \ge g^*$. □

Corollary II.25. *Assume that, for all $q \in \{r,p\}$ and $(s,a) \in \mathscr{Z}$, we choose $\mathscr{Q}_t(s,a)$ among **(C1-4)**. Then, with probability $1-3\delta$, for all $k \in K(T)$, we have $\mathfrak{g}_k \ge g^*$ and (2) $h^* \in \mathscr{H}_{t_k}$ and (3) for all (s,a), $(\hat{p}_{t_k}(s,a) - p(s,a))h^* \le \beta_{t_k}(s,a)$.*

Proof. By Lemma II.22, Assumption 1 is satisfied. Apply Lemma II.24. □

7.A.2.3 Sub-Weissman reward confidence region and Assumption 2

Although the kernel confidence region can even chosen to be trivial with (**C4**), in order to work, PMEVI-DT needs the reward confidence region to be sub-Weissman in the following sense:

Assumption 2. There exists a constant $C > 0$ such that for all $(s,a) \in \mathscr{S}$, for all $t \le T$, we have:

$$\mathscr{R}_t(s,a) \subseteq \left\{\tilde{r}(s,a) \in \mathscr{R}(s,a) : N_t(s,a)\|\hat{r}_t(s,a) - \tilde{r}(s,a)\|_1^2 \le C\log\left(\frac{2SA(1+N_t(s,a))}{\delta}\right)\right\}.$$

This is indeed the case if $\mathscr{R}_t(s,a)$ is chosen among (**C1-3**).

7.A.3 Convergence of EVI and Assumption 3

We start with a preliminary lemma on the speed of convergence of EVI. The Lemma II.26 is thought to be applied to extended MDPs. Below, when we claim that the action space is compact, we further claim that $a \in \mathscr{A}(s) \mapsto p(s,a)$ is a continuous map, so that the Bellman operator is continuous and that g^* and h^* are well-defined, see Puterman (1994).

Lemma II.26. *Let M a weakly-communicating MDP with finite state space $\mathbf{R}^{\mathscr{S}}$ and compact action space, and let L its Bellman operator. Assume that there exists $\gamma > 0$ such that, $\forall u \in \mathbf{R}^{\mathscr{S}}$,*

$$\forall s \in \mathscr{S}, \exists a \in \mathscr{A}(s), \quad Lu(s) = r(s,a) + p(s,a)u = r(s,a) + \gamma \max(u) + (1-\gamma) q_s^u u \qquad (*)$$

with $q_s^u \in \mathscr{P}(\mathscr{S})$. Then, for all $u \in \mathbf{R}^{\mathscr{S}}$ and all $\epsilon > 0$, if $\operatorname{sp}(L^{n+1}u - L^n u) \geq \epsilon$, then:

$$n \leq 2 + \frac{4\operatorname{sp}(w_0)}{\gamma\epsilon} + \frac{2}{\gamma}\log\left(\frac{2\operatorname{sp}(w_0)}{\epsilon}\right).$$

Proof. Since M is weakly communicating, has finitely many states and compact action space, it has well-defined gain g^* and bias h^* functions. Denote $u_{n+1} := L^n u$.

$$\begin{aligned} w_n &:= \max_{\pi\in\Pi}\{r_\pi + P_\pi u_{n-1}\} - ng^* - h^* \\ &= \max_{\pi\in\Pi}\{r_\pi - g^* + (P_\pi - I)h^* + P_\pi(u_{n-1} - h^* - (n-1)g^*)\} =: \max_{\pi\in\Pi}\{r'_\pi + P_\pi w_{n-1}\}. \end{aligned}$$

Observe that the policy achieving the maximum is the one achieving $u_n = r_\pi + P_\pi u_{n-1}$. Remark that $r'_\pi(s) = -\Delta^*(s,\pi(s)) \leq 0$ is the Bellman gap of the pair $(s,\pi(s))$, that we more simply write Δ_π. For all n, there exists $\pi_n \in \Pi$ such that $w_{n+1} = -\Delta_{\pi_n} + P_{\pi_n} w_n$. Moreover, by assumption, we have $P_{\pi_n} = \gamma \cdot e_{s_n}^\top e + (1-\gamma)Q_n$ where Q_n is a stochastic matrix. Moreover,

$$\big(\min(-\Delta_{\pi_n}) + \gamma w_n(s_n)\big)e + (1-\gamma)Q_n w_n \leq w_{n+1} \leq \big(\max(-\Delta_{\pi_n}) + \gamma w_n(s_n)\big)e + (1-\gamma)Q_n w_n.$$

Hence, $\operatorname{sp}(w_{n+1}) \leq (1-\gamma)\operatorname{sp}(w_n) + \operatorname{sp}(\Delta_{\pi_n})$. In addition, $w_n = L^n u - L^n h^*$, so by non-expansiveness of L in span semi-norm, $\operatorname{sp}(w_{n+1}) \leq \operatorname{sp}(w_n)$. Overall,

$$\operatorname{sp}(w_{n+1}) \leq \min\big((1-\gamma)\operatorname{sp}(w_n) + \operatorname{sp}(\Delta_{\pi_n}), \operatorname{sp}(w_n)\big). \qquad (II.9)$$

Fix $\epsilon > 0$, and let $n_\epsilon := \inf\{n : \operatorname{sp}(w_n) < \epsilon\}$.

Let π^* an optimal policy. We have $w_{n+1} \geq P_{\pi^*} w_n$ so by induction, $w_{n+1} \geq P_{\pi^*}^{n+1} w_0 \geq \min(w_0)e$. Meanwhile, we see that $\|w_n\|_1 \geq \sum_{k=0}^{n-1} \|\Delta_{\pi_k}\|_1 + S\min(w_0)$, so $\sum_{k=0}^{n-1} \|\Delta_{\pi_k}\|_1 \leq \operatorname{sp}(w_0)$. Since $\Delta_{\pi_k} \leq 0$ for all k, we have $\operatorname{sp}(\Delta_{\pi_k}) \leq \|\Delta_{\pi_k}\|_1$ so $\sum_{k=0}^{n-1} \operatorname{sp}(\Delta_{\pi_k}) \leq \operatorname{sp}(w_0)$.

By (II.9), either $\operatorname{sp}(w_{n+1}) \leq (1-\frac{1}{2}\gamma)\max(\epsilon, \operatorname{sp}(w_n))$ or $\operatorname{sp}(\Delta_{\pi_n}) \geq \frac{1}{2}\gamma\epsilon$, but because $\sum_{k=0}^{+\infty} \operatorname{sp}(\Delta_{\pi_k}) \leq \operatorname{sp}(w_0)$, the second case can happen at most $\frac{2\operatorname{sp}(w_0)}{\gamma\epsilon}$ times. We deduce that, for all $n \leq n_\epsilon$,

$$\operatorname{sp}(w_{n+1}) \leq \left(1 - \tfrac{1}{2}\gamma\right)^{n - \frac{2\operatorname{sp}(w_0)}{\gamma\epsilon}} \operatorname{sp}(w_0).$$

In particular, for $n = n_\epsilon - 1$, we get:

$$\epsilon \leq \left(1 - \tfrac{1}{2}\gamma\right)^{n_\epsilon - 2 - \frac{2\operatorname{sp}(w_0)}{\gamma\epsilon}} \operatorname{sp}(w_0).$$

We obtain:

$$n_\epsilon \leq 2 + \frac{2\operatorname{sp}(w_0)}{\gamma\epsilon} + \frac{2}{\gamma}\log\left(\frac{\operatorname{sp}(w_0)}{\epsilon}\right).$$

To conclude, check that $\operatorname{sp}(L^{n+1}u - L^n u) = \operatorname{sp}(w_{n+1} - w_n) \leq 2\operatorname{sp}(w_n)$. □

Before moving to the application of interest, remark that this result can be greatly improved if the supremum $\sup\{\Delta^*(s,a) : \Delta^*(s,a) < 0\}$ is not zero, to change the dominant term $\frac{4\mathrm{sp}(w_0)}{\gamma\epsilon}$ for a constant independent of ϵ.

Corollary II.27. *Assume that the $\mathscr{M}_t$ has non-empty interior, and that its Bellman operator satisfies the requirement of Lemma II.26, i.e., there exists $\gamma > 0$ such that, $\forall u \in \mathbf{R}^{\mathscr{S}}, \forall s \in \mathscr{S}, \exists a \in \mathscr{A}(s), \exists \tilde{r}_t(s,a) \in \mathscr{R}_t(s,a), \exists \tilde{p}_t(s,a) \in \mathscr{P}_t(s,a)$:*

$$\mathscr{L}_t u(s) = \tilde{r}_t(s,a) + \tilde{p}_t(s,a)u = \tilde{r}_t(s,a) + \gamma \max(u) + (1-\gamma)q_s^u u$$

for some $q_s^u \in \mathscr{P}(\mathscr{S})$. Then Assumption 3 is satisfied, and span fix-points $\tilde{h}_t$ of $\mathscr{L}_t$ are such that $g^*(\mathscr{M}_t) = \mathscr{L}_t\tilde{h}_t - \tilde{h}_t$.

Proof. If $\mathscr{M}_t$ is has non-empty interior, it means that for all (s,a), $\mathscr{P}_t(s,a)$ has non-empty interior. Therefore, for all state-action pair, there exists $\tilde{p}_t(s,a) \in \mathscr{P}_t(s,a)$ that is fully supported. It follows that $\mathscr{M}_t$ is communicating, and it follows from standard results Puterman (1994) that its span fix-points $\tilde{h}$ do exist and that $\tilde{g}_t := \mathscr{L}\tilde{h}_t - \tilde{h}_t \in \mathbf{R}e$ does not depend on the initial state.

Moreover, if $\widetilde{M} \in \mathscr{M}_t$ and $\pi \in \Pi$ with $\tilde{g}_\pi \equiv g(\pi, \mathscr{M}_t) \in \mathbf{R}e$, letting $\tilde{r}_\pi := r_\pi(\tilde{M})$ and $\tilde{P}_\pi := P_\pi(\tilde{M})$, we have:

$$\tilde{r}_\pi + \tilde{p}_\pi\tilde{h}_t \leq \mathscr{L}_t\tilde{h}_t \leq \tilde{g}_t e + \tilde{h}_t.$$

So by induction and since $\mathscr{L}_t$ is obviously monotone and linear, we show that:

$$\sum_{k=0}^{n} \tilde{P}_\pi^k \tilde{r}_\pi \leq n\tilde{g}_t e + (I - \tilde{P}_\pi^n)\tilde{h}_\pi.$$

Dividing by n and letting it go to infinity, we obtain $g(\pi, \mathscr{M}_t) \leq \tilde{g}_t$. Observe that we have equility by taking the policy achieving $(\tilde{g}_t, \tilde{h}_t)$.

To see that EVI converges indeed, simply observe that Lemma II.26 provides a finite bound on how much time is required until the $\mathrm{sp}(\mathscr{L}_t^{n+1}u - \mathscr{L}_t^n u) \leq \epsilon$. Hence $\mathrm{sp}(\mathscr{L}_t^{n+1}u - \mathscr{L}_t^n u)$ vanishes to 0. □

About Assumption 3. The assumptions made by Corollary II.27 are met if the kernel confidence regions are:

- Built out of Weissman's inequality (**C1**) (see the next section, also Auer et al. (2009));
- Built out of Bernstein's inequality (**C2**) (because the maximization algorithm to compute $\tilde{p}_t(s,a)u_i$ in EVI has the same greedy properties than with Weissman's inequality);
- Trivial (**C4**) obviously.

For confidence regions build with empirical likelihood estimates (**C3**), there is no guarantee of convergence (although we conjecture that one could be established), although the gain is still well-defined because $\mathscr{M}_t$ remains communicating. However, just like the original work of Filippi et al. (2010), the convergence is always met numerically.

7.A.4 Proof of Theorem II.16: Complexity of PMEVI with Weissman confidence regions

In this section, we show that when one is using Weissman confidence regions for kernels (**C1**), then the iterates of $\mathscr{L}_t$ converge to an ϵ span-fix-point quickly.

Proposition II.28. *Assume that* PMEVI-DT *uses kernel confidence regions of Weissman type* ***(C1)*** *satisfying Assumption 1. Then with probability* $1-\delta$*, the number of iterations of* PMEVI *(see Algorithm 2) is* $\mathrm{O}\big(D\sqrt{SAT}\big)$*, hence the algorithm has polynomial per-step amortized complexity.*

Proof. With Weissman type confidence regions for kernels, for all $t \le T$ and $(s,a) \in \mathscr{Z}$, we have

$$\mathscr{P}_t(s,a) \supseteq \left\{ \tilde{p}(s,a) \in \mathscr{P}(s,a) : \|\tilde{p}(s,a) - \hat{p}_t(s,a)\|_1 \le \sqrt{\frac{S\log(2SAT)}{T}} \right\}$$

It follows that, for all $t \le T$, the extended Bellman operator $\mathscr{L}_t$ satisfies the prerequisite $(*)$ of Lemma II.26 with

$$\gamma = \frac{1}{2}\sqrt{\frac{S\log(2SAT/\delta)}{T}} = \Omega\left(\sqrt{\frac{S\log(T/\delta)}{T}}\right).$$

Under Assumption 1, we have $M \in \mathscr{M}_t$ with probability $1-\delta$. Under this event, $\mathscr{M}_t$ is weakly communicating and $\mathrm{sp}(h^*(\mathscr{M}_t)) \le D(M)$, we can apply Lemma II.26 and conclude that every calls to PMEVI (Algorithm 2) takes

$$\mathrm{O}\left(\frac{\mathrm{sp}(w_0)\sqrt{T}}{\epsilon\sqrt{\frac{S\log(T/\delta)}{T}}}\right) = \mathrm{O}\left(\frac{DT}{\sqrt{S}\log(T)}\right)$$

where we use that $\epsilon = \sqrt{\frac{\log(SAT/\delta)}{T}}$, that $\mathrm{sp}(w_0) = \mathrm{O}(\mathrm{sp}(h^*(\mathscr{M}_t))) = \mathrm{O}(D(M))$ and that $\delta \ge \frac{1}{T}$. Since the number of episodes under the doubling trick (DT) is $\mathrm{O}(SA\log(T))$, we conclude accordingly. □

Every call to the projection operator solves a linear program. Although in theory, this time is polynomial (relying on recent work on the complexity of LP such as Cohen et al. (2020), it is the current matrix multiplication time $\mathrm{O}(S^{2.38})$), in practice, reducing the number of calls to the projection operator is key to run PMEVI-DT in reasonable time.

7.B Analysis of the projected mitigated Bellman operator

In this section, we fix the model region $\mathscr{M}$, the bias region $\mathscr{H}$ and the mitigation vector β, dropping the sub-script t for conciseness. We denote $\hat{r}, \hat{p}$ the respective empirical reward and kernel. Further assume that $\mathscr{H} = \mathscr{H}_0 + \mathbf{R}e$ with $\mathscr{H}_0$ a compact convex set. The associated projection operation (see Section 7.B.2) is denoted Γ. The (vanilla) extended Bellman operator $\mathscr{L}$ associated to $\mathscr{M}$ is given by $\mathscr{L}u(s) := \max_{a\in\mathscr{A}(s)}\{\sup \mathscr{R}(s,a) + \sup \mathscr{P}(s,a)u\}$. The *$\beta$-mitigated extended Bellman operator* associated to $\mathscr{M}$ is:

$$\mathscr{L}^\beta u(s) := \max_{a\in\mathscr{A}(s)} \sup_{\tilde{r}(s,a)\in\mathscr{R}(s,a)} \sup_{\tilde{p}(s,a)\in\mathscr{P}(s,a)} \Big\{ \tilde{r}(s,a) + \min\{\tilde{p}(s,a)u_i, \hat{p}(s,a)u_i + \beta(s,a)\} \Big\}. \quad \text{(II.10)}$$

The function Greedy$(\mathscr{M}, u, \beta)$ returns a stationary deterministic policy that picks its actions among the one reaching the maximum above. The projection of $\mathscr{L}^\beta$ to $\mathscr{H}$ is

$$\mathfrak{L} \equiv \mathfrak{L}^{\beta,\mathscr{H}} := \Gamma \circ \mathscr{L}^\beta. \quad \text{(II.11)}$$

The goal of this section is to establish Proposition II.13 and

- Proposition II.13 statement (1) is a consequence of Lemma II.33;
- Proposition II.13 statement (2) follows from Theorem II.36;

- Proposition II.13 statement (3) follows from Corollary II.38;
- Proposition II.13 statement (4) follows from Corollary II.32;
- Proposition II.13 prerequisites on the projection operator and Lemma II.15 follows from Lemma II.30

7.B.1 Finding an optimistic policy under bias constraints

The main goal is to find and optimistic policy under *bias constraints* (projection) and *bias error constraints* (mitigation). The bias constraints imply that we search for a policy π together with a model $\widetilde{M}$ such that $h(\pi,\widetilde{M}) \in \mathscr{H}$. The bias error means that, for $\tilde{h} \equiv h(\pi,\widetilde{M})$, we want in addition $\tilde{p}(s,\pi(s))\tilde{h} \le \hat{p}(s,\pi(s))\tilde{h} + \beta(s,\pi(s))$ where $\tilde{p}$ is the transition kernel of $\widetilde{M}$. In the end, our goal is to track the solution of the following optimization problem:

$$g^*(\mathscr{H},\beta,\mathscr{M}) := \sup\left\{ g\big(\pi,\widetilde{M}\big) : \begin{array}{c} \pi\in\Pi, \widetilde{M}\in\mathscr{M}, \\ \forall s\in\mathscr{S},\ \tilde{p}(s,\pi(s))\tilde{h} \le \hat{p}(s,\pi(s))\tilde{h} + \beta(s,\pi(s)), \\ \tilde{h}\equiv h(\pi,\widetilde{M})\in\mathscr{H},\ \mathrm{sp}(g\big(\pi,\widetilde{M}\big)) = 0 \end{array} \right\} \tag{II.12}$$

where the supremum is taken with respect to the product order $\mathbf{R}^{\mathscr{S}}$. In particular, if $\mathscr{U} \subseteq \mathscr{R}^{\mathscr{S}}$, check that $u^* = \sup \mathscr{U}$ is obtained as $u^*(s) := \sup\{v(s) : v \in \mathscr{U}\}$. The constraint $\mathrm{sp}(g\big(\pi,\widetilde{M}\big)) = 0$ is suggested by the work of Fruit (2019); Fruit et al. (2018) and is key for the problem to be solvable.

The bias and the β-constraints make the problem to handle with a "pure" extended MDP solution, which is why the extended Bellman operators are mitigated (with β) then projected (with Γ). The mitigation operation guarantees that the β-constraint is satisfied, while the projection on $\mathscr{H}$ makes sure that the bias constraint is satisfied. It is important for both operations to be compatible, i.e., that the β-constraint that $\mathscr{L}^\beta$ forces is not lost when applying Γ. As a matter of fact, projecting then mitigating would not work.

We now explain why $\mathfrak{L}$ can be used to solve (II.12).

7.B.2 Projection operation and definition of $\mathfrak{L}$

We start by discussing why $\mathfrak{L}$ is well-defined at all. The well-definition of $\mathscr{L}^\beta$ is obvious. The point is to explain why the projection onto $\mathscr{H}$ is possible while preserving mandatory structural properties such as monotony, non-expansivity, linearity and more. For general $\mathscr{H}$, such properties are impossible to meet. But the bias confidence region constructed with Algorithm 3 has a specific shape that makes the projection possible. The central property is the one below:

(**A1**) *The downward closure* $\{v \le u : v \in \mathscr{H}\}$ *of every* $u \in \mathbf{R}^{\mathscr{S}}$ *has a maximum in* $\mathscr{H}$.

The only order that we will be considering is the product order on $\mathbf{R}^{\mathscr{S}}$. Recall that a set $\mathscr{U} \subseteq \mathbf{R}^{\mathscr{S}}$ has a *maximum* if there exists $u \in \mathscr{U}$ such that $v \le u$ for all $u \in \mathscr{U}$. A *supremum* of $\mathscr{U}$ is a minimal upper-bound of $\mathscr{U}$, i.e., u such that (1) $v \le u$ for all $v \in \mathscr{U}$ and (2) no w satisfying (1) can be smaller than u. For the product order, the supremum of a subset $\mathscr{U}$ is unique and of the form $u(s) = \sup\{v(s) : v \in \mathscr{U}\}$.

Define the projection $\Gamma : \mathbf{R}^{\mathscr{S}} \to \mathscr{H}$ as such:

$$\Gamma u := \max\{v \le u : v \in \mathscr{H}\}. \tag{II.13}$$

In general, Assumption (**A1**) is satisfied when $\mathscr{H}$ admits a join, i.e., is stable by finite supremum: $u, v \in \mathscr{H} \Rightarrow \sup(u,v) \in \mathscr{H}$.

Lemma II.29. *If* $\mathscr{H}$ *is generated by constraints of the form* $\mathfrak{h}(s) - \mathfrak{h}(s') - c(s,s') \le d(s,s')$*, then it has a join and (**A1**) is satisfied. Moreover,* Γ *is then correctly computed with Algorithm 4.*

Proof. The first half of the result is well-known, see Zhang and Xie (2023), but we recall a proof for self-containedness. Let $v_1, v_2 \in \mathscr{H}$ and define $v_3 := \sup(v_1, v_2)$. Observe that $v_3(s) - v_3(s') \leq \max(v_1(s) - v_1(s'), v_2(s) - v_2(s')) \leq c(s,s') + d(s,s')$. So $v_3 \in \mathscr{H}$.

We continue by showing that if $\mathscr{H}$ has a join, then (II.13) is well-defined. For $s \in \mathscr{S}$, take a sequence v_n^s such that $v_n^s(s) \to \alpha(s) := \sup\{v(s) : v \leq u, v \in \mathscr{H}\}$. Because the span of every element of $\mathscr{H}$ is upper-bounded by $c := \sup\{\mathrm{sp}(v) : v \in \mathscr{H}\}$, it follows that v_n^s evolves in the compact region $\{v \leq u : v \in \mathscr{H}\} \cap \{v : \|v - \alpha s e\|_\infty = 1 + c\}$. We can therefore extract a convergent sequence of v_n^s, converging v_*^s that belongs to $\mathscr{H}$ since the latter is closed. By construction, $v_*^s(s) = \alpha(s)$. Because $\mathscr{H}$ has a join, $v_* := \sup\{v_*^s : s \in \mathscr{S}\} \in \mathscr{H}$. □

Lemma II.30. *Under assumption (A1), the operator $\Gamma u := \max\{v \leq u : v \in \mathscr{H}\}$ is well-defined, and is:*

(1) monotone: $u \leq v \Rightarrow \Gamma u \leq \Gamma v$;
(2) non span-expansive: $\mathrm{sp}(\Gamma u - \Gamma v) \leq \mathrm{sp}(u - v)$;
(3) linear: $\Gamma(u + \lambda e) = \Gamma u + \lambda e$;
(4) $\Gamma u \leq u$.

Proof. The well-definition of Γ is obvious from (**A1**). For (2), if $u \leq v$ then $w \leq u \Rightarrow w \leq v$. Hence $\Gamma u := \max\{w \leq u : w \in \mathscr{H}\} \leq \max\{w \leq v : w \in \mathscr{H}\} =: \Gamma v$. For (3), check that it follows from $\mathscr{H} = \mathscr{H} + \mathbf{R}e$. For (4), we obviously have $\Gamma u := \max\{v \leq u : v \in \mathscr{H}\} \leq u$.

The more difficult point is (2) span non-expansivity. Pick $u, v \in \mathbf{R}^{\mathscr{S}}$. By linearity, it suffices to show the result for $\sum_s u(s) = \sum_s v(s)$. In that case, we have $\mathrm{sp}(v-u) = \max(v-u) + \max(u-v)$. Observe that for all $w \leq u$, we have $w + \min(v-u)e \leq v$. Since $\mathscr{H} = \mathscr{H} + \mathbf{R}e$, it follows that:

$$\max\{w \leq u : u \in \mathscr{H}\} \leq \max\{w \leq v : w \in \mathscr{H}\} + \max(u-v)e.$$

Similarly, we have $\max\{w \leq u : w \in \mathscr{H}\} \geq \max\{w \leq v : w \in \mathscr{H}\} + \min(v-u)e$. Using them both at once, we find $\mathrm{sp}(\Gamma u - \Gamma v) \leq \mathrm{sp}(v-u)$. □

The properties (1), (3) and (4) are essential for $\mathfrak{L}$ to properly address the optimization problem (II.12). The property (2) is just as important, because it plays a central part in the convergence of value iteration. The next result shows similar properties for the β-mitigated extended Bellman operator $\mathscr{L}^\beta$. From now on, we will assume (A1), because it is almost-surely satisfied by the bias confidence region generated by Algorithm 3.

Lemma II.31. *The β-mitigated extended Bellman operator $\mathscr{L}^\beta$ is (1) monotone, (2) non-span-expansive and (3) linear.*

Proof. The properties (1) and (3) directly follow from the definition. We focus on (2). Fix $u, u' \in \mathbf{R}^{\mathscr{S}}$. By Lemma II.37, we can write $\mathscr{L}^\beta u = \tilde{r}_\pi + \tilde{P}_\pi u$ and $\mathscr{L}^\beta u' = \tilde{r}_{\pi'} + \tilde{P}_{\pi'} u'$. In the following, we write $\beta_\pi(s) := \beta(s, \pi(s))$. Check that:

$$\mathscr{L}^\beta u - \mathscr{L}^\beta u' = \tilde{r}_\pi + \tilde{P}_\pi u - \big(\tilde{r}_{\pi'} + \tilde{P}_{\pi'} u'\big) \leq \tilde{r}_\pi + \tilde{P}_\pi u - \big(\tilde{r}_\pi + \min\big\{\tilde{P}_\pi u', \hat{P}_\pi u' + \beta_\pi\big\}\big).$$

If the minimum is reached with $\tilde{P}_\pi u'$, then:

$$\mathscr{L}^\beta u - \mathscr{L}^\beta u' \leq \tilde{P}_\pi (u - u').$$

If the minimum is reached with $\hat{P}_\pi u' + \beta_\pi$, then upper-bound $\tilde{P}_\pi u$ by $\hat{P}_\pi u + \beta_\pi$ to obtain:

$$\mathscr{L}^\beta u - \mathscr{L}^\beta u' \leq \hat{P}_\pi (u - u').$$

Overall, we find that there exists $Q_\pi \in \mathscr{P}_\pi$ such that $\mathscr{L}^\beta u - \mathscr{L}^\beta u' \leq Q_\pi (u - u')$. Similarly, we find $Q_{\pi'} \in \mathscr{P}_{\pi'}$ such that $\mathscr{L}^\beta u - \mathscr{L}^\beta u' \geq Q_{\pi'}(u - u')$. We conclude that:

$$\mathrm{sp}(\mathscr{L}^\beta u - \mathscr{L}^\beta u') \leq \mathrm{sp}((Q_\pi - Q_{\pi'})(u - u')) \leq \mathrm{sp}(u - u').$$

This concludes the proof. □

By composition, we obtain the following result.

Corollary II.32. *$\mathfrak{L}$ is (1) monotone, (2) non-span-expansive and (3) linear. Moreover,* $\mathrm{sp}(\mathfrak{L}u - \mathfrak{L}v) \leq \mathrm{sp}(\mathscr{L}u - \mathscr{L}v)$ *for all* $u, v \in \mathbf{R}^{\mathscr{S}}$.

7.B.3 Fix-points of $\mathfrak{L}$ and (weak) optimism

Lemma II.33. *$\mathfrak{L}$ has a fix-point in span semi-norm, i.e.,* $\exists u \in \mathscr{H}, \mathrm{sp}(\mathfrak{L}u - u) = 0$.

Proof. The idea is to apply Brouwer's fix-point theorem in $\mathbf{R}^{\mathscr{S}}$ quotiented by the equivalence relation $u \sim v \Leftrightarrow \mathrm{sp}(u - v) = 0$, where $\mathrm{sp}(-)$ becomes a norm. By linearity (Corollary II.32), $\mathfrak{L}$ is well-defined in this quotient space, and if $\mathfrak{L}$ is shown continuous on $\mathbf{R}^{\mathscr{S}}$, so will it be on the quotient.

We show that $\mathfrak{L}$ is sequentially continuous on $\mathscr{H}$. Consider a sequence $u_n \in \mathscr{H}^{\mathbf{N}}$ converging to $u \in \mathscr{H}$ and fix $\epsilon > 0$. Provided that $n > N_\epsilon$ for N_ϵ large enough, we have $\|u_n - u\|_\infty < \epsilon$, i.e., $u_n - \epsilon e \leq u_n \leq u + \epsilon e$. Therefore, in the one hand, for all $v \leq u_n$, we have $v - \epsilon e \leq u$ so $\max\{v \leq u_n : v \in \mathscr{H}\} \leq \max\{v \leq u : v \in \mathscr{H}\} + \epsilon e$; And on the other hand, for all $v \leq u$, $v + \epsilon e \leq u_n$ so $\max\{v \leq u : v \in \mathscr{H}\} \leq \max\{v \leq u_n : v \in \mathscr{H}\} + \epsilon e$. Hence:

$$\|\max\{v \leq u : v \in \mathscr{H}\} - \max\{v \leq u_n : v \in \mathscr{H}\}\| \leq \epsilon.$$

It shows that Γ is continuous. The operator $\mathscr{L}^\beta$ is obviously continuous as well, so $\mathfrak{L} = \Gamma \circ \mathscr{L}^\beta$ is continuous by composition. Since $\mathscr{H} = \mathscr{H}_0 + \mathbf{R}e$ with $\mathscr{H}_0$ compact and ocnvex, the quotient $\mathscr{H}/\!\sim$ is compact and convex, and is preserved by $\mathfrak{L}/\!\sim$. By Brouwer's fix-point theorem, $\mathfrak{L}/\!\sim$ has a fix-point in $\mathscr{H}/\!\sim$. So $\mathfrak{L}$ has a span fix-point in $\mathscr{H}$. □

We write $\mathrm{Fix}(\mathfrak{L})$ the span fix-points of $\mathfrak{L}$.

Lemma II.34. *$\mathfrak{L}$ has well-defined growth. Specifically, if $\mathfrak{L}u = u + \mathfrak{g}e$, then:*

(1) There exists $c > 0$, s.t., for all $v \in \mathscr{H}_0$, $(n\mathfrak{g} - c)e + u \leq \mathfrak{L}^n v \leq (n\mathfrak{g} + c)e + u$;
(2) If $u' \in \mathrm{Fix}(\mathfrak{L})$, then $\mathfrak{L}u' - u' = \mathfrak{g}e$.

Proof. Setting $c := \max_{v \in \mathscr{H}_0} \|v - u\|_\infty < \infty$, one can check that $u - ce \leq v \leq u + ce$ for all $v \in \mathscr{H}_0$. this proves (1) for $n = 0$ and we then proceed by induction on $n \geq 0$. By induction, $\mathfrak{L}^n v \leq u + (n\mathfrak{g} + c)e$ and by Corollary II.32, $\mathfrak{L}$ is monotone, so we have:

$$\mathfrak{L}^{n+1} v \leq \mathfrak{L}\mathfrak{L}^n v \leq \mathfrak{L}(u + (n\mathfrak{g} + c)e) = u + ((n+1)\mathfrak{g} + c)e$$

where the last inequality use the linearity of $\mathfrak{L}$ together with $\mathfrak{L}u = u + \mathfrak{g}e$. The lower bound of $\mathfrak{L}^n v$ is shown similarly, establishing (1).

For (2), pick $u' \in \mathrm{Fix}(\mathfrak{L})$ with $\mathfrak{L}u' = u' + \mathfrak{g}'e$. Up to translating u', we can assume that $u' \in \mathscr{H}_0$ and apply (1). We get:

$$(n\mathfrak{g} - c)e + u \leq n\mathfrak{g}'e + u' \leq (n\mathfrak{g} + c)e + u.$$

Divided by n and let it go to infinity. We conclude that $\mathfrak{g} = \mathfrak{g}'$. □

We finally have everything in hand to claim that $\mathfrak{L}$ solves (II.12).

Corollary II.35. *The* growth *of $\mathfrak{L}$ given by $\mathfrak{g} = \mathfrak{L}u - u$ for $u \in \mathrm{Fix}(\mathfrak{L})$ is well-defined, and:*

$$\forall u \in \mathscr{H}, \quad \mathfrak{g}e = \liminf_{n \to \infty} \frac{\mathfrak{L}^n u}{n} = \limsup_{n \to \infty} \frac{\mathfrak{L}^n u}{n}.$$

Moreover, $\mathfrak{g} \geq g^*(\mathscr{H}, \beta, \mathscr{M})$.

Proof. The growth property is a direct consequence of Lemma II.34. We show $\mathfrak{g} \geq g^*(\mathscr{H}, \beta, \mathscr{M})$ which is defined in (II.12). Pick $\pi \in \Pi, \widetilde{M} \in \mathscr{M}$ its model with $\tilde{h} \equiv h(\pi, \widetilde{M})$ and $\tilde{P}_\pi \tilde{h} \leq \hat{P}_\pi \tilde{h} + \beta_\pi$ where $\beta_\pi(s) := \beta(s, \pi(s))$. Up to translation, we can assume that $\tilde{h} \in \mathscr{H}_0$.

We have $g(\pi, \widetilde{M}) = \tilde{g}e$ for $\tilde{g} \in \mathbf{R}$, so

$$\tilde{h} + \tilde{g}e = \tilde{r}_\pi + \tilde{P}_\pi \tilde{h} \leq \mathfrak{L}\tilde{h}$$

by definition. By monotony of $\mathfrak{L}$, see Corollary II.32, $n\tilde{g}e + \tilde{h} \leq \mathfrak{L}^n \tilde{h}$ follows by induction on $n \geq 0$. By Lemma II.34, we further have $\mathfrak{L}^n \tilde{h} \leq n(\mathfrak{g} + c)e + u$ where $u \in \mathrm{Fix}(\mathfrak{L})$. In tandem,

$$\tilde{g}e \leq \mathfrak{g}e + \frac{ce + u - \tilde{h}}{n}.$$

Letting $n \to \infty$, we deduce that $\tilde{g} \leq \mathfrak{g}$. Conclude by taking the best π and $\widetilde{M}$. □

The next theorem follows directly with the same proof technique, and guarantees optimism.

Theorem II.36. *Assume that $g^* + h^* \leq \mathfrak{L}h^*$. Then $\mathfrak{g} \geq g^*$.*

The condition "$g^* + h^* \leq \mathfrak{L}h^*$" can be referred to as a *weak* form of optimism. We qualify this version of optimism as *weak* because it is much weaker than optimism property suggested by Fruit (2019) $\mathscr{L} \geq L$ where L is the Bellman operator of the true MDP. Here, we only ask for $\mathfrak{L}h^* \geq Lh^*$, i.e., optimism at the fix-point of L. This condition is met as soon as $M \in \mathscr{M}$, $h^* \in \mathscr{H}$ and β is large enough, but is in fact much more general.

7.B.4 Modelization of the projected mitigated Bellman operator $\mathfrak{L}$

The aim of this paragraph is to establish Corollary II.38, stating that $\mathfrak{L}u$ can be viewed as a policy produced by $\texttt{Greedy}(\mathscr{M}, u, \beta)$.

Lemma II.37 (Modelization)**.** *For $\pi \in \Pi$, denote $\beta_\pi(s) := \beta(s, \pi(s))$, $\mathscr{R}_\pi := \prod_s \mathscr{R}(s, \pi(s))$ and $\mathscr{P}_\pi := \prod_s \mathscr{P}(s, \pi(s))$. Fix $u \in \mathbf{R}^{\mathscr{S}}$ and let $\pi := \texttt{Greedy}(\mathscr{M}, u, \beta)$.*

(1) If $\mathscr{P}$ is convex, then there exists $(\tilde{r}_\pi, \tilde{P}_\pi) \in \mathscr{R}_\pi \times \mathscr{P}_\pi$ such that $\mathscr{L}_\beta u = \tilde{r}_\pi + \tilde{P}_\pi u$.

(2) Assume that $\mathscr{L}_\beta u = \tilde{r}_\pi + \tilde{P}_\pi u$. There exists $r'_\pi \leq \tilde{r}_\pi$ such that $\mathfrak{L}u = r'_\pi + \tilde{P}_\pi u$.

The convexity requirement of (1) is always true if the kernel confidence region is chosen via (**C1-4**).

Proof. For (1), fix a state $s \in \mathscr{S}$, let $a := \pi(s)$ and $\rho := \min(\sup \mathscr{P}(s,a)u, \hat{p}(s,a)u + \beta(s,a))$. If $\rho = \sup \mathscr{P}(s,a)u$, then there is nothing to say because $\mathscr{P}$ is compact, hence the sup is a max and ρ is of the form $\tilde{p}(s,a)u$. Otherwise, let $\tilde{p}(s,a)u > \hat{p}(s,a)u + \beta(s,a)$ with $\tilde{p}(s,a) \in \mathscr{P}(s,a)$. Introduce, for $\lambda \in [0,1]$,

$$\tilde{p}_\lambda(s,a) := \lambda \tilde{p}(s,a) + (1-\lambda)\hat{p}(s,a).$$

By continuity, there exists $\lambda \in (0,1)$ such that $\tilde{p}_\lambda(s,a)u = \hat{p}(s,a)u + \beta(s,a)$ and by convexity of $\mathscr{P}(s,a)$, $\tilde{p}_\lambda(s,a) \in \mathscr{P}(s,a)$. This proves (1).

For (2), recall that $\mathfrak{L}u = \Gamma \mathscr{L}^\beta u = \Gamma(\tilde{r}_\pi + \tilde{P}_\pi u)$. Since $\Gamma v \leq v$, for $v \in \mathbf{R}^{\mathscr{S}}$, we have:

$$\Gamma(\tilde{r}_\pi + \tilde{P}_\pi u) \leq \tilde{r}_\pi + \tilde{P}_\pi u.$$

Set $r'_\pi := \Gamma(\tilde{r}_\pi + \tilde{P}_\pi u) - \tilde{P}_\pi u$. Check that r'_π satisfies $r'_\pi \leq \tilde{r}_\pi$ and $\mathfrak{L}u = r'_\pi + \tilde{P}_\pi u$. □

The last corollary bellow is crucial to claim that greedy policies are good choices in PMEVI-DT.

Corollary II.38 (Greedy modelization)**.** *Let $u \in \mathbf{R}^{\mathscr{S}}$ and fix $\pi := \texttt{Greedy}(\mathscr{M}, u, \beta)$. If $\mathscr{P}$ is convex, then with the notations of Lemma II.37, there exists $\tilde{r}_\pi \leq \sup \mathscr{R}_\pi$ and $\tilde{P}_\pi \in \mathscr{P}_\pi$ such that $\mathfrak{L}u = \tilde{r}_\pi + \tilde{P}_\pi u$.*

7.C Proof of Theorem II.16: Regret analysis of PMEVI-DT

We recall a few notations. At episode k, the played policy is denoted π_k. As a greedy response to $\mathfrak{h}_k$, by Proposition II.13 (3), there exists $\tilde{r}_k(s) \le \sup \mathscr{R}_{t_k}(s,\pi_k(s))$ and $\tilde{P}_k(s) \in \mathscr{P}_{t_k}(s,\pi(x))$ such that $\mathfrak{h}_k + \mathfrak{g}_k = \tilde{r}_k + \tilde{P}_k \mathfrak{h}_k$. The reward-kernel pair $\tilde{M}_k = (\tilde{r}_k, \tilde{P}_k)$ is referred to as the *optimistic model* of π_k. We write $P_k := P_{\pi_k}(M)$ the true kernel and $\hat{P}_k := P_{\pi_k}(\hat{M}_{t_k})$ the empirical kernel. Likewise, we define the reward functions r_k and $\hat{r}_k$. The optimistic gain and bias satisfy $\mathfrak{g}_k = g(\pi_k, \widetilde{M}_k)$ and $\mathfrak{h}_k = h(\pi_k, \widetilde{M}_k)$. We further denote $c_0 = T^{\frac{1}{5}}$.

Important remark. To slightly simplify the analysis, we assume that PMEVI is run with perfect precision $\epsilon = 0$, i.e., that $\mathfrak{h}_k = \texttt{PMEVI}(\mathscr{M}_{t_k}, \beta_{t_k}, \Gamma_{t_k}, 0)$ hence is a span fix-point of $\mathfrak{L}_{t_k}$. This assumption is mild and can be dropped by adding an extra error term that has to be carried out in the calculations.

7.C.1 Number of episodes under doubling trick (DT)

Lemma II.39 (Number of episodes, Auer et al. (2009))**.** *The number of episodes up to time $T \ge SA$ is upper-bounded by:*

$$K(T) \le SA \log_2\left(\tfrac{8T}{SA}\right).$$

7.C.2 Sum of bias variances

The Lemma II.40 below shows that $\sum_{t=0}^{T-1} \mathbf{V}(p(Z_t), h^*)$ scales as $T\,\mathrm{sp}(h^*)\mathrm{sp}(r) + \mathrm{sp}(h^*)\mathrm{Reg}(T)$ in probability.

Lemma II.40. *With probability at least $1-\delta$, we have:*

$$\sum_{t=0}^{T-1} \mathbf{V}(p(Z_t), h^*) \le 2\mathrm{sp}(h^*)\mathrm{sp}(r)T + \mathrm{sp}(h^*)^2\sqrt{\tfrac{1}{2}T\log\left(\tfrac{1}{\delta}\right)} + 2\mathrm{sp}(h^*)\sum_{t=0}^{T-1}\Delta^*(Z_t) + \mathrm{sp}(h^*)^2.$$

Proof. Using the Bellman equation $h^*(s) + g^*(s) = r(s,a) + p(s,a)h^* + \Delta^*(s,a)$, we have:

$$\mathbf{V}(p(Z_t), h^*) = \big(p(Z_t) - e_{S_t}\big)h^{*2} + 2h^*(S_t)(\Delta^*(Z_t) + r(Z_t) - g^*(S_t)).$$

Since $\mathrm{sp}(h^{*2}) \le \mathrm{sp}(h^*)^2$, we get:

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbf{V}(p(Z_t), h^*) &\le \sum_{t=0}^{T-1}\big(p(Z_t) - e_{S_t}\big)h^{*2} + 2\mathrm{sp}(h^*)\left(\mathrm{sp}(r)T + \sum_{t=0}^{T-1}\Delta^*(Z_t)\right) \\
&= \sum_{t=0}^{T-1}\big(p(Z_t) - e_{S_{t+1}}\big)h^{*2} + 2\mathrm{sp}(h^*)\left(\tfrac{1}{2}\mathrm{sp}(h^*)\mathrm{sp}(r)T + \sum_{t=0}^{T-1}\Delta^*(Z_t)\right) \\
\text{(Lemma I.19)} &\le 2\mathrm{sp}(h^*)\mathrm{sp}(r)T + \mathrm{sp}(h^*)^2\sqrt{\tfrac{1}{2}T\log\left(\tfrac{1}{\delta}\right)} + 2\mathrm{sp}(h^*)\sum_{t=0}^{T-1}\Delta^*(Z_t) + \mathrm{sp}(h^*)^2
\end{aligned}$$

where the last inequality holds with probability $1-\delta$. This concludes the proof. □

7.C.3 Regret and pseudo-regret: A tight relation

In this paragraph, we bound the regret with respect to the pseudo-regret (and conversely) up to a factor of order $(\mathrm{sp}(h^*)\mathrm{sp}(r)\log(\frac{T}{\delta}))^{1/2}$. Hence, in proofs, the pseudo-regret can be changed to the regret with ease.

Lemma II.41. *With probability $1-4\delta$, the regret and the pseudo-regret and linked as follows:*

$$\left|\sum_{t=0}^{T-1}(g^*-R_t)-\sum_{t=0}^{T-1}\Delta^*(Z_t)\right| \leq \left\{\begin{array}{c} 2\sqrt{\left(2\mathrm{sp}(h^*)\mathrm{sp}(r)+\frac{1}{8}\right)T\log\left(\frac{T}{\delta}\right)}+\sqrt{2\mathrm{sp}(h^*)\log\left(\frac{T}{\delta}\right)\sum_{t=0}^{T-1}\Delta^*(Z_t)} \\ +\mathrm{sp}(h^*)\left(\frac{1}{2}T\right)^{\frac{1}{4}}\log^{\frac{3}{4}}\left(\frac{T}{\delta}\right)+4\mathrm{sp}(h^*)\log\left(\frac{T}{\delta}\right)+2\mathrm{sp}(h^*) \end{array}\right\}.$$

Proof. We rely again on the Poisson equation $g^*(S_t)-r(Z_t)-\Delta^*(Z_t)=(p(Z_t)-e_{S_t})h^*$, so:

$$\begin{aligned} \mathrm{A} := \left|\sum_{t=0}^{T-1}(g^*-R_t-\Delta^*(Z_t))\right| &\leq \left|\sum_{t=0}^{T-1}\left(p(Z_t)-e_{S_t}\right)h^*\right|+\left|\sum_{t=0}^{T-1}(R_t-r(Z_t))\right| \\ &\leq \mathrm{sp}(h^*)+\left|\sum_{t=0}^{T-1}\left(p(Z_t)-e_{S_{t+1}}\right)h^*\right|+\left|\sum_{t=0}^{T-1}(R_t-r(Z_t))\right|. \end{aligned}$$

Up to the constant $\mathrm{sp}(h^*)$, the two error terms are respectively a navigation and a reward error. The second is bounded using Azuma's inequality (Lemma I.19), showing that with probability $1-2\delta$, we have:

$$\left|\sum_{t=0}^{T-1}(R_t-r(Z_t))\right| \leq \sqrt{\tfrac{1}{2}T\log\left(\tfrac{1}{\delta}\right)}.$$

We continue by using Freedman's inequality, instantiated in the form of Lemma I.21. With probability $1-\delta$, we have:

$$\left|\sum_{t=0}^{T-1}\left(p(Z_t)-e_{S_{t+1}}\right)h^*\right| \leq \sqrt{2\sum_{t=0}^{T-1}\mathbf{V}(p(Z_t),h^*)\log\left(\tfrac{T}{\delta}\right)}+4\mathrm{sp}(h^*)\log\left(\tfrac{T}{\delta}\right).$$

The quantity $\sum_{t=0}^{T-1}\mathbf{V}(p(Z_t),h^*)$ is a classical one that appears at several places throughout the analysis. Using Lemma II.40, we bount it explicitely. Further simplifying the bound with $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$, we get that with probability $1-4\delta$, we have:

$$\mathrm{A} \leq \left\{\begin{array}{c} \sqrt{2\mathrm{sp}(h^*)\mathrm{sp}(r)T\log\left(\frac{T}{\delta}\right)}+\sqrt{\frac{1}{2}T\log\left(\frac{1}{\delta}\right)}+\sqrt{2\mathrm{sp}(h^*)\log\left(\frac{T}{\delta}\right)\sum_{t=0}^{T-1}\Delta^*(Z_t)} \\ +\mathrm{sp}(h^*)\left(\frac{1}{2}T\right)^{\frac{1}{4}}\log^{\frac{3}{4}}\left(\frac{T}{\delta}\right)+4\mathrm{sp}(h^*)\log\left(\frac{T}{\delta}\right)+2\mathrm{sp}(h^*) \end{array}\right\}.$$

Bound $\log(\frac{1}{\delta})$ by $\log(\frac{T}{\delta})$ and use $\sqrt{a}+\sqrt{b}\leq 2\sqrt{a+b}$ to merge the terms in $\sqrt{T\log(\frac{T}{\delta})}$ under a single square-root. □

Overall, Lemma II.41 states that the regret $\sum_{t=0}^{T-1}(g^*-R_t)$ and the pseudo-regret $\sum_{t=0}^{T-1}\Delta^*(Z_t)$ differ by about $(\mathrm{sp}(h^*)T\log(\frac{T}{\delta}))^{1/2}$ in probability (up to asymptotically negligible additional terms). In general, the precise form of Lemma II.41 is not convenient to use because it is of form form $x\leq y+\alpha\sqrt{y}+\beta$ that is not linear in y. Corollary II.42 factorizes the result into one which will be more convenient in proofs.

Corollary II.42. *Denote $x:=\sum_{t=0}^{T-1}(g^*-R_t)$ and $y:=\sum_{t=0}^{T-1}\Delta^*(Z_t)$. Further introduce:*

$$\begin{aligned} \alpha &:= \sqrt{2\mathrm{sp}(h^*)\log\left(\tfrac{T}{\delta}\right)} \\ \beta &:= 2\sqrt{\left(2\mathrm{sp}(h^*)\mathrm{sp}(r)+\tfrac{1}{2}\right)T\log\left(\tfrac{T}{\delta}\right)}+\mathrm{sp}(h^*)\left(\tfrac{1}{2}T\right)^{\frac{1}{4}}\log^{\frac{3}{4}}\left(\tfrac{T}{\delta}\right)+2\mathrm{sp}(h^*)\left(2\log\left(\tfrac{T}{\delta}\right)+1\right). \end{aligned}$$

Then, with probability $1-4\delta$, we have $\sqrt{x}\leq\sqrt{y}+\frac{1}{2}\alpha+\sqrt{\beta}$ and $\sqrt{y}\leq\sqrt{x}+\alpha+\sqrt{\beta}$.

Proof. This is straight forward algebra from the result of Lemma II.41. □

7.C.4 Proof of Lemma II.17, reward optimism

We start by getting rid of the reward noise. We have:

$$\begin{aligned}\mathrm{Reg}(T) &:= \sum_{t=0}^{T-1}(g^* - R_t) = \sum_{t=0}^{T-1}(g^* - r(Z_t)) + \sum_{t=0}^{T-1}(r(Z_t) - R_t)\\ &\leq \sum_{t=0}^{T-1}(g^* - r(Z_t)) + \sqrt{\tfrac{1}{2}T\log\left(\tfrac{1}{\delta}\right)}\end{aligned}$$

with probability $1-\delta$ by Azuma's inequality (Lemma I.19). We are left with $\sum_{t=0}^{T-1}(g^* - r(Z_t))$. We continue by splitting the regret episodically and invoking optimism. By Lemma II.24, with probability $1-4\delta$, we have $\sum_{t=0}^{T-1}(g^* - r(Z_t)) \leq \sum_k \sum_{t=t_k}^{t_{k+1}-1}(\mathfrak{g}_k - r(Z_t))$. Introduce

$$B_0(T) := \sum_k \sum_{t=t_k}^{t_{k+1}-1}(\mathfrak{g}_k - r(Z_t)). \tag{II.14}$$

We focus on bounding $B_0(T)$. By Assumption 2, $\tilde{r}_k(s,a)$ is of the form $\hat{r}_k(s,a) + \sqrt{C\log(2SAT/\delta)/N_{t_k}(s,a)} - \eta_k(s,a)$ with $\eta_k(s,a) \in \mathbf{R}$. By the statement (3) of Proposition II.13, $\eta_k(s,a) \geq 0$. Therefore,

$$\begin{aligned}B_0(T) &= \sum_k \sum_{t=t_k}^{t_{k+1}-1}(\mathfrak{g}_k - \tilde{r}_k(Z_t)) + \sum_k \sum_{t=t_k}^{t_{k+1}-1}(\tilde{r}_k(Z_t) - r(Z_t))\\ &\leq \sum_k \sum_{t=t_k}^{t_{k+1}-1}(\mathfrak{g}_k - \tilde{r}_k(Z_t)) + SA + \sum_k \sum_{t=t_k}^{t_{k+1}-1}\mathbf{1}(N_{t_k}(Z_t) \geq 1)\left(\hat{r}_k(Z_t) - r(Z_t) + \sqrt{\frac{C\log\left(\frac{2SAT}{\delta}\right)}{N_{t_k}(Z_t)}}\right)\\ &\overset{(*)}{\leq} \sum_k \sum_{t=t_k}^{t_{k+1}-1}(\mathfrak{g}_k - \tilde{r}_k(Z_t)) + SA + \sum_k \sum_{t=t_k}^{t_{k+1}-1}\mathbf{1}(N_{t_k}(Z_t) \geq 1)\left(\sqrt{\frac{2\log\left(\frac{2SAT}{\delta}\right)}{N_{t_k}(s,a)}} + \sqrt{\frac{C\log\left(\frac{2SAT}{\delta}\right)}{N_{t_k}(s,a)}}\right)\end{aligned}$$

where $(*)$ holds with probability $1-\delta$ following Lemma I.23. By the doubling trick rule (DT), we have $N_t(Z_t) \leq 2N_{t_k}(Z_t)$ for $t < t_{k+1}$, so, with probability $1-\delta$,

$$\begin{aligned}B_0(T) &\leq \sum_k \sum_{t=t_k}^{t_{k+1}-1}(\mathfrak{g}_k - \tilde{r}_k(Z_t)) + SA + 2\sum_k \sum_{t=t_k}^{t_{k+1}-1}\mathbf{1}(N_{t_k}(Z_t) \geq 1)\sqrt{\frac{(2+C)\log\left(\frac{2SAT}{\delta}\right)}{N_{t_k}(s,a)}}\\ &\leq \sum_k \sum_{t=t_k}^{t_{k+1}-1}(\mathfrak{g}_k - \tilde{r}_k(Z_t)) + SA + 2\sqrt{(2+C)\log\left(\tfrac{2SAT}{\delta}\right)}\sum_{s,a}\sum_{n=1}^{N_T(s,a)-1}\sqrt{\tfrac{1}{n}}\\ &\leq \sum_k \sum_{t=t_k}^{t_{k+1}-1}(\mathfrak{g}_k - \tilde{r}_k(Z_t)) + SA + 4\sqrt{(2+C)\log\left(\tfrac{2SAT}{\delta}\right)}\sum_{s,a}\sqrt{N_T(s,a)}\\ \text{(Jensen)} &\leq \sum_k \sum_{t=t_k}^{t_{k+1}-1}(\mathfrak{g}_k - \tilde{r}_k(Z_t)) + SA + 4\sqrt{(2+C)SAT\log\left(\tfrac{2SAT}{\delta}\right)}.\end{aligned}$$

We conclude that with probability $1-6\delta$, we have:

$$\mathrm{Reg}(T) \leq \sum_k \sum_{t=t_k}^{t_{k+1}-1}(\mathfrak{g}_k - \tilde{r}_k(Z_t)) + 4\sqrt{(2+C)SAT\log\left(\tfrac{2SAT}{\delta}\right)} + \sqrt{\tfrac{1}{2}T\log\left(\tfrac{2SAT}{\delta}\right)} + SA. \tag{II.15}$$

This concludes the proof. □

7.C.5 Proof of Lemma II.18, navigation error

We have:

$$\sum_k \sum_{t=t_k}^{t_{k+1}-1} (p_k(S_t) - e_{S_t})\mathfrak{h}_k \le \sum_k \sum_{t=t_k}^{t_{k+1}-1} (p_k(S_t) - e_{S_{t+1}})\mathfrak{h}_k + \sum_k \mathrm{sp}(\mathfrak{h}_k)$$
$$\le \underbrace{\sum_k \sum_{t=t_k}^{t_{k+1}-1} (p_k(S_t) - e_{S_{t+1}})(\mathfrak{h}_k - h^*)}_{\mathrm{A}_1} + \underbrace{\sum_k \sum_{t=t_k}^{t_{k+1}-1} (p_k(S_t) - e_{S_{t+1}})h^*}_{\mathrm{A}_2} + \sum_k \mathrm{sp}(\mathfrak{h}_k).$$

The last term is $\mathrm{O}(c_0 SA\log(T))$ by Lemma II.39, hence is $\mathrm{O}(T^{1/5}\log(T))$.

(**STEP 1**) We start by bounding A_1. By Lemma II.24, with probability $1-4\delta$, we have $h^* \in \mathscr{H}_{t_k}$ for all $k \le K(T)$. So $\mathrm{sp}(\mathfrak{h}_k - h^*) \le \mathrm{sp}(\mathfrak{h}_k) + \mathrm{sp}(h^*) \le 2c_0$. By Freedman's inequality invoked in the form of Lemma I.21, we have with probability $1-5\delta$,

$$\mathrm{A}_1 \le \sqrt{2\sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathrm{V}(p(Z_t), \mathfrak{h}_k - h^*)\log\left(\tfrac{T}{\delta}\right) + 8c_0\log\left(\tfrac{T}{\delta}\right)}$$

It suffices to bound the first term. Recall that e is the vector full of ones. We have:

$$\sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathrm{V}(p(Z_t), \mathfrak{h}_k - h^*) = \sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathrm{V}(p(Z_t), \mathfrak{h}_k - h^* - (\mathfrak{h}_k(S_t) - h^*(S_t))\cdot e)$$
$$\le \sum_k \sum_{t=t_k}^{t_{k+1}-1} \sum_{s'\in\mathscr{S}} p(s'|Z_t)\big(\mathfrak{h}_k(s') - h^*(s') - (\mathfrak{h}_k(S_t) - h^*(S_t))\big)^2$$
$$\overset{(*)}{\le} 3\sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathbf{E}\left[\sum_{s'\in\mathscr{S}} p(s'|Z_t)\big(\mathfrak{h}_k(s') - h^*(s') - (\mathfrak{h}_k(S_t) - h^*(S_t))\big)^2 \middle| \mathscr{O}_t\right] + 16c_0^2\log\left(\tfrac{1}{\delta}\right)$$
$$= 3\sum_k \sum_{t=t_k}^{t_{k+1}-1} (\mathfrak{h}_k(S_{t+1}) - h^*(S_{t+1}) - (\mathfrak{h}_k(S_t) - h^*(S_t)))^2 + 16c_0^2\log\left(\tfrac{1}{\delta}\right).$$

Here the inequality $(*)$ holds with probability $1-\delta$ following Lemma I.28. We will bound the summand with the bias estimation error $\mathrm{error}(c_k, s, s')$ that spawns the inner regret estimation $B_0(t_k) = \sum_{\ell=1}^{k-1}\sum_{t=t_\ell}^{t_{\ell+1}-1}(\mathfrak{g}_\ell - R_t)$. This inner estimation is linked to $B(T) := \sum_{k,t}(\mathfrak{g}_k - R_t)$ the overall optimistic regret by:

$$B_0(t_k) \le \sum_{\ell=1}^{K(T)}\sum_{t=t_\ell}^{t_{\ell+1}-1}(\mathfrak{g}_k - R_t) - \sum_{\ell=k}^{K(T)}\sum_{t=t_\ell}^{t_{\ell+1}-1}(\mathfrak{g}_k - R_t)$$
$$\overset{(*)}{\le} \sum_{\ell=1}^{K(T)}\sum_{t=t_\ell}^{t_{\ell+1}-1}(\mathfrak{g}_k - R_t) - \sum_{\ell=k}^{K(T)}\sum_{t=t_\ell}^{t_{\ell+1}-1}(g^* - R_t)$$
$$\le \sum_{\ell=1}^{K(T)}\sum_{t=t_\ell}^{t_{\ell+1}-1}(\mathfrak{g}_k - R_t) - \sum_{\ell=k}^{K(T)}\sum_{t=t_k}^{T-1}\big(\Delta^*(Z_t) + \big(p(Z_t) - e_{S_t}\big)h^* + r(Z_t) - R_t\big)$$
$$\le \sum_{\ell=1}^{K(T)}\sum_{t=t_\ell}^{t_{\ell+1}-1}(\mathfrak{g}_k - R_t) + \mathrm{sp}(h^*) - \sum_{\ell=k}^{K(T)}\sum_{t=t_k}^{T-1}\big(\big(p(Z_t) - e_{S_{t+1}}\big)h^* + r(Z_t) - R_t\big)$$

$$\overset{(\dagger)}{\leq} \sum_{\ell=1}^{K(T)} \sum_{t=t_\ell}^{t_{\ell+1}-1} (\mathfrak{g}_k - R_t) + \mathrm{sp}(h^*) + (1+\mathrm{sp}(h^*))\sqrt{\tfrac{1}{2}T\log\left(\tfrac{1}{\delta}\right)}$$

$$=: B(T) + \mathrm{sp}(h^*) + (1+\mathrm{sp}(h^*))\sqrt{\tfrac{1}{2}T\log\left(\tfrac{1}{\delta}\right)}.$$

In the above, $(*)$ holds with probability $1-4\delta$ uniformly on k following Lemma II.24 and $(\dagger)$ holds, also uniformly on k, with probability $1-\delta$ by applying Azuma-Hoeffding's inequality (Lemma I.19). Continuing, still on the event specified by Lemma II.24, we have with probability $1-6\delta$:

$$\begin{aligned}
\sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathbf{V}(p(Z_t), \mathfrak{h}_k - h^*) &\leq 3\sum_k \sum_{t=t_k}^{t_{k+1}-1} \frac{3c_0 + (1+c_0)\sqrt{8t_k\log\left(\frac{2}{\delta}\right)} + 2B_0(t_k)}{N_{t_k}(S_{t+1} \leftrightarrow S_t)} + 16c_0^2\log\left(\tfrac{1}{\delta}\right) \\
&\leq 3\sum_k \sum_{t=t_k}^{t_{k+1}-1} \frac{4c_0 + (1+c_0)\sqrt{32T\log\left(\frac{2}{\delta}\right)} + 2B(T)}{N_{t_k}(S_t, A_t, S_{t+1})} + 16c_0^2\log\left(\tfrac{1}{\delta}\right) \\
(\mathrm{DT}) &\leq 12c_0^2 S^2 A + 3\left(4c_0 + (1+c_0)\sqrt{32T\log\left(\tfrac{2}{\delta}\right)} + 2B(T)\right)S^2 A\log(T) \\
&\quad + 16c_0^2\log\left(\tfrac{1}{\delta}\right).
\end{aligned}$$

(**STEP 2**) For A_2, by Freedman's inequality invoked in the form of Lemma I.21 again, we have with probability $1-\delta$,

$$\begin{aligned}
\mathrm{A}_2 &\leq \sqrt{2\sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathbf{V}(p_k(S_t), h^*)\log\left(\tfrac{T}{\delta}\right) + 8c_0\log\left(\tfrac{T}{\delta}\right)} \\
&\leq \sqrt{2\sum_{t=0}^{T-1} \mathbf{V}(p(Z_t), h^*)\log\left(\tfrac{T}{\delta}\right) + 8c_0\log\left(\tfrac{T}{\delta}\right)}.
\end{aligned}$$

We recognize the sum of variance $\sum_{t=0}^{T-1} \mathbf{V}(p(Z_t), h^*)$ that we leave as is.

(**STEP 3**) As a result, with probability $1-7\delta$, we have:

$$\sum_k \sum_{t=t_k}^{t_{k+1}-1} (p_k(S_t) - e_{S_t})\mathfrak{h}_k \leq \sqrt{2\sum_{t=0}^{T-1} \mathbf{V}(p(Z_t), h^*)\log\left(\tfrac{T}{\delta}\right)} + 2SA^{\frac{1}{2}}\sqrt{3B(T)}\log\left(\tfrac{T}{\delta}\right) + \mathrm{O}\left(SA^{\frac{1}{2}}T^{\frac{7}{20}}\log^{\frac{3}{4}}\left(\tfrac{T}{\delta}\right)\right)$$

when $c_0 = T^{\frac{1}{5}}$. □

7.C.6 Proof of Lemma II.19, empirical bias error

Because h^* is a fixed vector, Bennett's inequality (see Lemma I.27) guarantees that $(\hat{p}_k(S_t) - p_k(S_t))h^*$ is small as follows. By doing a union bound over Lemma I.27 with confidence $\frac{\delta}{SAT}$ over all pairs (s,a) and visits counts $N(s,a) \leq T$, we see that with probability $1-\delta$, for all k, we have:

$$\begin{aligned}
\sum_{t=t_k}^{t_{k+1}-1} (\hat{p}_k(S_t) - p_k(S_t))h^* &\leq \mathrm{sp}(h^*)SA + \sum_{t=t_k}^{t_{k+1}-1} \mathbf{1}(N_{t_k}(Z_t) \geq 1)\left(\sqrt{\frac{2\mathbf{V}(p(Z_t),h^*)\log\left(\frac{SAT}{\delta}\right)}{N_{t_k}(Z_t)}} + \frac{\mathrm{sp}(h^*)\log\left(\frac{SAT}{\delta}\right)}{3N_{t_k}(Z_t)}\right) \\
(\text{by doubling trick}) &\leq \mathrm{sp}(h^*)SA + 2\sum_{t=t_k}^{t_{k+1}-1} \mathbf{1}(N_t(Z_t) \geq 1)\left(\sqrt{\frac{2\mathbf{V}(p(Z_t),h^*)\log\left(\frac{SAT}{\delta}\right)}{N_t(Z_t)}} + \frac{\mathrm{sp}(h^*)\log\left(\frac{SAT}{\delta}\right)}{3N_t(Z_t)}\right).
\end{aligned}$$

Summing this over k and factorizing over state-action pairs, we get that with probability $1-\delta$,

$$\sum_k (2k) \leq \mathrm{sp}(h^*)SA + 2\sum_{s,a}\left(\sum_{n=1}^{N_T(s,a)} \sqrt{\frac{2\mathbf{V}(p(s,a),h^*)\log\left(\frac{SAT}{\delta}\right)}{n}} + \sum_{n=1}^{N_T(s,a)} \frac{\mathrm{sp}(h^*)\log\left(\frac{SAT}{\delta}\right)}{n}\right)$$

$$\leq \mathrm{sp}(h^*)SA + 4\sum_{s,a}\sqrt{N_T(s,a)\mathbf{V}(p(s,a),h^*)\log\left(\tfrac{SAT}{\delta}\right)} + 2\mathrm{sp}(h^*)SA\log\left(\tfrac{SAT}{\delta}\right)\log(T)$$

$$\text{(Jensen)} \leq \mathrm{sp}(h^*)SA + 4\sqrt{SA\sum_{s,a}\mathbf{V}(p(s,a),h^*)\log\left(\tfrac{SAT}{\delta}\right)} + 2\mathrm{sp}(h^*)SA\log\left(\tfrac{SAT}{\delta}\right)\log(T)$$

$$= \mathrm{sp}(h^*)SA + 4\sqrt{\sum_{t=0}^{T-1}\mathbf{V}(p(Z_t),h^*)\log\left(\tfrac{SAT}{\delta}\right)} + 2\mathrm{sp}(h^*)SA\log\left(\tfrac{SAT}{\delta}\right)\log(T)$$

We recognize the sum of variances $\sum_{t=0}^{T-1}\mathbf{V}(p(Z_t),h^*)$, that is left to be upper-bounded later on. □

7.C.7 Proof of Lemma II.20, optimism overshoot

Because of the β-mitigation generated by Algorithm 5, the quantity $(\tilde{p}_k(S_t)-\hat{p}_k(S_t))\mathfrak{h}_k$ is shown to be directly related to $\mathbf{V}(p(Z_t),h^*)$ up to a provably negligible error. Denote h'_k the reference point $\texttt{BiasProjection}(\mathscr{H}_{t_k}, c_{t_k}(-,s_0))$ used in Algorithm 5 (denoted h_0 in the algorithm). By Lemma II.24, with probability $1-4\delta$, we have $h^* \in \mathscr{H}_{t_k}$ for all k. To lighten up notations, we write $d_{t_k}(s',s)$ instead of $\mathrm{error}(c_{t_k},s',s)$.

(**STEP 1**) Denote $\mathrm{A} := (\tilde{p}_k(S_t)-\hat{p}_k(S_t))\mathfrak{h}_k$. By construction of $\tilde{p}_k$, we have $\mathrm{A} \leq \beta_{t_k}(Z_t)$, so:

$$\begin{aligned}
\mathrm{A} &\leq \beta_{t_k}(Z_t)\\
&=: \sqrt{\frac{2\left(\mathbf{V}(\hat{p}_k(S_t),h'_k) + 8c_0\sum_{s'\in\mathscr{S}}\hat{p}_k(s'|S_t)d_{t_k}(s',S_t)\log\left(\tfrac{SAT}{\delta}\right)\right)}{N_{t_k}(Z_t)}} + \frac{3c_0\log\left(\tfrac{SAT}{\delta}\right)}{N_{t_k}(Z_t)}\\
&\leq \underbrace{\sqrt{\frac{2\mathbf{V}(\hat{p}_k(S_t),h'_k)}{N_{t_k}(Z_t)}}}_{\mathrm{A}_1} + \underbrace{\sqrt{\frac{16c_0\sum_{s'\in\mathscr{S}}\hat{p}_k(s'|S_t)d_{t_k}(s',S_t)\log\left(\tfrac{SAT}{\delta}\right)}{N_{t_k}(Z_t)}}}_{\mathrm{A}_2} + \frac{3c_0\log\left(\tfrac{SAT}{\delta}\right)}{N_{t_k}(Z_t)}.
\end{aligned}$$

The rightmost term of A is of order $\mathrm{O}(\log^2(T))$ hence is negligible. We focus on the other two. The analysis of A_1 will spawn a term similar to A_2, hence we start by the second. Recall that d_{t_k} is the bias error provided by Algorithm 3 and that the inner regret estimation is $B_0(t_k) = \sum_{\ell=1}^{k-1}\sum_{t=t_\ell}^{t_{\ell+1}-1}(\mathfrak{g}_\ell - R_t)$. Now, remark that:

$$\begin{aligned}
B_0(t_k) &\leq \sum_{\ell=1}^{K(T)}\sum_{t=t_\ell}^{t_{\ell+1}-1}(\mathfrak{g}_k - R_t) - \sum_{\ell=k}^{K(T)}\sum_{t=t_\ell}^{t_{\ell+1}-1}(\mathfrak{g}_k - R_t)\\
&\overset{(*)}{\leq} \sum_{\ell=1}^{K(T)}\sum_{t=t_\ell}^{t_{\ell+1}-1}(\mathfrak{g}_k - R_t) - \sum_{\ell=k}^{K(T)}\sum_{t=t_\ell}^{t_{\ell+1}-1}(g^* - R_t)\\
&\leq \sum_{\ell=1}^{K(T)}\sum_{t=t_\ell}^{t_{\ell+1}-1}(\mathfrak{g}_k - R_t) - \sum_{\ell=k}^{K(T)}\sum_{t=t_k}^{T-1}\left(\Delta^*(Z_t) + \left(p(Z_t) - e_{S_t}\right)h^* + r(Z_t) - R_t\right)\\
&\leq \sum_{\ell=1}^{K(T)}\sum_{t=t_\ell}^{t_{\ell+1}-1}(\mathfrak{g}_k - R_t) + \mathrm{sp}(h^*) - \sum_{\ell=k}^{K(T)}\sum_{t=t_k}^{T-1}\left(\left(p(Z_t) - e_{S_{t+1}}\right)h^* + r(Z_t) - R_t\right)\\
&\overset{(\dagger)}{\leq} \sum_{\ell=1}^{K(T)}\sum_{t=t_\ell}^{t_{\ell+1}-1}(\mathfrak{g}_k - R_t) + \mathrm{sp}(h^*) + (1+\mathrm{sp}(h^*))\sqrt{\tfrac{1}{2}T\log\left(\tfrac{1}{\delta}\right)}\\
&=: B(T) + \mathrm{sp}(h^*) + (1+\mathrm{sp}(h^*))\sqrt{\tfrac{1}{2}T\log\left(\tfrac{1}{\delta}\right)}.
\end{aligned}$$

In the above, $(*)$ holds with probability $1-4\delta$ uniformly on k following Lemma II.24 and $(\dagger)$ holds, also uniformly on k, with probability $1-\delta$ by applying Azuma-Hoeffding's inequality (Lemma I.19). Therefore, with probability $1-5\delta$, for all k and $t \in \{t_k, \dots, t_{k+1}-1\}$, we have:

$$\begin{aligned}
\sqrt{\frac{16c_0 \sum_{s'\in\mathcal{S}} \hat{p}_k(s'|S_t) d_{t_k}(s', S_t) \log\left(\frac{SAT}{\delta}\right)}{N_{t_k}(Z_t)}} &\leq \frac{\sqrt{16c_0 \log\left(\frac{SAT}{\delta}\right) \sum_{s'\in\mathcal{S}} N_{t_k}(S_t, A_t, s') d_{t_k}(s', S_t)}}{N_{t_k}(Z_t)} \\
&\leq \frac{\sqrt{16c_0 \log\left(\frac{SAT}{\delta}\right) \sum_{s'\in\mathcal{S}} N_{t_k}(S_t \longleftrightarrow s') d_{t_k}(s', S_t)}}{N_{t_k}(Z_t)} \\
&\leq \frac{\sqrt{16c_0 \log\left(\frac{SAT}{\delta}\right) S\left(3c_0 + (1+c_0)\left(1+\sqrt{8T\log\left(\frac{2}{\delta}\right)}\right) + 2B_0(t_k)\right)}}{N_{t_k}(Z_t)} \\
&\leq \frac{\sqrt{16c_0 \log\left(\frac{SAT}{\delta}\right) S\left((1+c_0)\left(3+2\sqrt{8T\log\left(\frac{2}{\delta}\right)}\right) + 2B(T)\right)}}{N_{t_k}(Z_t)} \\
&\leq \frac{\sqrt{16c_0 \log\left(\frac{SAT}{\delta}\right) S\left((1+c_0)\left(3+2\sqrt{8T\log\left(\frac{2}{\delta}\right)} + 2B(T)\right)\right)}}{N_{t_k}(Z_t)}.
\end{aligned}$$

This bound will be enough. We move on to A_1. We have:

$$\begin{aligned}
\sqrt{\mathbf{V}(\hat{p}_k(S_t), h'_k)} &\leq \sqrt{\left|\mathbf{V}(\hat{p}_k(S_t), h'_k) - \mathbf{V}(p(Z_t), h^*)\right|} + \sqrt{\mathbf{V}(p(Z_t), h^*)} \\
&\leq \sqrt{\left|\mathbf{V}(\hat{p}_k(S_t), h'_k) - \mathbf{V}(\hat{p}_k(Z_t), h^*)\right|} \sqrt{\left|\mathbf{V}(\hat{p}_k(S_t), h^*) - \mathbf{V}(p(Z_t), h^*)\right|} + \sqrt{\mathbf{V}(p(Z_t), h^*)} \\
&\overset{(*)}{\leq} \sqrt{8c_0 \sum_{s'\in\mathcal{S}} \hat{p}_k(s'|S_t) d_k(s', S_t)} + \mathrm{sp}(h^*) \sqrt{\|\hat{p_k}(S_t) - p_k(S_t)\|_1} + \sqrt{\mathbf{V}(p(Z_t), h^*)} \\
&\overset{(\dagger)}{\leq} \sqrt{8c_0 \sum_{s'\in\mathcal{S}} \hat{p}_k(s'|S_t) d_k(s', S_t)} + \mathrm{sp}(h^*) \left(\frac{S \log\left(\frac{SAT}{\delta}\right)}{N_{t_k}(Z_t)}\right)^{\frac{1}{4}} + \sqrt{\mathbf{V}(p(Z_t), h^*)} \\
&\leq \frac{\mathrm{A}_2}{\sqrt{2N_{t_k}(Z_t)}} + \mathrm{sp}(h^*) \left(\frac{S \log\left(\frac{SAT}{\delta}\right)}{N_{t_k}(Z_t)}\right)^{\frac{1}{4}} + \sqrt{\mathbf{V}(p(Z_t), h^*)}
\end{aligned}$$

where $(*)$ is obtained by applying Lemma II.23 and $(\dagger)$ holds with probability $1-\delta$ by applying Weissman's inequality, see Lemma I.23. All together, with probability $1-6\delta$, A is upper-bounded by:

$$\mathrm{A} \leq \sqrt{\frac{2\mathbf{V}(p(Z_t), h^*) \log\left(\frac{SAT}{\delta}\right)}{N_{t_k}(Z_t)}} + 2\mathrm{A}_2 + \underbrace{\mathrm{sp}(h^*) \sqrt{\frac{2\log\left(\frac{SAT}{\delta}\right) \sqrt{S \log\frac{SAT}{\delta}}}{N_{t_k}(Z_t)\sqrt{N_{t_k}(Z_t)}} + \frac{3c_0 \log\left(\frac{SAT}{\delta}\right)}{N_{t_k}(Z_t)}}}_{\mathrm{A}_3(k,t)}.$$

(**STEP 2**) The number of visits $N_k(Z_t)$ is lower-bounded by $\frac{1}{2}N_t(Z_t)$ when $N_k(Z_t) \geq 1$ by doubling trick (DT). By summing over t and k, we find that with probability $1-6\delta$,

$$\begin{aligned}
\sum_k (3k) &\leq SAc_0 + \sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathbf{1}_{N_{t_k}(Z_t)\geq 1} \sqrt{\frac{2\mathbf{V}(p(Z_t), h^*) \log\left(\frac{SAT}{\delta}\right)}{N_{t_k}(Z_t)}} + \sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathbf{1}_{N_{t_k}(Z_t)\geq 1} (2\mathrm{A}_2(k,t) + \mathrm{A}_3(k,t)) \\
(\mathrm{DT}) &\leq SAc_0 + 2\sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathbf{1}_{N_{t_k}(Z_t)\geq 1} \sqrt{\frac{2\mathbf{V}(p(Z_t), h^*) \log\left(\frac{SAT}{\delta}\right)}{N_t(Z_t)}} + \sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathbf{1}_{N_{t_k}(Z_t)\geq 1} (2\mathrm{A}_2(k,t) + \mathrm{A}_3(k,t))
\end{aligned}$$

$$\leq SAc_0 + 4\sqrt{2SA\sum_{t=0}^{T-1} \mathrm{V}(p(Z_t), h^*)\log\left(\tfrac{SAT}{\delta}\right)} + \sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathbf{1}_{N_{t_k}(Z_t)\geq 1}(2\mathrm{A}_2(k,t) + \mathrm{A}_3(k,t))$$

where the last inequality is obtained with computations that are similar to those detailed in the proof of Lemma II.19. We recognize the variance that we will leave as is. We finish the proof by bounding the lower order terms A_2 and A_3.

(**STEP 3**) We start with A_2. We have:

$$\sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathbf{1}_{N_{t_k}(Z_t)\geq 1}\mathrm{A}_2(k,t) := \sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathbf{1}_{N_{t_k}(Z_t)\geq 1} \frac{\sqrt{16c_0\log\left(\frac{SAT}{\delta}\right)S\left((1+c_0)\left(3+2\sqrt{8T\log\left(\frac{2}{\delta}\right)}+2B(T)\right)\right)}}{N_{t_k}(Z_t)}$$

$$\text{(DT)} \leq 2\sqrt{16c_0 S\log\left(\tfrac{SAT}{\delta}\right)\left((1+c_0)\left(3+2\sqrt{8T\log\left(\tfrac{2}{\delta}\right)}+2B(T)\right)\right)SA\log(T)}$$

$$\leq 8(1+c_0)S^{\frac{3}{2}}A\log^{\frac{3}{2}}\left(\tfrac{SAT}{\delta}\right)\left(2+4T^{\frac{1}{4}}\log^{\frac{1}{4}}\left(\tfrac{SAT}{\delta}\right)+\sqrt{2B(T)}\right).$$

(**STEP 4**) We are left with A_3. We have:

$$\sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathbf{1}_{N_{t_k}(Z_t)\geq 1}\mathrm{A}_3(k,t) := \sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathbf{1}_{N_{t_k}(Z_t)\geq 1}\left(\mathrm{sp}(h^*)\sqrt{\frac{2\log\left(\frac{SAT}{\delta}\right)\sqrt{S\log\frac{SAT}{\delta}}}{N_{t_k}(Z_t)\sqrt{N_{t_k}(Z_t)}}} + \frac{3c_0\log\left(\frac{SAT}{\delta}\right)}{N_{t_k}(Z_t)}\right)$$

$$\text{(DT)} \leq \sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathbf{1}_{N_{t_k}(Z_t)\geq 1}\left(\mathrm{sp}(h^*)\sqrt{\frac{2\log\left(\frac{SAT}{\delta}\right)\sqrt{S\log\frac{SAT}{\delta}}}{N_{t_k}(Z_t)\sqrt{N_{t_k}(Z_t)}}} + \frac{3c_0\log\left(\frac{SAT}{\delta}\right)}{N_{t_k}(Z_t)}\right)$$

$$\leq C\mathrm{sp}(h^*)S^{\frac{5}{4}}AT^{\frac{1}{4}}\log^{\frac{3}{4}}\left(\tfrac{SAT}{\delta}\right) + 6c_0SA\log\left(\tfrac{SAT}{\delta}\right)$$

$$= \mathrm{O}\left(\mathrm{sp}(h^*)S^{\frac{5}{4}}AT^{\frac{1}{4}}\log\left(\tfrac{SAT}{\delta}\right)\right).$$

This concludes the proof. □

7.C.8 Proof of Lemma II.21, second order error

Recall that by Lemma II.24, with probability $1-4\delta$, $h^* \in \mathscr{H}_{t_k}$ for all k, hence $\mathrm{sp}(\mathfrak{h}_k - h^*) \leq 2c_0$ for all k on the same event. Therefore, with probability $1-4\delta$,

$$\sum_k (4k) := 2c_0SA + \sum_k \sum_{t=t_k}^{t_{k+1}-1} \mathbf{1}_{N_{t_k}(Z_t)\geq 1}(\hat{p}_k(S_t) - p_k(S_t))(\mathfrak{h}_k - h^*)$$

$$= 2c_0SA + \sum_k \sum_{t=t_k}^{t_{k+1}-1} \sum_{s'\in\mathscr{S}} \mathbf{1}_{N_{t_k}(Z_t)\geq 1}(\hat{p}_k(s'|S_t) - p_k(s'|S_t))(\mathfrak{h}_k - h^*(s'))$$

$$\overset{(*)}{\leq} 2c_0SA + 2\sum_k \sum_{t=t_k}^{t_{k+1}-1} \sum_{s'\in\mathscr{S}} \mathbf{1}_{N_{t_k}(Z_t)\geq 1}(\hat{p}_k(s'|S_t) - p_k(s'|S_t))d_{t_k}(s', S_t)$$

$$\overset{(\dagger)}{\leq} 2c_0SA + 2\sum_k \sum_{t=t_k}^{t_{k+1}-1} \sum_{s'\in\mathscr{S}} \mathbf{1}_{N_{t_k}(Z_t)\geq 1}\left(d_k(s', S_t)\sqrt{\frac{2\hat{p}_k(s'|S_t)\log\left(\frac{S^2AT}{\delta}\right)}{N_{t_k}(Z_t)}} + 3d_k(s'|S_t)\frac{\log\left(\frac{S^2AT}{\delta}\right)}{N_{t_k}(Z_t)}\right)$$

$$\leq 2c_0SA + 2\sum_k \sum_{t=t_k}^{t_{k+1}-1} \sum_{s'\in\mathscr{S}} \mathbf{1}_{N_{t_k}(Z_t)\geq 1}\left(\sqrt{c_0}\sqrt{\frac{2\hat{p}_k(s'|S_t)d_k(s', S_t)\log\left(\frac{S^2AT}{\delta}\right)}{N_{t_k}(Z_t)}} + \frac{3c_0\log\left(\frac{S^2AT}{\delta}\right)}{N_{t_k}(Z_t)}\right)$$

$$\leq 2c_0SA + 4\sum_k \sum_{t=t_k}^{t_{k+1}-1} \sum_{s'\in\mathcal{S}} \mathbf{1}_{N_{t_k}(Z_t)\geq 1}\left(\sqrt{c_0}\sqrt{\frac{2\hat{p}_k(s'|S_t)d_k(s',S_t)\log\left(\frac{S^2AT}{\delta}\right)}{N_t(Z_t)}} + \frac{3c_0\log\left(\frac{S^2AT}{\delta}\right)}{N_t(Z_t)}\right)$$

where $(*)$ uses that $h^* \in \mathcal{H}_{t_k}$, and $(\dagger)$ is obtained by applying the empirical Bernstein's inequality, see Lemma I.24, to $\hat{p}_k(s'|S_t) - p_k(s'|S_t)$, and holds with probability $1-\delta$. The rightmost term's sum is upper-bounded by:

$$4\sum_k \sum_{t=t_k}^{t_{k+1}-1} \sum_{s'\in\mathcal{S}} \frac{3c_0\log\left(\frac{S^2AT}{\delta}\right)}{N_t(Z_t)} \leq 12S^2A\log(T)\log\left(\frac{S^2AT}{\delta}\right).$$

For the other term, follow the line of the proof of Lemma II.20 (term A_2). We have with probability $1-5\delta$ (4δ of which is by invoking Lemma II.24):

$$\begin{aligned}\hat{p}_k(s'|S_t)d_k(s',S_t) &= \frac{N_{t_k}(S_t,A_t,s')\left((1+c_0)\left(1+\sqrt{8t_k\log\left(\frac{2}{\delta}\right)}\right)+2B_0(t_k)\right)}{N_{t_k}(S_t\leftrightarrow s')N_{t_k}(Z_t)} \\ &\leq \frac{\left((1+c_0)\left(3+2\sqrt{8T\log\left(\frac{2}{\delta}\right)}+2B(T)\right)\right)}{N_{t_k}(Z_t)}.\end{aligned}$$

Therefore,

$$\sqrt{c_0}\sqrt{\frac{2\hat{p}_k(s'|S_t)d_{t_k}(s',S_t)\log\left(\frac{S^2AT}{\delta}\right)}{N_t(Z_t)}} \leq \frac{4(1+c_0)\sqrt{\left(3+2\sqrt{8T\log\left(\frac{2}{\delta}\right)}+2B(T)\right)\log\left(\frac{S^2AT}{\delta}\right)}}{N_t(Z_t)}.$$

Summing over k, t, s', with probability $1-6\delta$, we have:

$$\sum_k(4k) \leq \left\{\begin{array}{c} 16S^2A(1+c_0)\log^{\frac{1}{2}}\left(\frac{S^2AT}{\delta}\right)\left(\sqrt{2B(T)}+2\left(8T\log\left(\frac{2}{\delta}\right)\right)^{\frac{1}{4}}\right) \\ +32S^2A\left(\log(T)\log\left(\frac{S^2AT}{\delta}\right)+(1+c_0)\log^{\frac{1}{2}}\left(\frac{S^2AT}{\delta}\right)\right)\end{array}\right\}$$

This concludes the proof. □

7.D Details on experiments

7.D.1 River swim

Experiments are run on n-states river-swim. Such MDPs are, despite their size, known to be hard to learn. They consists in n states aligned in a straight line with two playable actions RIGHT and LEFT whose dynamics are given in the figure below. Rewards are Bernoulli and null everywhere excepted for $r(s_n, \text{RIGHT}) = 0.95$ and $r(s_0, \text{LEFT}) = 0.05$.

3-state river-swim. The gain is $g^* \approx 0.82$ and $h^* \approx (-4.28, -2.24, 0.4)$.

5-state river-swim. The gain is $g^* \approx 0.82$ and $h^* \approx (-9.62, -7.58, -4.96, -2.27, 0.45)$.

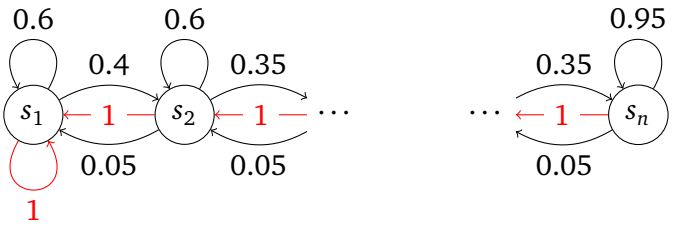


Figure 7.D.1: The kernel of a n-state river-swim.

Part III

Instance Optimal Regret in Average Reward MDPs

This part is dedicated to the first model dependent regret lower bound for communicating Markov decision processes, which appears to be much more challenging than in the recurrent setting. We start with the lower bound in Chapter 8. In Chapter 9, we show that the lower bound is intractable, hence that any asymptotically optimal algorithmic solution should, morally, avoid computing it. Regardless of these computational difficulties, Chapter 10 provides an algorithmic scheme named ECoE that manages to approach the lower bound arbitrarily close. This shows that the lower bound of Chapter 8 is optimal, or roughly speaking, that it is the "true" lower bound for communicating Markov decision processes. In its current form, ECoE is not reasonably implementable because it solves NP-hard problems at every time step. The question of tractable solutions are left for future works.

This part is the conclusion of works done in collaboration with Odalric-Ambrym Maillard.

Chapter 8

The instance dependent lower bound

8.1 A preliminary example

We begin by taking a look at a generic example and try to address a few essential questions. How should an efficient planner behave? How can the way they navigate the model be described?

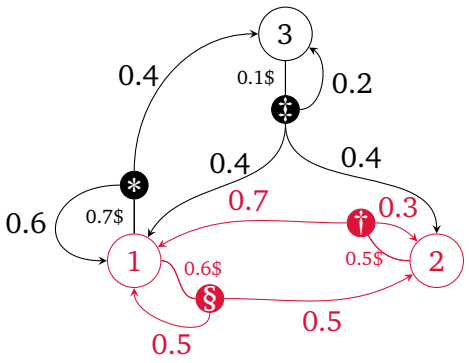


Figure 8.1.1: In the above model, there are two deterministic policies that are determined by the choice of action at state 1, identified by the vectors $(*, \dagger, \ddagger)$ and $(\S, \dagger, \ddagger)$. The optimal pairs $\mathscr{Z}^{**} = \{(1, \S), (2, \dagger)\}$ are represented in red and the optimal policy is $(\S, \dagger, \ddagger)$.

In the reference model M given in Figure 8.1.1, the unique optimal policy is $(\S, \dagger, \ddagger)$ whose recurrent pairs are $\mathscr{Z}^{**} = \{(1, \S), (2, \dagger)\}$, and the Bellman gaps satisfy $\Delta(z) = 0$ with $\Delta(z) > 0$ only for $z = (1, *)$. Any planner with sublinear regret on M will therefore mostly play pairs among $(1, \S), (2, \dagger)$ and $(3, \ddagger)$ and avoid playing $(1, *)$. The structure of M enforces such a planner to spend most of its time playing pairs of $\mathscr{Z}^{**}$, looping around between states 1 and 2. Although the planner mostly plays $(1, \S)$ and $(2, \dagger)$, a **consistent** learner will also make sure that the plausibility that $*$ is a better action than $\S$ from state 1 vanishes, meeting an old idea from Thompson (1933). Otherwise, at least from a Bayesian viewpoint, the planner is subjected to mistake $(1, *)$ for a suboptimal pair in alternative models where it actually is optimal. This will be made formal later on; A consistent planner must play $(1, *)$ at least infinitely often. Of course, an efficient planner will limit the number of tries to $(1, *)$ to the bare minimum, that shall therefore be visited a negligible amount in front of $(1, \S)$ and $(2, \dagger)$.

Overall, we see that any consistent planner will explore M in an unbalanced way, spending most of its time alternating between $(1, \S)$ and $(2, \dagger)$, only rarely trying $(1, *)$ and taking $(3, \ddagger)$

to come back to playing optimal pairs. Formally, we see that we should have:

$$N_T(z) = \Theta(T) \text{ if } z \in \mathscr{Z}^{**} \text{ and } N_T(z) = o(T) \text{ if } z \notin \mathscr{Z}^{**}. \tag{III.1}$$

The way the environment may be explored is determined by the environment, and an efficient learner shall control this environment in order to gather as much information as possible with the smallest cost possible. We are therefore interested in the structure imposed on the vector of visits counts N_T by the environment. Say that $N_T(z) \sim \alpha(z) f(T)$ when $z \notin \mathscr{Z}^{**}$ and for some dimensionless function $f(T) = o(T)$, that will happen to be $f(T) = \log(T)$. Then, this α is what one is really interested in, because the first order regret is given by what happens outside of $\mathscr{Z}^{**}$. This provides a solid motivation to look for a way to **erase** what happens on $\mathscr{Z}^{**}$ during play. Not only are pairs of $\mathscr{Z}^{**}$ zero-cost, but we additionally see that in M of Figure 8.1.1, any state of $\mathscr{S}(\mathscr{Z}^{**})$ is reachable from another state of $\mathscr{S}(\mathscr{Z}^{**})$ using only pairs of $\mathscr{Z}^{**}$. Hence, at any point in time, one does not really mind whether the planner is currently in state 1 or in state 2, because zero-cost pairs can be played to reach one state from the other; And so does a planner that has correctly identified the optimal pairs. Intuitively, optimal pairs can be used to take "shortcuts" in the environment and every state of $\mathscr{S}(\mathscr{Z}^{**})$ accounts for the same.

From a high level viewpoint, N_T is a **quasi-flow** and more precisely, it satisfies $\sum_z N_T(z;s') = \sum_{a' \in \mathscr{A}(s')} N_T(s',a') \pm 1$ for all $s' \in \mathscr{S}$, where $N_T(z;s') := \sum_{t=0}^{T-1} \mathbf{1}(Z_t = z, S_{t+1} = s')$ is the observed number of transitions to s' upon playing z. Informally, the number of visits counts exiting a state is equal to the number of visits that entered that state, up to an error due to the first and last visited states. We are not interested in what happens on $\mathscr{Z}^{**}(M)$, which is likely to hold the dominant mass of N_T seen as a flow. So we artificially remove the mass on $\mathscr{Z}^{**}(M)$, see Figure 8.1.2. The obtained object is not a quasi-flow of M and is instead a quasi-flow of the Markov decision process obtained by merging the states 1 and 2.

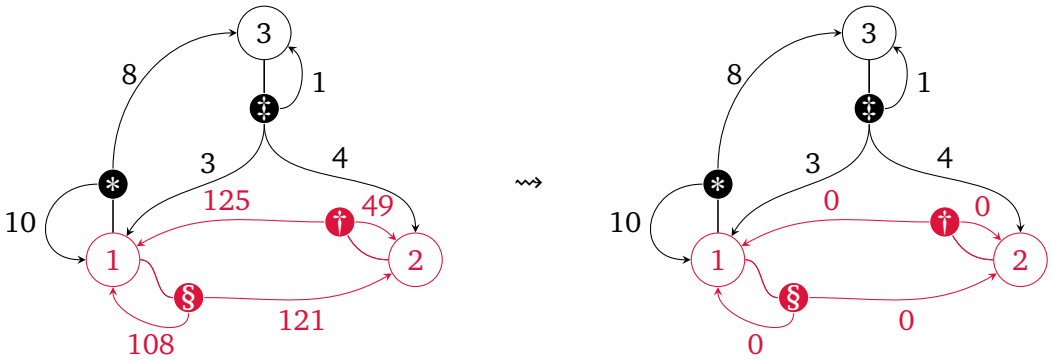


Figure 8.1.2: A quasi-flow of visit counts (to the left) and the object obtained by removing the mass on $\mathscr{Z}^{**}(M)$.

So, **the states of $\mathscr{S}(\mathscr{Z}^{**})$ are merged** into a single quotient state. The kernel of M is reworked to fit the new state-space, as pictured in Figure 8.1.3.

The resulting quotient model is called a **minor** of M, in reference to graph theory, and is denoted $M/\mathscr{Z}^{**}$. Because the visit counts on $M/\mathscr{Z}^{**}$ truncated to $\mathscr{Z} \setminus \mathscr{Z}^{**}(M)$ are a quasi-flow, it happens that the vector α introduced upstream must be an invariant measure of $M/\mathscr{Z}^{**}$. It means that an efficient consistent planner will explore $\mathscr{Z} \setminus \mathscr{Z}^{**}$ following (1) an invariant measure α of $M/\mathscr{Z}^{**}$ that is (2) informative enough so that when used to dictate exploration, all sub-optimal pairs are visited enough to eliminate the plausibility that they are optimal; and (3) makes the regret as small as possible. Indeed, the first order regret is given by $\sum_{z \notin \mathscr{Z}^{**}} N_T(z)\Delta^*(z) \sim \sum_{z \notin \mathscr{Z}^{**}} \alpha(z)\Delta^*(z) \cdot f(T)$ grows with the dot product between α and Δ^*. Hence the smaller is $\sum_{z \in \mathscr{Z}^{**}} \alpha(z)\Delta^*(z)$, the smaller the regret. The best possible behavior that a planner may

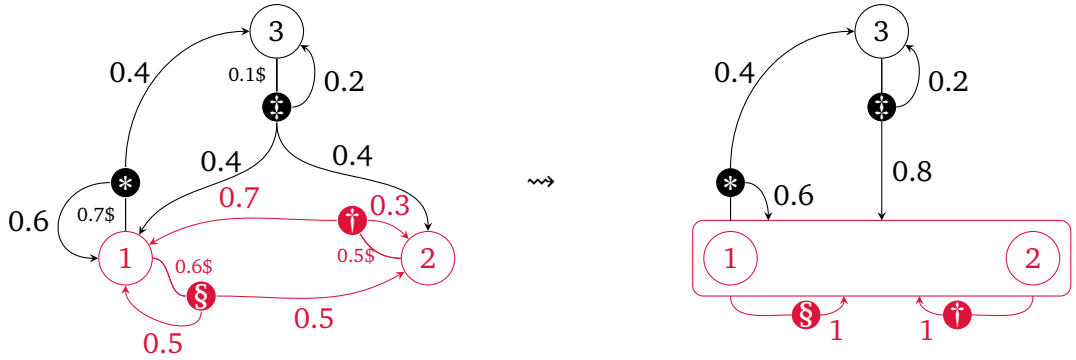


Figure 8.1.3: The contraction of M (see Figure 8.1.1) by $\mathscr{Z}^{\star\star}$, written $M/\mathscr{Z}^{\star\star}$.

follow is therefore described as the optimization problem given by (1-3). In this optimization problem, (1) is reminiscent of what is called the **navigation constraint** in the literature, (2) the **information constraint** and (3) is the **objective function**.

In the sequel, we first investigate information constraints in Section 8.2, then focus on navigation constraints and the role of minors Section 8.3, and finally state the regret lower bound in Section 8.4.

8.2 Confusing models and information constraints

In Section 8.1, we have introduced the idea that a consistent planner must collect enough information to reject the plausibility that seemingly suboptimal pairs are indeed suboptimal. This idea goes back to Lai and Robbins (1985) although our proof machinery is more modern and borrows the formalism of Kaufmann et al. (2016). The core of the argument is a change of measure (Lemma I.18). For all measurable function $f : \mathscr{O}_T \to [0,1]$ of observations, we have

$$\sum_{z\in\mathscr{Z}} \mathbf{E}^{(\pi_t),M}[N_T(z)]\mathrm{KL}(M(z)||M^\dagger(z)) \geq \mathrm{kl}\Big(\mathbf{E}^{(\pi_t),M}[f(O_T)], \mathbf{E}^{(\pi_t),M^{(\pi_t)}}[f(O_T)]\Big) \tag{III.2}$$

where $\mathrm{KL}(M(z)||M^\dagger(z)) := \mathrm{KL}(r(z)||r^\dagger(z)) + \mathrm{KL}(p(z)||p^\dagger(z))$. In (III.2), the LHS is the expected log-likelihood ration between M and $M^\dagger$, and the RHS can be interpreted as a log-likelihood ratio for the value of a test function. This inequality was used to establish the minimax regret lower bound in Part II. Here, it is used to establish the information constraints of the model dependent lower bound. The idea is that if M and $M^\dagger$ have different optimal policies, then a consistent planner should behave differently on M and $M^\dagger$. Therefore, with the right test function f, we may show $\mathbf{E}^{(\pi_t),M}[f(O_T)] \approx 1$ while $\mathbf{E}^{(\pi_t),M^\dagger}[f(O_T)] \ll 1$, to conclude that:

$$\sum_{z\in\mathscr{Z}} \mathbf{E}^{(\pi_t),M}[N_T(z)]\mathrm{KL}(M(z)||M^\dagger(z)) \gg 1.$$

Obviously, all this needs to be precisely quantified. We start by introducing **alternative sets**.

Definition III.1. *For $\mathscr{M} \in \mathfrak{M}(\mathscr{Z})$ a model space and $M \in \mathscr{M}$, we introduce:*

(1) *The **alternative models** of M, denoted $\mathrm{Alt}(M;\mathscr{M})$ as all the $M^\dagger \in \mathscr{M}$ such that $M \ll M^\dagger$ and $\Pi^*(M) \cap \Pi^*(M^\dagger) = \varnothing$;*
(2) *The **confusing models** of M, denoted $\mathrm{Cnf}(M;\mathscr{M})$, as all the alternative $M^\dagger \in \mathscr{M}$ such that $M = M^\dagger$ on $\mathscr{Z}^{\star\star}(M)$.*

When $\mathscr{M}$ is clear in the context, we write $\mathrm{Alt}(M)$ *and* $\mathrm{Cnf}(M)$ *instead.*

In the sequel of the section, we show that if an algorithm is consistent, then it must explore the suboptimal part of the model sufficiently enough to statistically reject **confusing** models.[1] We conjecture that the result below can be generalized to $M^\dagger \in \mathrm{Alt}(M)$. Although this generalization is not necessary to obtain the regret lower bound, it may be of interest for more general purposes, for e.g., the identification of gain optimal policies. Our proof technique is simpler by using structure that confusing models have over mere alternative models and rely on strong consistency (Definition I.22). Recall that a planner is **strongly consistent** on $\mathscr{M}$ if, for all $\epsilon > 0$ and $M \in \mathscr{M}$, $\mathbf{E}^{(\pi_t),M}[\mathrm{Reg}(T)] = \mathrm{o}(T^\epsilon)$.

The result is the following.

Proposition III.1 (Information constraint)**.** *Let $M \in \mathscr{M}$ and let $M^\dagger \in \mathrm{Cnf}(M)$. Then, every strongly consistent learner satisfies:*

$$\liminf_{T\to\infty} \frac{\sum_z \mathbf{E}^M[N_T(z)]\mathrm{KL}_z\left(M||M^\dagger\right)}{\log(T)} \geq 1.$$

This result is established using the notion of **pair covering**, or **covering**. This notion is related to closed state-action spaces in a dynamical sense. A subspace $\mathscr{Z}_0 \subseteq \mathscr{Z}$ is **forward closed** if $p(-|x)$ is supported in $\mathscr{S}(\mathscr{Z}_0)$ for all $z \in \mathscr{Z}_0$. If $\mathscr{S}(\mathscr{Z}_0) = \mathscr{S}$ and $\mathscr{Z}_0$ is forward closed, then it is the graph of the randomized policy π_0 obtained as $\pi(-|s)$ is the uniform distribution on $\{a : (s,a) \in \mathscr{Z}_0\}$; And, conversely, the graph of any randomized policy defines a forward closed pair space. For instance, $\mathscr{Z}_*(M)$ is forward closed in M.

Definition III.2 (Covering of a forward closed space)**.** *Let $\mathscr{Z}_0$ a forward closed state-action space of M. A **covering** of $\mathscr{Z}_0$ is a subset $\mathscr{Z}_c \subseteq \mathscr{Z}_0$ such that, for every policy π whose recurrent pairs belong to $\mathscr{Z}_0$, for every invariant probability measure μ of π seen as an element of $\mathbf{R}^{\mathscr{Z}}$, $\mathrm{supp}(\mu) \cap \mathscr{Z}_c \neq \emptyset$.*

The idea is that if $\mathscr{Z}_c$ covers $\mathscr{Z}_0$, then provided $\mathscr{Z}_0$ is visited a lot, so will $\mathscr{Z}_c$ in probability.

Lemma III.2 (Visits of coverings)**.** *Let $M \in \mathscr{M}$. Let $\mathscr{Z}_0$ a closed state-action space of M and let $\mathscr{Z}_c$ a covering of it. Then, there exists constraints $\epsilon_c, D_c > 0$ such that, whatever the planner, we have:*

$$\forall u \geq 0, \quad \mathbf{P}^M\left(\sum_{x\in\mathscr{Z}_c} N_{T+1}(z) \leq \epsilon_c T - u - D_c \sum_{x\notin\mathscr{Z}_0} N_{T+1}(z)\right) \leq \exp\left(-\frac{2u^2}{TD_c^2}\right).$$

Proof of Lemma III.2. Consider the revised version M_f of M with revised reward vector:

$$f(z) := \begin{cases} -1 & \text{if } z \in \mathscr{Z}_c \\ 0 & \text{otherwise} \end{cases}$$

Remark that $\mathscr{Z}_0(M)$ is closed in M_f. Therefore, $M_f|_{\mathscr{Z}_0(M)}$ is well-defined and we can consider an optimal policy π_f^0 of it which is extended to M_f as π_f by setting it to the uniform policy everywhere it is undefined. Denote g_f, h_f and Δ_f its gain, bias and gap functions.

[1] Alternative sets play a secondary role, but will come in useful later on.

By optimality of π_f on $M_f|_{\mathscr{Z}_0(M)}$, we have $\Delta_f(z) \geq 0$ for all $z \in \mathscr{Z}_0(M)$ and because $\mathscr{Z}_c$ is a covering of $\mathscr{Z}_0(M)$, we also have $g_f(s) < 0$ for all $s \in \mathscr{S}(\mathscr{Z}_0(M)$, hence for $s \in \mathscr{S}$. Let $\epsilon_c := -\max_{s\in\mathscr{S}} g_f(s) < 0$, and denote $D_c := \max\{\mathrm{sp}(h^f), \max_{z\in\mathscr{Z}}|\Delta_f(z)|\} < \infty$. We have:

$$\begin{aligned}\sum_{z\in\mathscr{Z}_c} N_{T+1}(z) &= -\sum_{t=1}^{T} f(Z_t)\\ &\overset{(*)}{=} -\sum_{t=1}^{T}\big(g_f(S_t) + \big(e_{S_t} - p(Z_t)\big)h_f - \Delta_f(Z_t)\big)\\ &\geq \epsilon_c T + \sum_{t=1}^{T}\big(e_{S_{t+1}} - p(Z_t)\big)h_f - D_c\left(1 + \sum_{z\notin\mathscr{Z}_0} N_{T+1}(z)\right)\end{aligned}$$

where $(*)$ invokes the Poisson equation $g_f(s) + h_f(s) = f(s,a) + p(s,a)h_f + \Delta_f(s,a)$. By Azuma-Hoeffding's inequality, the MDS term satisfies:

$$\forall u \geq 0, \quad \mathbf{P}^M\left(\sum_{t=1}^{T}\big(e_{S_{t+1}} - p(Z_t)\big)h_f \leq -u\right) \leq \exp\left(-\frac{2u^2}{TD_c^2}.\right)$$

This concludes the proof. □

We may henceforth prove the proposition of interest.

Proof of Proposition III.1. Let $M^\dagger \in \mathrm{Alt}(M)$, fix $T \geq 1$ and $\eta > 0$. We from that Lemma I.18 that if $\mathscr{E}$ is a $\mathscr{F}_T$-measurable event, then

$$\sum_{z\in\mathscr{Z}} \mathbf{E}^M[N_T(z)]\mathrm{KL}_z\big(M||M^\dagger\big) \geq \mathrm{kl}\Big(\mathbf{P}^M(\mathscr{E}), \mathbf{P}^{M^\dagger}(\mathscr{E})\Big). \tag{III.3}$$

The goal of the proof is to find an event $\mathscr{E}$ that is very likely under M and very unlikely under $M^\dagger$. The RHS of the above equation will be large and the result will follow. The construction of this event is motivated by the idea that the recurrent states of optimal policies of M and $M^\dagger$ are not the same, hence the pairs a consistent algorithm will spend most of its time on will not be the same on M and $M^\dagger$.

(**STEP 1**) The fact that $M^\dagger$ is a confusing model of M implies two important things. First, the structure of $\mathscr{Z}_{**}(M)$ is preserved on $M^\dagger$ and $M^\dagger \gg M$, hence every $\pi \in \Pi^*(M)$ eventually converges to $\mathscr{Z}_{**}(M)$ on $M^\dagger$ and has gain equal to $g(\pi, M^\dagger) = g^*(M)$. Second, no policy of $\Pi^*(M)$ is optimal in $M^\dagger$ by definition, hence the gain achieved on $\mathscr{Z}_{**}(M)$ on $M^\dagger$ is lower than $g^*(M^\dagger)$. Both together, it follows that every policy that converges to $\mathscr{Z}_{**}(M)$ in $M^\dagger$ has sub-optimal gain. Therefore,

$$\mathscr{Z}_c := \mathscr{Z}_{**}(M) \cap \mathscr{Z}_{-}(M^\dagger)$$

is a *covering* of $\mathscr{Z}_*(M)$ in M and in $M^\dagger$, and Lemma III.2 is applicable. Provided $T \geq (4D_c(M)/\epsilon_c(M))^3$, we have:

$$\mathbf{P}^M\left(\sum_{z\in\mathscr{Z}_c} N_{T+1}(z) \leq \frac{1}{2}\epsilon_c(M)T - D_c(M)\left(1 + \sum_{z\in\mathscr{Z}_{-}(M)} N_{T+1}(z)\right)\right) \leq \frac{1}{T}. \tag{III.4}$$

We will be looking at the event $\mathscr{E} := \sum_{z\in\mathscr{Z}_c} N_{T+1}(z) \geq \frac{1}{4}\epsilon_c(M)T$, showing that it has high probability in M and small probability in $M^\dagger$.

(**STEP 2**) We start by looking at what happens in M. We have:

$$\mathbf{P}^M\left(D_c(M)\sum_{z\in\mathscr{Z}_{-}(M)} N_{T+1}(z) \geq \tfrac{1}{8}\epsilon_c(M)T\right) \leq \frac{D_c(M)\mathbf{E}^M\Big[\sum_{z\in\mathscr{Z}_{-}(M)} N_{T+1}(z)\Big]}{\frac{1}{8}\epsilon_c(M)T}$$

$$= \frac{D_c(M)\sum_{z\in\mathscr{Z}_{\_}(M)}\Delta^*(z,M)^{-1}\Delta^*(z,M)\mathbf{E}^M[N_{T+1}(z)]}{\frac{1}{8}\epsilon_c(M)T}$$
$$\leq \frac{D_c(M)\mathbf{E}^M[\mathrm{Reg}(T)]}{\Delta^*_{\min}(M)\cdot\frac{1}{8}\epsilon_c(M)T} = \mathrm{o}\big(T^{\eta-1}\big)$$

where the last equality follows by consistency. Moreover, if T is large enough, we have $D_c(M) \leq \frac{1}{8}\epsilon_c(M)T$, so all combined with (III.4), we obtain:

$$\mathbf{P}^M\left(\sum_{z\in\mathscr{Z}_c} N_{T+1}(z) \geq \tfrac{1}{4}\epsilon_c(M)T\right) = 1-\mathrm{o}\big(T^{\eta-1}\big).$$

(**STEP 3**) Meanwhile, by construction, we have $\mathscr{Z}_c \subseteq \mathscr{Z}\setminus\mathscr{Z}^*(M^\dagger)$, so:

$$\mathbf{P}^{M^\dagger}\left(\sum_{z\in\mathscr{Z}_c} N_{T+1}(z) \geq \tfrac{1}{4}\epsilon_c(M)T\right) \leq \mathbf{P}^{M^\dagger}\left(\sum_{z\notin\mathscr{Z}^*(M^\dagger)} N_{T+1}(z) \geq \tfrac{1}{4}\epsilon_c(M)T\right)$$
$$\leq \frac{\mathbf{E}^{M^\dagger}\big[\sum_{z\notin\mathscr{Z}^*(M^\dagger)} N_{T+1}(z)\big]}{\frac{1}{4}\epsilon_c(M)T}$$
$$= \frac{\sum_{z\notin\mathscr{Z}^*(M^\dagger)}\Delta^*(z,M^\dagger)^{-1}\Delta^*(z,M^\dagger)\mathbf{E}^{M^\dagger}[N_{T+1}(z)]}{\frac{1}{4}\epsilon_c(M)T}$$
$$\leq \frac{\mathbf{E}^{M^\dagger}[\mathrm{Reg}(T)]}{\Delta^*_{\min}(M^\dagger)\frac{1}{4}\epsilon_c(M)T} = \mathrm{o}\big(T^{\eta-1}\big)$$

where the last equality follows by consistency again.

(**STEP 4**) We conclude with the likelihood-ratio inequality:

$$\sum_{x\in\mathscr{Z}}\mathbf{E}^M[N_{T+1}(z)]\mathrm{KL}_z(M||M^\dagger) \geq \mathrm{kl}\Big(\mathbf{P}^M(\mathscr{E}),\mathbf{P}^{M^\dagger}(\mathscr{E})\Big)$$
$$= \mathrm{kl}\big(1-\mathrm{o}\big(T^{\eta-1}\big),\mathrm{o}\big(T^{\eta-1}\big)\big) \gtrsim (1-\eta)\log(T).$$

Divide by $\log(T)$ and take the liminf in T. Conclude by making $\eta\to 0$. ■

8.3 Minors and navigation constraints

In this section, we provide a formal descriptions of minors (Definition III.4) behind the notation $M/\mathscr{Z}^{**}$ in Figure 8.1.3. They generalize edge contraction on graphs to Markov decision processes and are close in spirit to the state reduction/aggregation of Ortner (2013). Minors are obtained by contracting subsets of pairs of the initial model and in opposition to classical graph theory, we won't allow for the contraction of arbitrary subsets of pairs. The contracted subset $\mathscr{Z}'$ must be **closed** (Definition III.3), meaning that (1) one remains in the states spawned by $\mathscr{Z}'$ by playing pairs of $\mathscr{Z}'$, and (2) it does not contain transient states. The first property implies that M can be restricted to $\mathscr{Z}'$ and the states it spawns, and (2) that the obtained model is a union of communicating models.

Definition III.3 (Closed set of M)**.** *A subset $\mathscr{Z}_0\subseteq\mathscr{Z}$ with corresponding states $\mathscr{S}(\mathscr{Z}_0)$ is a **closed** set of M if it is (1) **forward closed** meaning $p(-|z)$ is supported in $\mathscr{S}(\mathscr{Z}_0)$ for all $z\in\mathscr{Z}_0$, and (2) **backward closed**, meaning the model M constrained to $\mathscr{Z}_0$[2] is a union of communicating components (no transient states).*

[2]It is well-defined by forward closeness (1).

A simple, yet illuminating observation, is that a subset $\mathscr{Z}_0$ is closed if, and only if it is the set of recurrent pairs under some randomized stationary policy. The most important example of closed set, in this chapter, is the set of optimal pairs $\mathscr{Z}^{**}(\mathscr{M})$, which is obtained as the recurrent pairs of the policy π^* given by $\pi^*(s)$ as the uniform distribution on $\{a \in \mathscr{A}(s) : (s,a) \in \mathscr{Z}^*(M)\}$.

Definition III.4 (Minors/Contractions)**.** *Up to re-labeling actions, assume that $\mathscr{A}(s) \cap \mathscr{A}(s') = \emptyset$ for $s \neq s'$. Let $M \in \mathscr{M}$ a model and fix $\mathscr{Z}_0 \subseteq \mathscr{Z}$ a closed set of M. The **contraction of** M **by** $\mathscr{Z}_0$ is the model $M/\mathscr{Z}_0$ obtained by merging every communicating component of $\mathscr{Z}_0$ into single states. More formally, letting $\mathscr{S}_1, \dots, \mathscr{S}_k$ the communicating components of $\mathscr{Z}_0$, we have:*

(1) *The state space is $\mathscr{S}(M/\mathscr{Z}_0) := \{\mathscr{S}_1, \dots, \mathscr{S}_k\} \cup \{\{s\} : s \in \mathscr{S} \text{ and } \forall i, s \notin \mathscr{S}_i\}$ and contracted states are denoted $[s]$;*
(2) *The action space is, for $[s] \in \mathscr{S}(M/\mathscr{Z}_0)$, $\mathscr{A}(M/\mathscr{Z}_0)[s] := \bigcup_{s' \in [s]} \mathscr{A}(s')$; Because in M, the choice of an action uniquely determines a state, the state-action space $\mathscr{Z}(M/\mathscr{Z}_0)$ is canonically isomorphic to $\mathscr{Z}(M)$, by associating $([z], a)$ to (s', a) where s' is the unique state such that $a \in \mathscr{A}(s')$.*
(3) *The kernel is $[p]([s_1]|[s_0], a) := \sum_{s_1' \in [s_1]} p(s_1'|s_0', a)$;*
(4) *The reward is $[r]([s_0], a) := r(s_0', a)$.*

*We also say that $M/\mathscr{Z}_0$ is a **minor** of $\mathscr{M}$.*

Minors generically provide a descriptive decomposition of how a Markov decision process can be explored. In Section 8.1, we have claimed that if $N_T(z) \sim \alpha(z) f(T)$ when $z \notin \mathscr{Z}^{**}$ for some $\alpha \in \mathbf{R}^{\mathscr{Z}}$ and function $f(T) = o(T)$, then α must be an invariant measure of $M/\mathscr{Z}^{**}$. This follows from a much more general principle. The collection of optimal pairs $\mathscr{Z}^{**}$ is an example of closed set (Definition III.3), that themselves are recurrent pairs of randomized policies. Independently of the way a planner explores the model, the normalized ratio of visits outside of a closed set converges to an invariant measure of the minor induced by that closed set, provided that the outside is visited at least logarithmically often.

Proposition III.3. *Let π a randomized policy and let $\mathscr{Z}_\pi$ its recurrent pairs. Assume that $\mathbf{E}[\sum_{z \notin \mathscr{Z}_\pi} N_T(z)] = \Omega(\log(T))$. The vector given by*

$$\mu_t(z) := \frac{\mathbf{E}[N_t(z)]\mathbf{1}(z \notin \mathscr{Z}_\pi)}{\mathbf{E}\left[\sum_{z' \notin \mathscr{Z}_\pi} N_t(z')\right]}$$

converges to the space of invariant measures (see Definition I.23) of $M/\mathscr{Z}_\pi$, i.e., every limit point of (μ_t) is an invariant measure of $M/\mathscr{Z}_\pi$.

This is result is coupled with a second observation. It can be shown that if the expected visits counts can be written as $\mathbf{E}^{(\pi_t)}[N_T(z)] = \alpha(z)T + o(T)$, then $\mathscr{Z}_{(\pi_t)} := \{z : \alpha(z) > 0\}$ is a closed set that, if the planner is consistent, is a subset of $\mathscr{Z}^{**}$. Proposition III.3 supports the previously motivated idea that the "sublinear" part of visit counts is easier to understand after contracting the model by $\mathscr{Z}^{**}(M)$.

It will be used in the following form.

Corollary III.4 (Navigation constraints)**.** *Consider a strongly consistent algorithm and let $M \in \mathscr{M}$ such that $\mathrm{Cnf}(M) \neq \varnothing$. Then the vector given by*

$$\mu_T(z) := \frac{\mathbf{E}^M[N_T(z)]\mathbf{1}(z \notin \mathscr{Z}_{**}(M))}{\mathbf{E}^M\Big[\sum_{z' \notin \mathscr{Z}_{**}(M)} N_T(z')\Big]}$$

converges to $\mathrm{Inv}(M/\mathscr{Z}_{**}(M)) \cap \mathscr{P}(\mathscr{Z})$ when $T \to \infty$.

Proof of Corollary III.4. Since $\mathrm{Cnf}(M) \neq \varnothing$, there exists $M^\dagger \in \mathrm{Cnf}(M)$ to which is associated at least one $z \notin \mathscr{Z}_{**}(M)$ such that $\mathrm{KL}_z(M||M^\dagger) \in (0,\infty)$. By Proposition III.1, we have $\mathbf{E}^M[N_T(z)] \gtrsim \mathrm{KL}_z(M||M^\dagger)^{-1}\log(T)$, and from this follows that $\mathbf{E}^M[\sum_{z \notin \mathscr{Z}_{**}(M)} N_T(z)] = \Omega(\log(T))$. Moreover, remark that $\mathscr{Z}_{**}(M)$ is closed as the set of recurrent pairs of the policy π^* such that $\pi^*(-|s)$ is uniform on $\{a : (s,a) \in \mathscr{Z}_{**}(M)\}$ if $s \in \mathscr{S}(\mathscr{Z}_{**}(M))$ and uniform on $\mathscr{A}(s)$ otherwise. Accordingly, we can apply Proposition III.3, to see that

$$\mu'_T(z) := \frac{\mathbf{E}^M[N_T(z)]\mathbf{1}(z \notin \mathscr{Z}_{**}(M))}{\mathbf{E}^M\Big[\sum_{z' \notin \mathscr{Z}_{**}(M)} N_T(z')\Big]}$$

converges to $\mathrm{Inv}(M/\mathscr{Z}_{**}(M))$ as $T \to \infty$. The fact that it is a probability vector is obvious. □

We now move on to the proof of Proposition III.3.

Proof of Proposition III.3. Denote $f(t) := \mathbf{E}[\sum_{z' \notin \mathscr{Z}_\pi} N_t(z')]$ for short. By definition, the states $[M] := M/\mathscr{Z}_\pi$ are subsets of states of the original model M and its pair-space is canonically identical to the one of M. Introduce the $\mathscr{Z}_\pi^c$-truncated visit counts:

$$N'_t(z) := \mathbf{1}(z \notin \mathscr{Z}_\pi)N_t(z)$$

with the induced state-wise visits $N'_t(s) := \sum_a N'_t(s,a)$, and in the minor $N'_t[s] := \sum_{s' \in [s]} N'_t(s')$. By definition, we have $\mu_t(s,a) \equiv \mathbf{E}[N'_t(s,a)]f(t)^{-1}$. Let $\mu_t(s) := \sum_{a \in \mathscr{A}(s)} \mu_t(s,a) = \mathbf{E}[N'_t(s)]$ and, for $[s] \in \mathscr{S}[M]$ a state of the contraction, $\mu_t[s] := \sum_{s' \in \mathscr{S}} \mu_t(s') = \mathbf{E}[N'_t[s]]$. Regarding N'_t as a vector indexed by states of $[M]$, a remarkable property of N'_t is that it satisfies the quasi-flow property (what goes in is what goes out):

$$N'_t[s] = \sum_{s' \in [s]} \sum_z N'_t(z;s') + \mathbf{1}(S_0 \in [s]) - \mathbf{1}(Z_t \in \mathscr{Z}_\pi, S_t \in [s]) = \sum_{s' \in [s]} \sum_{a \in \mathscr{A}(s')} N'_t(s',a) \qquad \text{(III.5)}$$

where $N'_{t+1}(z;s') := \mathbf{1}(z \notin \mathscr{Z}_\pi)N_{t+1}(z;s')$. This is established by induction on $t \geq 0$. This is obvious for $t = 0$, and for $t \geq 1$, we have:

$$\begin{aligned}
(-) &:= \sum_{s' \in [s]} \sum_{a \in \mathscr{A}(s')} N'_t(s',a) \\
&\equiv N'_t[s] = N'_{t-1}[s] + \mathbf{1}(S_t \in [s], X_t \notin \mathscr{Z}_\pi) \\
&\overset{(*)}{=} \mathbf{1}(S_0 \in [s]) - \mathbf{1}(X_{t-1} \in \mathscr{Z}_\pi, S_{t-1} \in [s]) + \sum_{s' \in [s]} \sum_z N'_{t-1}(z;s') + \mathbf{1}(S_t \in [s], X_t \notin \mathscr{Z}_\pi) \\
&= \mathbf{1}(S_0 \in [s]) + \sum_{s' \in [s]} \sum_z N'_t(z;s') \\
&\quad - \mathbf{1}(X_{t-1} \in \mathscr{Z}_\pi, S_{t-1} \in [s]) + \mathbf{1}(S_t \in [s], X_t \notin \mathscr{Z}_\pi) - \mathbf{1}(S_t \in [s], X_{t-1} \notin \mathscr{Z}_\pi)
\end{aligned}$$

where $(*)$ is obtained by induction. We focus on the RHS:

$$\alpha = -\mathbf{1}(X_{t-1} \in \mathscr{Z}_\pi, S_{t-1} \in [s]) + \mathbf{1}(S_t \in [s], X_t \notin \mathscr{Z}_\pi) - \mathbf{1}(S_t \in [s], X_{t-1} \notin \mathscr{Z}_\pi)$$

$$\overset{(\dagger)}{=} -\mathbf{1}(Z_t \in \mathscr{Z}_\pi, S_t \in [s]).$$

The equality (†) is shown by distinguishing cases.

- If $X_{t-1} \in \mathscr{Z}_\pi$ and $Z_t \in \mathscr{Z}_\pi$, then because the states of $\mathscr{Z}_\pi$ are closed by playing pairs of $\mathscr{Z}_\pi$, it follows that $[S_t] = [S_{t-1}]$ correspond to the same recurrent class of π. If $[S_t] = [S_{t-1}] \neq [s]$, then we get (†) : $0 = 0$ and if $[S_t] = [S_{t-1}] = [s]$, then (†) : $-1 = -1$.

- If $X_{t-1} \notin \mathscr{Z}_\pi$ and $Z_t \notin \mathscr{Z}_\pi$, then (†) : $0 = 0$.

- If $X_{t-1} \in \mathscr{Z}_\pi$ and $Z_t \notin \mathscr{Z}_\pi$, then similarly $[S_{t-1}] = [S_t]$. If equal to $[s]$ then (†) : $0 = 0$ and otherwise (†) : $0 = 0$.

- If $X_{t-1} \notin \mathscr{Z}_\pi$ and $Z_t \in \mathscr{Z}_\pi$, then $[S_t]$ and $[S_{t-1}]$ can be equal or different. If (1) $[S_t] = [S_{t-1}] = [s]$, we have (†) : $-1 = -1$; If (2) $[S_t] = [S_{t-1}] \neq [s]$, we have (†) : $0 = 0$; If (3) $[S_t] = [s] \neq [S_{t-1}]$, we have (†) : $-1 = -1$; And if (4) $[S_t] \neq [s] = [S_{t-1}]$, we have (†) : $0 = 0$.

So (†) is established the quasi-flow property follows immediately.

Introduce π'_t the policy of $[M]$ as any state-wise probability distribution with $\mathbf{E}[N'_t[s]]\pi'_t(a|[s]) = \mathbf{E}[N'_t([s],a)]$, which is uniquely defined when $\mathbf{E}[N'_t[s]] > 0$. For all $[s'] \in \mathscr{S}[M]$, we have:

$$\begin{aligned}
\sum_{z \equiv ([s],a) \in \mathscr{Z}[M]} \mu_t[s]\pi'_t(x|[s])p([s']|x) &= \sum_{z \equiv ([s],a) \in \mathscr{Z}[M]} \frac{\mathbf{E}[N'_t[s]]\pi'_t(x|[s])}{f(t)} p([s']|x) \\
&= \sum_{z \notin \mathscr{Z}_\pi} \frac{\mathbf{E}[N_t(z)]}{f(t)} p([s']|x) \\
&= \sum_{z \notin \mathscr{Z}_\pi} \sum_{s'' \in [s']} \frac{\mathbf{E}[N_t(z)]}{f(t)} p(s''|x) \\
&= \sum_{z \notin \mathscr{Z}_\pi} \sum_{s'' \in [s']} \frac{\mathbf{E}[N_t(z)(p(s''|x) - \hat{p}_t(s''|x))] + \mathbf{E}[N_t(z)\hat{p}_t(s''|x)]}{f(t)} \\
\text{(Lemma III.16)} &= \sum_{z \notin \mathscr{Z}_\pi} \sum_{s'' \in [s']} \left(\frac{\mathbf{E}[N_{t+1}(z;s'))]}{f(t)} + \mathrm{o}\left(1 + \frac{\mathbf{E}[N_t(z)]}{f(t)}\right)\right) \\
&= \sum_{z \notin \mathscr{Z}_\pi} \sum_{s'' \in [s']} \left(\frac{\mathbf{E}[N_{t+1}(z;s'')]}{f(t)} + \mathrm{o}(1)\right) \\
&\equiv \sum_{z \notin \mathscr{Z}_\pi} \sum_{s'' \in [s']} \left(\frac{\mathbf{E}[N'_{t+1}(z;s'')]}{f(t)} + \mathrm{o}(1)\right) \\
\text{(quasi-flow property (III.5))} &= \sum_{s'' \in [s']} \left(\frac{\mathbf{E}[N'_{t+1}(s'')]}{f(t)} + \mathrm{o}(1)\right) = \mu[s'] + \mathrm{o}(1).
\end{aligned}$$

This concludes the proof. □

8.4 The model dependent lower bound of the regret

We finally have all the material to state and prove the model dependent lower bound.

Theorem III.5 (Regret lower bound). *Let $M \in \mathscr{M}$. The regret of every strongly consistent algorithm on $\mathscr{M}$ satisfies $\liminf_{T\to\infty} \mathbf{E}^M[\mathrm{Reg}(T)]/\log(T) \geq K(M)$ where $K(M) \in [0,\infty]$ is the solution of the optimization problem:*

$$\inf_{\mu\in \mathrm{Inv}(M/\mathscr{Z}^{**}(M))} \sum_{z\in\mathscr{Z}} \mu(z)\Delta^*(z) \quad s.t. \quad \forall M^\dagger \in \mathrm{Cnf}(M),\ \sum_{z\in\mathscr{Z}} \mu(z)\mathrm{KL}_z(M||M^\dagger) \geq 1. \quad \text{(III.6)}$$

Proof of Theorem III.5. Consider a consistent algorithm and pick $M \in \mathscr{M}$.

(**STEP 1**) If $\mathrm{Cnf}(M) = \varnothing$, then there are no information constraints and $\mu := 0^{\mathscr{Z}} \in \mathrm{Inv}(M/\mathscr{Z}_{**}(M))$ is the optimal exploration measure and the provided lower bound is trivial.

(**STEP 2a**) Moving on from this special case, assume that $\mathrm{Cnf}(M) \neq \varnothing$. By Corollary III.4, the vector

$$\mu_T(z) := \frac{\mathbf{1}(z \notin \mathscr{Z}_{**}(M))\mathbf{E}^M[N_T(z)]}{\mathbf{E}^M[\sum_{z'\notin\mathscr{Z}_{**}(M)} N_T(z')]}$$

converges to $\mathrm{Inv}(M/\mathscr{Z}_{**}(M))\cap\mathscr{P}(\mathscr{Z})$ when $T\to\infty$. Introduce $\lambda(T) := \mathbf{E}^M[\sum_{z'\notin\mathscr{Z}_{**}(M)} N_T(z')]\cdot \log^{-1}(T)$. By consistency of the algorithm, from Proposition III.1 follows that:

$$\forall M^\dagger \in \mathrm{Cnf}(M), \quad \sum_{z\in\mathscr{Z}} \lambda(T)\mu_T(z)\mathrm{KL}_z\big(M||M^\dagger\big) \geq 1.$$

Meanwhile, the regret satisfies:

$$\frac{\mathbf{E}^M[\mathrm{Reg}(T)]}{\log(T)} = \frac{1}{\log(T)}\sum_{z\in\mathscr{Z}} \mathbf{E}^M[N_T(z)]\Delta^*(z,M) = \sum_{z\in\mathscr{Z}} \lambda(T)\mu_T(z)\Delta^*(z,M).$$

If the infemum limit of the above quantity is infinite, there is nothing to prove, so assume that it is finite.

(**STEP 2b**) We claim the liminf of $\psi(T) := \mathbf{E}^M[\mathrm{Reg}(T)]\log^{-1}(T) = \sum_{z\in\mathscr{Z}} \lambda_\infty\mu_\infty(z)\Delta^*(z,M)$ with μ_∞ is a limit point of μ_T, hence an element of $\mathrm{Inv}(M/\mathscr{Z}_{**}(M))\cap\mathscr{P}(\mathscr{Z})$ and $\lambda_\infty < \infty$.

Indeed, let (T_n) a sequence of times such that $\psi(T_n)\to\liminf\psi(T)$. Because $\mathrm{Inv}(M/\mathscr{Z}_{**}(M))\cap\mathscr{P}(\mathscr{Z})$ is compact, we can assume that μ_{T_n} converges to a $\mu_\infty \in \mathrm{Inv}(M/\mathscr{Z}_{**}(M))\cap\mathscr{P}(\mathscr{Z})$ up to extracting a sub-sequence of (T_n). Now, by construction $\mu_T = 0$ on $\mathscr{Z}_{**}(M)$, hence $\mu_\infty = 0$ on $\mathscr{Z}_{**}(M)$ as well and by Lemma III.15, we must have $\sum_{z\notin\mathscr{Z}_*(M)} \mu_\infty(z) = c > 0$, i.e., μ_∞ puts a positive mass on sub-optimal pairs. We get:

$$\limsup_{n\to\infty} \lambda(T_n)c\Delta^*_{\min}(M) \leq \liminf\psi(T) < \infty$$

hence $\limsup\lambda(T_n) < \infty$. So, up to extracting a sub-sequence of (T_n) again, we can assume that $\lambda(T_n)$ converges to some $\lambda_\infty < \infty$.

(**STEP 2c**) We conclude that $\nu_\infty := \lambda_\infty\mu_\infty \in \mathrm{Inv}(M/\mathscr{Z}_{**}(M))$ is such that

$$\liminf_{T\to\infty} \frac{\mathbf{E}^M[\mathrm{Reg}(T)]}{\log(T)} = \sum_{z\in\mathscr{Z}} \nu_\infty(z)\Delta^*(z,M) \quad \text{and} \quad \forall M^\dagger \in \mathrm{Cnf}(M), \quad \sum_{z\in\mathscr{Z}} \nu_\infty(z)\mathrm{KL}_z\big(M||M^\dagger\big) \geq 1.$$

This concludes the proof. □

The lower bound of Theorem III.5 is tight and in Chapter 10, we present a strongly consistent algorithm with regret scaling as $K(M)\log(T)$. We conclude the section by showing that the contraction can be dropped in the lower bound.

Important remark. The quantity $K(M)$ actually depends on $\mathscr{M}$. In general, $\mathscr{M}$ is obvious in the context but whenever it is not, we write $K(M;\mathscr{M})$ to make the dependency clear.

8.4.1 From contracted invariant measures to invariant measures

While minors play a central role in the decomposition of the execution of consistent algorithm (Section 8.3), they can be dropped, slightly simplifying the lower bound and greatly simplifying the design of optimal planners. In fact, using that $\Delta^*(z)$ is null for $z \in \mathscr{Z}^{**}(M)$, invariant measures of the minor $M/\mathscr{Z}^{**}(M)$ **represented** by invariant measures of the initial model M, leading to:

Proposition III.6 (Removing the contraction)**.** *The regret lowerbound $K(M)$ is equal to:*

$$K(M) = \inf\left\{ \sum_{z\in\mathscr{Z}} \mu(z)\Delta^*(z) : \mu \in \mathrm{Inv}(M) \text{ and } \inf_{M^\dagger\in\mathrm{Cnf}(M)} \sum_{z\in\mathscr{Z}} \mu(z)\mathrm{KL}_z(M||M^\dagger) \geq 1 \right\}. \tag{III.7}$$

This is a direct consequence of the following remarkable result.

Lemma III.7. *Denote $u|_{\mathscr{Z}\setminus\mathscr{Z}^{**}(M)}$ the truncation of $u \in \mathbf{R}^{\mathscr{Z}}$ to pairs of $\mathscr{Z}\setminus\mathscr{Z}^{**}(M)$ and extend the notations to subsets of $\mathbf{R}^{\mathscr{Z}}$. Then $\mathrm{Inv}(M)|_{\mathscr{Z}\setminus\mathscr{Z}^{**}(M)} = \mathrm{Inv}(M/\mathscr{Z}^{**}(M))|_{\mathscr{Z}\setminus\mathscr{Z}^{**}(M)}$.*

Proof. Since $\mathrm{Inv}(M/\mathscr{Z}^{**}(M)) \supseteq \mathrm{Inv}(M)$, one inclusion is obvious and we focus on the other. Up to iterating the process on the communicating components of $\mathscr{Z}^{**}(M)$, we assume that $\mathscr{Z}_0 := \mathscr{Z}^{**}(M)$ induces a communicating model $M|_{\mathscr{Z}_0}$. Pick $[\mu] \in \mathrm{Inv}(M/\mathscr{Z}_0)$. We show that there μ such that $[\mu](z) = \mu(z)$ for $z \notin \mathscr{Z}_0$. Denote $\mathscr{S}_0 := \{s : \exists a, (s,a) \in \mathscr{Z}_0\}$ the states in which $\mathscr{Z}_0$ is rooted. In $[M] := M/\mathscr{Z}_0$, we have $[p]([\mathscr{S}_0]|z_0) = 1$ for all $z_0 \in \mathscr{Z}_0$ so that we can assume that $[\mu](z_0) = 0$ without loss of generality. For $s_0 \in \mathscr{S}_0$, introduce

$$\alpha(s) := \sum_{a\in\mathscr{A}(s_0)} [\mu](s_0,a) - \sum_{z\in\mathscr{Z}} [\mu](z)p(s_0|z). \tag{III.8}$$

Observe that $\sum_{s_0\in\mathscr{S}_0} \alpha(s_0) = 0$ since $[\mu]$ is an invariant measure of $[M] \equiv M/\mathscr{Z}_0$. It is enough to find $\mu_0 \in \mathbf{R}^{\mathscr{Z}_0}$ such that

$$\mu_0 \geq 0 \quad \text{and} \quad \forall s_0 \in \mathscr{S}_0,\ \sum_{z_0\in\mathscr{Z}_0} \mu_0(z_0)p(s_0|z_0) = \sum_{a_0:(s_0,a_0)\in\mathscr{Z}_0} \mu(s_0,a_0) + \alpha(s_0); \tag{III.9}$$

then $\mu \in \mathbf{R}^{\mathscr{Z}}$ given by $\mu(z) = \mu_0(z)$ if $z \in \mathscr{Z}_0$ and $[\mu](z)$ if $z \notin \mathscr{Z}_0$ will be solution. Assume that (III.9) has no solution. By Farkas' Lemma, there exists $\nu_0 \in \mathbf{R}^{\mathscr{S}_0}$ such that:

$$\sum_{s_0\in\mathscr{S}_0} \nu_0(s_0)\alpha(s_0) < 0 \quad \text{and} \quad \forall (s_0,a_0) \in \mathscr{Z}_0,\ \sum_{s_0'\in\mathscr{S}_0} \nu_0(s_0)\big(p(s_0'|s_0,a_0) - \mathbf{1}(s_0' = s_0)\big) \geq 0. \tag{III.10}$$

Let π_0 the policy picking its actions uniformly in $\mathscr{Z}_0$ from $\mathscr{S}_0$. The second condition can be rewritten as $p_{\pi_0}(M|_{\mathscr{Z}_0})\nu_0 \geq \nu_0$ so by induction, $\overline{p}_{\pi_0}(M|_{\mathscr{Z}_0})\nu_0 \geq \nu_0$. Because $\mathscr{Z}_0$ is communicating, $\overline{p}_{\pi_0}(M|_{\mathscr{Z}_0})$ has full support and $\nu_0 \in \mathbf{R}e$. But $\sum_{s_0\in\mathscr{S}_0} \alpha(s_0) = 0$ so $\sum_{s_0\in\mathscr{S}_0} \nu_0(s_0)\alpha(s_0) = 0$, contradicting (III.10). □

8.5 Examples and links to existing results

The statements of the lower bound in its two forms (Theorem III.5 and Proposition III.6) are a bit dry. In this section, its consequences are investigated and we discuss links with the existing literature. We explore how the informational constraints can be decomposed, and how the structure of the minor $M/\mathscr{Z}^{\star\star}(M)$ tells how easy exploration may conceptually be.

8.5.1 Example: Multi-armed bandits

Recall that multi-armed bandits are nothing less than state-less Markov decision processes, with pair space of the form $\mathscr{Z} = \{1\} \times \{1, \dots, A\}$. Let $\mathscr{M}$ the space of all Markov decision processes with Bernoulli rewards and pair space $\mathscr{Z}$ and fix $M \in \mathscr{M}$.

An optimal policy is a choice of optimal action (or arm), and the gain of a policy $\pi(1) = a$ is equal to $r(a) \equiv r(1,a)$ and the Bellman gap of $z \in \mathscr{Z}$ is $\Delta^\star(z) = g^\star(M) - r(z)$. Any contraction of a multi-armed bandit is a multi-armed bandit hence $\mathrm{Inv}(M/\mathscr{Z}^{\star\star}) = \mathbf{R}_+^{\mathscr{Z}}$. Moreover, confusing models $M^\dagger$) are precisely models such that $\exists z \notin \mathscr{Z}^{\star\star}$ such that $r^\dagger(z) > \max_{z'} r(z') = g^\star(M)$. Combined, Theorem III.5 can be rewritten as:

$$K(M;\mathscr{M}) = \inf\left\{ \sum_{z\in\mathscr{Z}} \mu(z)\Delta^\star(z;M) : \mu \in \mathbf{R}_+^{\mathscr{Z}} \text{ and } \forall M^\dagger \in \mathrm{Cnf}(M), \sum_{z\in\mathscr{Z}} \mu(z)\mathrm{kl}(r(z)||r^\dagger(z)) \geq 1 \right\}.$$

Now, remark that any $M^\dagger \in \mathrm{Cnf}(M)$ can be changed into a $M^\ddagger \in \mathrm{Cnf}(M)$ where M and $M^\ddagger$ only differ at one pair (the suboptimal pair of M that is made optimal in $M^\ddagger$). Using the continuity of the function $r^\dagger(z) \mapsto \mathrm{kl}(r(z)||r^\dagger(z))$ on $(0,1)$, we retrieve closed form of the regret lower bound of Lai and Robbins (1985):

$$K(M;\mathscr{M}) = \sum_{z\in\mathscr{Z}} \frac{\Delta^\star(z;M)}{\mathrm{kl}(r(z)||\max(r))}. \tag{III.11}$$

We also retrieve the interior condition "$r < 1$". Obviously, we have restricted the analysis to Bernoulli rewards models, but this result could be generalized to more general reward distributions because the technique does not really rely on the Bernoulli nature of reward distributions. For more general lower bounds on multi-armed bandits, see Honda and Takemura (2015) for example.

Several observations can be made regarding (III.11).

(1) The regret lower bound has a closed-form expression that can easily be evaluated.
(2) Exploration constraints are trivial in multi-armed bandits (i.e., $\mathrm{Inv}(M/\mathscr{Z}^{\star\star}) = \mathbf{R}_+^{\mathscr{Z}}$).
(3) Information constraints are pair-wisely decoupled (indexed) and are equivalent to $|\mathscr{Z}|$ linear constraints on μ, all of the form $\mu(z) \geq (\mathrm{kl}(r(z)||\max(r)))^{-1}$.

The properties (1-3) are not always satisfied. As a matter fact, (1) and (3) can break by considering $K(M;\mathscr{M}')$ for $\mathscr{M}' \subsetneq \mathscr{M}$ (**structured** bandit problems). Meanwhile, (2) only depends on M rather than $\mathscr{M}$. All together, these properties are arguably what makes multi-armed bandits significatively easier to learn than Markov decision processes. However, multi-armed bandits are not the only classes of model spaces with such properties.

8.5.2 Example: (Optimally) Recurrent models, or navigation-free models

The property (2) above, stating that $\mathrm{Inv}(M/\mathscr{Z}^{\star\star}) = \mathbf{R}_+^{\mathscr{Z}}$, only depend on M and makes it **navigation-free** models, because the minor $M/\mathscr{Z}^{\star\star}$ is a bandit. Such models are those such that

once the planner has correctly identified $\mathscr{Z}^{\star\star}$, navigating the environment causes no trouble anymore, because any state can be reached from any other with zero cost. This motivates the definition below.

Definition III.5. *We say that a model M is **optimally recurrent** if there exists a gain optimal policy π^* whose recurrent state is $\mathscr{S}$, or equivalently, if $\mathscr{Z}^{\star\star}(M)$ covers all the states of M.*

By definition, M is optimally recurrent if, and only if $M/\mathscr{Z}^{\star\star}$ is state-less. For instance, ergodic models are optimally recurrent. The navigation and information constraints can be heavily simplified when the model is optimally recurrent. Given $M \in \mathscr{M}$ an optimally recurrent model and $(s,a) \in \mathscr{Z}^{-}(M)$, denote:

$$C(M,s,a) := \inf_{\tilde{r}_{s,a},\tilde{p}_{s,a}} \left\{ \mathrm{KL}\big(r_{s,a} \| \tilde{r}_{s,a}\big) + \mathrm{KL}\big(p_{s,a} \| \tilde{p}_{s,a}\big) : \tilde{r}(s,a) + \tilde{p}(s,a)h^* > g^*(s) + h^*(s) \right\} \quad \text{(III.12)}$$

where g^* and h^* are respectively the optimal gain and bias vectors of M.

Proposition III.8 (Lower bound for optimally recurrent models)**.** *Assume that $\mathscr{M}$ the space of all models with state-action space $\mathscr{Z}$. If $M \in \mathscr{M}$ is an optimally recurrent model, then $K(M)$ is equal to:*

$$K(M) = \sum_{(s,a) \notin \mathscr{Z}^{\star\star}(M)} \frac{\Delta^*(s,a)}{C(M,s,a)}. \quad \text{(III.13)}$$

In particular, when $M \in \mathscr{M}$ is an optimally recurrent model, then the navigation constraints become trivial ($\mu \in \mathbf{R}_+^{\mathscr{Z}}$) and the information constraints are decoupled along sub-optimal pairs. This bound strictly generalizes the work of Agrawal et al. (1988); Burnetas and Katehakis (1997) that where specific to ergodic models. It also generalizes (III.11).

Proof. When M is optimally recurrent, $M/\mathscr{Z}^{\star\star}(M)$ is a single-state Markov decision process, hence $\mathrm{Inv}(M/\mathscr{Z}^{\star\star}(M)) = \mathbf{R}_+^{\mathscr{Z}}$ meaning that the navigation constraints are trivial. We now simplify the information constraints using a policy improvement argument which is similar to Burnetas and Katehakis (1997). Let $\pi^* \in \Pi^*(M)$ with recurrent class $\mathscr{S}$ and pick a confusing model $M^\dagger \in \mathrm{Cnf}(M)$. We have $(s,\pi^*(s)) \in \mathscr{Z}^{\star\star}(M)$ for all $s \in \mathscr{S}$, and because M and $M^\dagger$ coincide on $\mathscr{Z}^{\star\star}(M)$, it follows that the gain, bias, reward and kernel of π^* are preserved in $M^\dagger$. Yet π^* is not gain-optimal in $M^\dagger$, hence is not Bellman optimal in $M^\dagger$ (Proposition I.4). Accordingly, there must exists $(s,a) \in \mathscr{Z}$ such that:

$$r^\dagger(s,a) + p^\dagger(s,a)h^*(M) > g^*(s,M) + h^*(s,M).$$

Meanwhile, all states are recurrent under π^* which is optimal in M, hence (1) $\mathscr{Z}^{\star\star}(M) = \mathscr{Z}^{\star}(M)$ and (2) it is a fixpoint of the Bellman operator of M, in particular $r(s,a) + p(s,a)h^*(M) \leq g^*(s,M) + h^*(s,M)$. By (1), M and $M^\dagger$ coincide on $\mathscr{Z}^{\star}(M)$ so invoking (2), we must have $(s,a) \notin \mathscr{Z}^{\star}(M)$. In the end, we obtain:

$$K(M) = \inf\left\{ \sum_{z \in \mathscr{Z}} \mu(z)\Delta^*(z) : \mu \in \mathbf{R}_+^{\mathscr{Z}} \text{ and } \forall (s,a) \in \mathscr{Z}^{-}(M),\ \mu(s,a)C(M,s,a) \geq 1 \right\}$$

of which the solution is obvious. □

We recover a regret lower-bound that is in closed form, navigation free and with pair-wisely decoupled information constraints. In other words, the regret lower bound of optimally recurrent models is morally the same than than bandits'. Although the regret lower bound is navigation

free, the model itself is not navigation free because not every policy is recurrent, and an efficient planner has to be careful of that fact. If the model is recurrent however, then all states are visited independently of the chosen actions. This makes recurrent models essentially similar to bandits, and the design of efficient planners essentially similar in the two model classes.

On a side note, efficient planners for recurrent model are harder to design than those for bandits, because the Bellman equations is differently complex on the two classes. For bandits, the optimal bias vector is $h^* = 0$. Although gain and bias optimalities coincide when M is recurrent, the optimal bias vector is non-trivial and the Bellman gaps depend on the bias. Correctly estimating the bias requires a careful technique and carelessly trying to estimate h^* with few data can lead to disastrous performance. Therefore, although the planners provided by Burnetas and Katehakis (1997) and Pesquerel and Maillard (2022) (`IMED-RL`) are indexed algorithms that are reminiscent of known planners for stochastic bandits (respectively Lai and Robbins (1985) and Honda and Takemura (2015) (`IMED`)), they make use of specialized mechanisms to estimate the optimal bias function and their analysis is more intricate than their bandit analogues.

8.5.3 Example: Fixed kernel spaces

In this last paragraph, we detail what insight can the lower bound of Theorem III.5 provide in the setting investigated by Ortner (2010); Saber et al. (2024); Tranos and Proutiere (2021): Deterministic transition models and more generally models where the kernel is known.

Definition III.6. *Fix a pair space $\mathscr{Z}$. A model space $\mathscr{M} \in \mathfrak{M}(\mathscr{Z})$ is called a **fixed kernel space** if all elements of $\mathscr{M}$ have the same transition kernel, i.e., $\forall M, M' \in \mathscr{M}$, $p = p'$. We say that it is **deterministic kernel space** if, in addition, $p(s'|s,a) \in \{0,1\}$ for every transition triplet (s,a,s').*

For deterministic kernel space, (Tranos and Proutiere, 2021, Theorem 1) provided a model dependent lower bound, that already make apparent the information and navigation constraints in a form that are similar to Theorem III.5, although the objective function is written differently. In our set of notations, their result is the following.

Theorem III.9 (Tranos and Proutiere (2021))**.** *Let $\mathscr{M}$ a deterministic kernel space. Then $K(M;\mathscr{M})$ is at least:*

$$\inf\left\{\sum_{z\in\mathscr{Z}}\mu(z)(g^*-r(z)) : \mu\in \mathrm{Inv}(M) \text{ and } \inf_{M^\dagger\in \mathrm{Cnf}(M;\mathscr{M})}\sum_{z\in\mathscr{Z}}\mu(z)\mathrm{KL}(M(z)||M^\dagger(z)) \geq 1\right\}. \tag{III.14}$$

This lower bound is a special case of the one provided in Theorem III.5, because

$$\begin{aligned}
\sum_{z\in\mathscr{Z}}\mu(z)(g^*-r(z)) &\overset{(*)}{=} \sum_{(s,a)\in\mathscr{Z}}\mu(s,a)(\Delta^*(s,a)+(p(s,a)-e_s)h^*) \\
&= \sum_{z\in\mathscr{Z}}\mu(z)\Delta^*(z) + \sum_{s'\in s} h^*(s')\sum_{(s,a)\in\mathscr{Z}}\mu(s,a)p(s'|s,a) - \sum_{s\in\mathscr{S}}h^*(s)\sum_{a\in\mathscr{A}(s)}\mu(s,a) \\
&\overset{(\dagger)}{=} \sum_{z\in\mathscr{Z}}\mu(z)\Delta^*(z) + \sum_{s'\in\mathscr{S}} h^*(s')\sum_{a\in\mathscr{A}(s')}\mu(s',a) - \sum_{s\in\mathscr{S}}h^*(s)\sum_{a\in\mathscr{A}(s)}\mu(s,a) \\
&= \sum_{z\in\mathscr{Z}}\mu(z)\Delta^*(z)
\end{aligned}$$

where $(*)$ follows from the Poisson equation and $(\dagger)$ uses that $\mu \in \mathrm{Inv}(M)$.

However, Tranos and Proutiere (2021) is specific to deterministic kernel spaces and do not provide general tightness guarantees of their lower bound (only on special instances). Still for deterministic spaces, Ortner (2010) provide a variant of UCRL2, named UCYCLE, with sub-optimal regret guarantees of order $O(\log(T))$. In the broader setting of fixed kernel spaces, the recent Saber et al. (2024) tries to go closer to the lower bound with IMED-KD, a strategy based on IMED, but this work does not aim at quantifying the gap between their regret guarantees and the regret lower bound; Actually, no lower bound where available for their setting upon publication. Their algorithm, however, tries indeed to track a lower bound. In an episodic manner, IMED-KD checks if it lacks information somewhere by considering every policy in a small pool of policy candidates Π_τ. For every policy in Π_τ, IMED-KD computes an index inspired from IMED Honda and Takemura (2015) that quantifies the likelihood that this policy is better than the current empirically optimal policy (up to rescaling), from which an informative policy is computed. The role of this policy is to gather the estimated missing information quickly. This policy is iterated until the algorithm considers that enough information has been gathered, then switches back to exploiting the current empirically optimal policy.

Two ideas are essential in the design of IMED-KD.

(1) The alternative set can always be decomposed **policy-wise**, by writing:

$$\mathrm{Alt}(M) = \bigcup_{\pi \notin \Pi^*(M)} \left\{ M^\dagger \in \mathrm{Alt}(M) : \pi \in \Pi^*(M^\dagger) \right\}.$$

Hence, one can check that enough information has been gathered on M by checking that enough information has been gathered for every sub-optimal policy. This decomposition is not specific to fixed kernel spaces, and in general, to every policy π, one can associate a **most confusing** model that is the harder to reject such that π is optimal. Yet, when the kernel is fixed, finding this most confusing model is much easier, making indexed planners like IMED-KD computationally affordable.

(2) There are too many policies. Checking that enough information has been gathered for every sub-optimal policy is difficult a priori, because $|\Pi|$ is exponentially large, hence the introduction of the policy pool Π_τ of "good" candidate policies. Saber et al. (2024) discusses how this pool may be chosen.

Overall, the design of IMED-KD takes into account a concern that we so far have dismissed: The tractability of the lower bound. This is the subject of the next chapter.

Chapter 9

Intractability of the lower bound

In this chapter, we discuss the computational difficulties related to the regret lower bound $K(M;\mathscr{M})$ of Theorem III.5. Recall that it is given as the solution of the optimization problem:

$$\inf_{\mu\in \mathrm{Inv}(M/\mathscr{Z}^{**})} \sum_z \mu(z)\Delta^*(z;M) \quad \text{such that} \quad \inf_{M^\dagger\in \mathrm{Cnf}(M)} \sum_z \mu(z)\mathrm{KL}(M(z)||M^\dagger(z)) \geq 1.$$

While the objective function is linear, the constraints are non-convex in general. In this chapter, we show that if the model space is discrete, the optimal value is computationally hard.

9.1 CRITICAL-MODEL: A NP-complete problem

A simple question that arises from the regret optimization problem (III.6) is, given a measure μ, is it easy to check the condition $\forall M^\dagger \in \mathrm{Cnf}(M), \sum_z \mu(z)\mathrm{KL}(M(z)||M^\dagger(z)) \geq 1$? The answer is no: The problem is coNP difficult, see the problem CRITICAL-MODEL thereafter. This problem is a sub-problem that asymptotically optimal planners must indirectly solve, in order to determine whether they lack information or not. This problem appears in the design of the nearly optimal algorithm provided in the next chapter.

CRITICAL-MODEL: *Given a space of MDPs $\mathscr{M}$, a reference model $M \in \mathscr{M}$ and a pair of scalars $\alpha, \beta \geq 0$, is there a $M^\dagger \in \mathscr{M}$ such that:*

$$\sum_z \mu(z)\mathrm{KL}(M(z)||M^\dagger(z)) < \alpha \text{ and } g^*(M^\dagger) > \beta, \tag{III.1}$$

where μ is an invariant probability measure of M?

To be formal, we have to specify how a space of MDPs may be fed to an algorithm. If $\mathscr{M}$ has polynomial number of models, and every element of it has polynomial size, then the enumeration of $\mathscr{M}$ is polysized – but in this kind of setting, one can show that CRITICAL-MODEL is P. Therefore, we have to focus on spaces $\mathscr{M}$ that cannot be enumerated in polynomial time. We assume that $\mathscr{M}$ is given by its state-action space, as well as a polynomially many linear constraints on its kernel and rewards. Then, we have the following result:

Theorem III.10. CRITICAL-MODEL *is NP-complete.*

Notation. In the proof below, we write $\mathrm{KL}_{M||M'}(z)$ for $\mathrm{KL}(M(z)||M'(z))$.

Proof. Proving that it is NP is not a problem, because the optimal gain of a MDP is the solution of a linear program, and so is its invariant measure Puterman (1994).

(**STEP 1**) To prove that the problem is NP-hard, it is reduced from the Knapsack Problem (KP). Recall that an instance of KP is given by a collection of n items of integer values $\{v_1, \dots, v_n\}$ and integer weights $\{w_1, \dots w_n\}$, as well as a capacity W and a value threshold V, both integers. The problem is to determine whether there exists $\mathscr{K} \subseteq [n]$ such that $\sum_{k\in\mathscr{K}} w_k \leq W$ and $\sum_{k\in\mathscr{K}} v_k \geq V$.

Fix $\epsilon, \sigma, \delta > 0$ to be tuned later on. Given an instance of KP, consider the MDP $\mathscr{M}$ whose structure is given by n (CHOOSE k) widgets connected in a ring fashion.

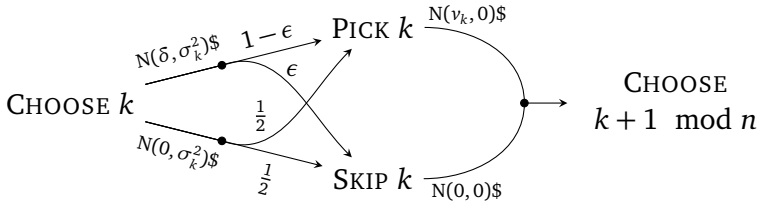


Figure 9.1.1: The (CHOOSE k) widget, where $\sigma_k^2 := \frac{\sigma^2}{w_k}$.

From the state (CHOOSE k) are two actions: The top action that is likely to go to (PICK k) that shall be referred to as action PICK, and the bottom action called SKIP. From every over state, there is a single action that denoted $*$. A (deterministic) policy of $\mathscr{M}$ is analogue to a subset $\mathscr{K} \subseteq \{1, ..., n\}$, written $\pi_{\mathscr{K}}$, which is given by $\pi_{\mathscr{K}}(\text{PICK}|\text{CHOOSE } k) := \mathbf{1}(k \in \mathscr{K})$. We get:

$$g(\pi_{\mathscr{K}}) = \frac{1}{2n}\left(\frac{1}{2}\sum_{k=1}^{n} v_k + \sum_{k\in\mathscr{K}}\left(\left(\tfrac{1}{2}-\epsilon\right)v_k + \delta\right)\right) = \frac{\|v\|_1}{4n} + \frac{1}{2n}\sum_{k\in\mathscr{K}}\left(\left(\tfrac{1}{2}-\epsilon\right)v_k + \delta\right). \tag{III.2}$$

(**STEP 2**) Every policy $\pi_{\mathscr{K}}$ can equivalently be seen as a *single-action* Markov decision process $M_{\mathscr{K}}$, i.e., the model of a policy over the state-space

$$\mathscr{S} := \{(\text{CHOOSE } k), (\text{PICK } k), (\text{SKIP } k) : k = 1, \dots, n\}.$$

The choice of an action is equivalently the choice of a kernel distribution. The set of stationary deterministic policies of $\mathscr{M}$, denoted $\Pi^{\text{MD}}(\mathscr{M})$, can therefore be seen as the set of Markov reward processes $\mathscr{M}^{\text{MD}} := \{M_{\mathscr{K}} : \mathscr{K} \subseteq \{1, \dots, n\}\}$. Provided that the parameters ϵ, σ, δ are polynomial in n, v, w, this (structured) set of mdps is described in polynomial size in n, v, w.

Consider $M_\varnothing \in \mathscr{M}^{\text{MD}}$. Because $\mathscr{M}^{\text{MD}}$ is a space of single-action MDPs, we don't make any distinction between a state and a pair of $M_{\mathscr{K}} \in \mathscr{M}^{\text{MD}}$. Now, we see that $g(M_\varnothing) = \frac{\|v\|_1}{4n}$ and the invariant measure μ of the unique policy of $M_\varnothing$ is:

$$\mu(\text{CHOOSE } k) = \tfrac{1}{2n} \text{ and } \mu(\text{PICK } k) = \mu(\text{SKIP } k) = \tfrac{1}{4n}.$$

Moreover, check that for $\mathscr{K} \subseteq \{1, \dots, n\}$, the only states such that $\text{KL}_{M_\varnothing \| M_{\mathscr{K}}}(s) \neq 0$ are (CHOOSE k) states, with:

$$\text{KL}_{M_\varnothing \| M_{\mathscr{K}}}(\text{CHOOSE } k) = \mathbf{1}(k \in \mathscr{K})\left(\log\left(\tfrac{1}{4\epsilon(1-\epsilon)}\right) + w_k\left(\tfrac{\delta}{\sigma}\right)^2\right).$$

Hence:

$$\sum_z \mu(z)\text{KL}_{M_\varnothing \| M_{\mathscr{K}}}(z) = \frac{1}{2n}\sum_{k\in\mathscr{K}}\left(\log\left(\tfrac{1}{4\epsilon(1-\epsilon)}\right) + w_k\left(\tfrac{\delta}{\sigma}\right)^2\right).$$

(**STEP 3**) We want (1) to be able to retrieve the value of $\sum_{k\in\mathscr{K}} v_k$ from $g(M_{\mathscr{K}})$; (2) to be able to retrieve the value of $\sum_{k\in\mathscr{K}} w_k$ from $\sum_z \mu(z)\text{KL}_{M_\varnothing \| M_k}(z)$. For simplicity and because it will eventually work with it, fix $\epsilon \equiv \frac{1}{4}$.

The condition (1) holds when $\delta = \frac{1}{16n}$. Indeed, then:

$$g(M_{\mathscr{K}}) = \frac{\|v\|_1}{4n} + \frac{1}{8n}\sum_{k\in\mathscr{K}}(v_k + 4\delta)$$
$$= \frac{1}{8n}\left(2\|v\|_1 + \sum_{k\in\mathscr{K}} v_k \pm \frac{1}{4}\right)$$

where $\pm\frac{1}{4}$ denotes an arbitrary quantity in the range of $[-\frac{1}{4}, \frac{1}{4}]$. Rearranging, we get $\sum_{k\in\mathscr{K}} v_k = 8ng(M_{\mathscr{K}}) - 2\|v\|_1 \pm \frac{1}{4} = [8ng(M_{\mathscr{K}}) - 2\|v\|_1]$ where $[\lambda]$ denotes the rounding operation (nearest integer).

The condition (2) is satisfied when

$$\sigma^2 = \frac{\delta^2}{4n\log\left(\frac{1}{4\epsilon(1-\epsilon)}\right)} \equiv \frac{1}{1024n^3\log\left(\frac{4}{3}\right)}.$$

Indeed, then we have

$$\sum_z \mu(z)\mathrm{KL}_{M_\emptyset\|M_{\mathscr{K}}}(z) = \frac{1}{2n}\sum_{k\in\mathscr{K}}\left(\log\left(\frac{1}{4\epsilon(1-\epsilon)}\right) + w_k\left(\frac{\delta}{\sigma}\right)^2\right)$$
$$= \frac{\delta^2}{2n\sigma^2}\sum_{k\in\mathscr{K}}\left(w_k + \frac{1}{4n}\right) = \frac{\delta^2}{2n\sigma^2}\left(\sum_{k\in\mathscr{K}} w_k \pm \frac{1}{4}\right).$$

Rearranging, we find $\sum_{k\in\mathscr{K}} w_k = [\frac{2n\sigma^2}{\delta^2}\sum_z \mu(z)\mathrm{KL}_{M_\emptyset\|M_{\mathscr{K}}}(z)]$ where $[\lambda]$ also denotes the rounding operation.

(**STEP 4**) Remark that this choice of ϵ, σ, δ is polynomial in the size of n. Following this remark, it should be clear that $\mathscr{M}^{\mathrm{MD}}$ can be encoded in polynomial size.[1] Finally set $\alpha = 2\log(4/3)(W + \frac{1}{3})$ and $\beta = \frac{1}{8n}(2\|v\| + V)$. Then, we claim that there is $M^\dagger \in \mathscr{M}^{\mathrm{MD}}$ such that

$$\sum_z \mu(z)\mathrm{KL}(M(z)||M^\dagger(z)) \leq \alpha, \text{ and } g^*(M^\dagger) \geq \beta \tag{III.3}$$

if, and only if the KP instance (v, w, V, W) has a solution.

This is just a commodity to check using the formulas established so far. If the KP instance has solution $\mathscr{K}$, then $M_{\mathscr{K}}$ is by construction a solution of (III.3), because

$$\sum_z \mu(z)\mathrm{KL}(M(z)||M^\dagger(z)) = \frac{\delta^2}{2n\sigma^2}\sum_{k\in\mathscr{K}}\left(w_k + \frac{1}{4n}\right) \leq 2\log\left(\frac{4}{3}\right)\left(W + \frac{|\mathscr{K}|}{4n}\right) < \alpha$$

and

$$g(M_{\mathscr{K}}) = \frac{1}{8n}\left(2\|v\|_1 + \sum_{k\in\mathscr{K}}\left(v_k + \tfrac{1}{4n}\right)\right) > \frac{1}{8n}(2\|v\|_1 + V) = \beta.$$

Conversely, if M_k is a solution of (III.3), then we have

$$\alpha = 2\log\left(\tfrac{4}{3}\right)\left(W + \tfrac{1}{3}\right) \geq \frac{\delta^2}{2n\sigma^2}\sum_{k\in\mathscr{K}}\left(w_k + \frac{1}{4n}\right) \geq 2\log\left(\tfrac{4}{3}\right)\sum_{k\in\mathscr{K}} w_k$$

hence $\sum_{k\in\mathscr{K}} w_k \leq W + \frac{1}{3}$, so $\sum_{k\in\mathscr{K}} w_k \leq W$; and similarly

$$\beta = \frac{1}{8n}(2\|v\| + V) \leq \frac{1}{8n}\left(2\|v\| + \frac{1}{4} + \sum_{k\in\mathscr{K}} v_k\right)$$

so $\sum_{k\in\mathscr{K}} v_k \geq V - \frac{1}{4}$, so $\sum_{k\in\mathscr{K}} v_k \geq V$. □

[1] If you are not convinced, just check that you can efficiently generate $M_{\mathscr{K}}$ on your computer.

9.2 REGRET: Checking solutions is co-NP-hard

With the problem CRITICAL-MODEL being NP-complete, it does not sound good for the tractability of the regret bound, because computing it is essentially harder than CRITICAL-MODEL. Remark that the intractability of CRITICAL-MODEL does not prove the intractability of the regret lower bound. If μ, α, β are close to the values corresponding to the solution of regret optimization problem, is the problem still difficult? Yes, and computing the regret is difficult.

REGRET: *Given a space of MDPs $\mathscr{M}$, a reference model $M \in \mathscr{M}$ and a scalar $\rho \geq 0$, does there exists a $\mu \in \mathrm{Inv}(M/\mathscr{Z}^{**}(M))$ such that*

$$\sum_{z \in \mathscr{Z}} \mu(z)\Delta^*(z;M) \leq \rho \text{ and } \inf_{M^\dagger \in \mathrm{Cnf}(M;\mathscr{M})} \sum_{z \in \mathscr{Z}} \mu(z)\mathrm{KL}(M(z)||M^\dagger(z)) \geq 1? \tag{III.4}$$

We have the following result.

Theorem III.11. *Checking a solution of* REGRET *is co-NP complete.*

Proof. We provide a reduction from the co-Knapsack problem (co-KP), which is coNP-complete because KP is NP-complete. An instance of co-KP is given by a collection of n items of integer values $\{v_1, \ldots, v_n\}$ and integer weights $\{w_1, \ldots, w_n\}$, as well as a capacity and a value threshold V, both integers. The problem is to determine if, for all $\mathscr{K} \subseteq [n]$, we either have $\sum_{k \in \mathscr{K}} w_k \geq W$ or $\sum_{k \in \mathscr{K}} v_k \leq V$.

The reduction is very similar to CRITICAL-MODEL's. Fix $\epsilon, \sigma, \delta, \theta$ to be tuned later on and consider an instance of co-KP. Consider the MDP $\mathscr{M}$ whose structure is as given by Figure 9.2.1.

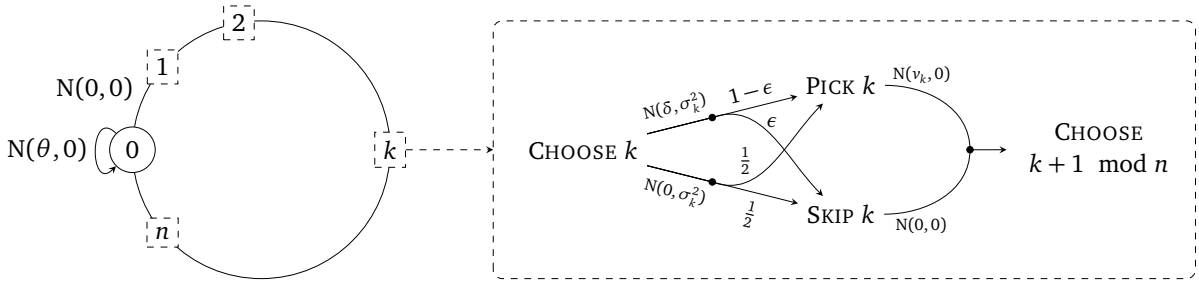


Figure 9.2.1: Embedding a knapsack problem in a Markov decision process.

The change regarding the reduction of CRITICAL-MODEL is the state (0), in between (Choose n) and (Choose 1). From (0) you can either loop with the action Loop scoring θ, or go to (Choose 1) with the action Cycle scoring 0, hence entering the big cycle. The state (0) is a special state. From the state (Choose k) are two actions: The top action that is likely to go to (Pick k) that we shall refer to as action Pick, and the bottom action that we shall call Skip. From every over state, there is a single action that denoted $*$. The special policy looping on (0) is denoted π^* and will model the optimal policy later on. The other (deterministic) policies of $\mathscr{M}$ are analogue to a subset $\mathscr{K} \subseteq \{1, ..., n\}$, written $\pi_{\mathscr{K}}$, and are given $\pi_{\mathscr{K}}(\text{Pick}|\text{Choose } k) := \mathbf{1}(k \in \mathscr{K})$ with $\pi(\text{Cycle}|0) = 1$. We get:

$$g(\pi_{\mathscr{K}}) = \frac{1}{2(n+1)}\left(\frac{1}{2}\sum_{k=1}^{n} v_k + \sum_{k \in \mathscr{K}}\left(\left(\tfrac{1}{2} - \epsilon\right)v_k + \delta\right)\right) = \frac{\|v\|_1}{4(n+1)} + \frac{1}{2(n+1)}\sum_{k \in \mathscr{K}}\left(\left(\tfrac{1}{2} - \epsilon\right)v_k + \delta\right). \tag{III.5}$$

Every policy $\pi_{\mathscr{K}}$ can equivalently be seen as a single-action Markov decision process $M_{\mathscr{K}}$. The choice of an action is equivalently the choice of a kernel distribution. The set of stationary deterministic policies of $\mathscr{M}$, denoted $\Pi^{\text{MD}}(\mathscr{M})$, can therefore be seen as the set of Markov reward processes $\mathscr{M}^{\text{MD}} := \{M_{\mathscr{K}} : \mathscr{K} \subseteq \{1,\dots,n\}\}$. Now, we see that $g(M_\varnothing) = \frac{\|v\|_1}{4(n+1)}$ and the invariant measure $\mu_\varnothing$ of the unique policy of $M_\varnothing$ is:

$$\mu_\varnothing(0) = \tfrac{1}{n+1},\ \mu_\varnothing(\textsc{Choose}\ k) = \tfrac{1}{2(n+1)},\ \text{and}\ \mu_\varnothing(\textsc{Pick}\ k) = \mu_\varnothing(\textsc{Skip}\ k) = \tfrac{1}{4(n+1)}.$$

Moreover, check that:

$$\sum_z \mu_\varnothing(z)\mathrm{KL}_{M_\varnothing \| M_{\mathscr{K}}}(z) = \frac{1}{2(n+1)} \sum_{k\in\mathscr{K}} \Big(\log\Big(\tfrac{1}{4\epsilon(1-\epsilon)}\Big) + w_k\big(\tfrac{\delta}{\sigma}\big)^2\Big).$$

We find the values:

$$\epsilon = \frac{1}{4},\ \delta = \frac{1}{16n},\ \sigma^2 = \frac{\delta^2}{4(n+1)\log\big(\frac{4}{3}\big)},\ \theta = \frac{2\|v\|_1 + V}{8(n+1)},\ \text{and}\ \rho = \frac{V}{16\log\big(\frac{4}{3}\big)W}.$$

Consider $\mathscr{M}_*^{\text{MD}}$ the copy of M^{MD} with each element augmented with the action Loop at 0, scoring $\mathrm{N}(\theta, 0)$ and pick the reference model $M_\varnothing \in \mathscr{M}^{\text{MD}}$ (In abuse of notations, we write the elements of $\mathscr{M}^{\text{MD}}$ and $\mathscr{M}_*^{\text{MD}}$ similarly because the two sets are obviously isomorphic, so $M_\varnothing$ contains the policies $\pi_\varnothing$ and π_*). We show that the initial co-KP problem is reduced to the REGRET instance $(\mathscr{M}_*^{\text{MD}}, M_\varnothing, \rho)$.

First, remark that π^* is the optimal policy of $M_\varnothing$. Then, We show that given $\mu \in \mathrm{Inv}(M_\varnothing/\mathscr{Z}^{**})$ such that $\sum_z \mu(z)\Delta^*(z; M_\varnothing) \le \rho$, we have

(1) $\sum_{k\in\mathscr{K}} v_k \le V$ if, and only if $g(M_{\mathscr{K}}) > \theta$, i.e., $M_{\mathscr{K}} \in \mathrm{Cnf}(M_\varnothing; \mathscr{M}_*^{\text{MD}})$; and
(2) $\sum_{k\in\mathscr{K}} w_k \ge W$ if, and only if $\sum_z \mu(z)\mathrm{KL}_{M_\varnothing \| M_{\mathscr{K}}}(z) \ge 1$.

We start with (1). If $\sum_{k\in\mathscr{K}} v_k \le V$, then

$$g(\pi_{\mathscr{K}}) > \frac{\|v\|_1}{4(n+1)} + \frac{V}{8(n+1)} = \theta.$$

Conversely, if $g(\pi_{\mathscr{K}}) > \theta$, then

$$\frac{2\|v\|_1 + V}{8(n+1)} < \frac{1}{8(n+1)}\left(2\|v\|_1 + \sum_{k\in\mathscr{K}}\left(v_k + \frac{1}{4n}\right)\right) \le \frac{2\|v\|_1 + \frac{1}{4} + \sum_{k\in\mathscr{K}} v_k}{8(n+1)},$$

so $\sum_{k\in\mathscr{K}} v_k \ge V - \frac{1}{4}$, so $\sum_{k\in\mathscr{K}} v_k \ge V$.

For (2), first remark that the only positive Bellman-gap of $M_\varnothing$ is at the state-action pair $(0, \textsc{Cycle})$ with $\Delta^*((0, \textsc{Cycle}); M_\varnothing) = \frac{V}{8}$. Moreover, every element of $\mathrm{Inv}(M_\varnothing/\mathscr{Z}^{**})$ is of the form $c\mu_\varnothing$ where $c > 0$. So, having $\sum_z \mu(z)\Delta^*(z; M_\varnothing) \le \rho$ means that $\mu = c\mu_\varnothing$ with $c \le \frac{8\rho}{V} = (2\log(\frac{4}{3})W)^{-1}$. With this in mind, if $\sum_{k\in\mathscr{K}} w_k \ge W$, then

$$\sum_z \mu(z)\mathrm{KL}_{M_\varnothing \| M_{\mathscr{K}}}(z) = \frac{c\delta^2}{2(n+1)\sigma^2} \sum_{k\in\mathscr{K}}\left(w_k + \frac{1}{4n}\right) \ge \frac{2\log\big(\frac{4}{3}\big)W}{2\log\big(\frac{4}{3}\big)W} \ge 1.$$

Conversely, if $\sum_z \mu(z)\mathrm{KL}_{M_\varnothing \| M_{\mathscr{K}}}(z) \ge 1$, then

$$1 \le \frac{c\delta^2}{2(n+1)\sigma^2} \sum_{k\in\mathscr{K}}\left(w_k + \frac{1}{4n}\right) \le \frac{2\log\big(\frac{4}{3}\big)\Big(\sum_{k\in\mathscr{K}} w_k + \frac{|\mathscr{K}|}{4n}\Big)}{2\log\big(\frac{4}{3}\big)W}$$

so $\sum_{k\in\mathcal{K}} w_k \geq W - \frac{1}{4}$, so $\sum_{k\in\mathcal{K}} w_k \geq W$.

We readily obtain that: "every $\mathcal{K} \subseteq [n]$ satisfies either $\sum_{k\in\mathcal{K}} w_k \geq W$ or $\sum_{k\in\mathcal{K}} v_k \leq V$" is equivalent to $\mu \equiv (2W\log(\frac{4}{3}))^{-1}\mu_\varnothing$ satisfying:

$$\forall M^\dagger \in \mathrm{Cnf}(M_\varnothing; \mathscr{M}_*^{\mathrm{MD}}),\ \sum_z \mu(z)\mathrm{KL}_{M_\varnothing \| M_{\mathcal{K}}}(z) \geq 1,$$

and this μ is the unique $\mu \in \mathrm{Inv}(M_\varnothing/\mathscr{Z}^{**})$ such that $\sum_z \mu(z)\Delta^*(z; M_\varnothing) = \rho$. □

9.3 Discussion of the result

If we focus on convex model classes, the proof of intractability fails. The previous reduction doesn't work anymore, because by taking the convex hull of $\mathscr{M}^{\mathrm{MD}}$, we obtain a space that is very close to the space of randomized policies of $\mathscr{M}$ instead of deterministic ones, which is essential in the proof. Also, the rational relaxed Knapsack problem is a linear program, so is solvable in polynomial time – but the optimization problem related to $\mathscr{M}^{\mathrm{MR}}$ is not exactly the relaxation of the reduced KP, hence we cannot easily claim that $K(M; \mathrm{Conv}(\mathscr{M}^{\mathrm{MD}}))$ is tractable. It is only natural to raise the following question.

> **Open problem.** *If $\mathscr{M}$ is convex,[a] does* REGRET *remain computationally difficult?*
>
> [a] e.g., a polyhedron.

I conjecture that it is even harder, but then the problem must be reduced to smooth computational problems.

Remark that for many model classes, $K(M; \mathscr{M})$ is tractable. This is the case if $\mathscr{M}$ is a space of bandit or, more generally, a space of optimally recurrent models (see Definition III.5). If $\mathscr{M}$ is a fixed kernel space, then $\inf_{M^\dagger \in \mathrm{Cnf}(M;\mathscr{M})} \sum_z \mu(z)\mathrm{KL}(M(z)||M^\dagger(z))$ is a convex function of μ as the infemum of linear functions. This infemum is then computed with gradient descent, because $\mathrm{Cnf}(M; \mathscr{M})$ is a half space and the objective function is convex in $M^\dagger$ when $\mathscr{M}$ is a fixed kernel space. However, when optimizing in μ, one must minimize a linear function subjected to concave constraints. Such problems are hard in general, because their dual problems correspond to convex maximization subjected to linear constraints. However, approximating the solution may be possible. This is left for future work.

Chapter 10

ECoE: A nearly asymptotically optimal algorithmic scheme

In Chapter 9, we have shown that the lower bound provided by Theorem III.5 is intractable in general. Disregarding any concern of computability, is the bound even tight? Is there a consistent algorithm achieving a regret upper bound that matches the lower bound? The answer is: Yes.

10.1 The Exploration-CoExploration-Exploitation trilemma and ECoE

The algorithm is inspired from the lower bound. One of the important takeaways of the lower bound is that exploration must be randomized, or behave like a randomized policy in the long run at least. One of the noticeable gamble of the lower bound is the assumption that the planner will be able to have nearly perfect information uniformly on $\mathscr{Z}^{**}$. It is by no means obvious. In many models, $\mathscr{Z}^{**}$ is split into several components and sub-optimal pairs must be played to go from one component to another. For instance, in the literature on minimax regret (see Part II), most algorithms reduce their number of episodes to amortize the cost endured by switching policies. Yet, the lower bound of Theorem III.5 dismisses this extra navigational difficulty.

Perhaps one of the most important takeaways of what is to come downstream, is that this **switching cost** is not due to exploration. Exploration is about visiting sub-optimal pairs to make sure that optimal pairs are correctly guessed, and is embodied by the **exploration measure** μ that appears in (III.7). This exploration measure is generally fully supported, hence corresponds to an invariant measure of a unique and fully supported randomized policy, that we call the **exploration policy**. By being fully supported, this policy is recurrent and induces no "switching cost" upon playing. Switching costs are due to something else. It cannot be exploitation either, because exploitation is about iterating the optimal policy, and by definition this never increases the first order regret (unless the optimal policy is wrongly guessed, but this is another matter).

Switching costs are due to something in between exploration and exploitation, that we refer to as **co-exploration**. Co-exploration is about visiting, or traveling to, a recurrent class of optimal pairs over which information is estimated to be lacking. Actually, the current lower bound (III.7) claims that the cost due to co-exploration is sub-logarithmic, hence asymptotically negligible. Hence co-exploration must be managed by a dedicated mechanism, whose purpose is to make sure that enough information is gathered (uniformly enough) on $\mathscr{Z}^{**}$. Upon co-exploring, optimal pairs are played not to score maximally, but rather because of the possibility that they are wrongly estimated.

This **Exploration-CoExploration-Exploitation trichotomy** is new and specific to Markov decision processes. Following this observation, we suggest the algorithmic scheme ECoE to learn optimally.

Algorithm III.1 The ECoE framework.

for episodes $k = 1, 2, \dots$ **do**
  Estimate optimal pairs $\mathscr{Z}_t^{**}$ out of observations;
  Update exploitation π_t^+ and exploration π_t^- policies;
  if S_t is not recurrent under π_t^+ or lacking information on $\mathscr{Z} \setminus \mathscr{Z}_t^{**}$ **then**
    Explore with π_t^- one step; ▷ *Vanilla exploration*
  else if lacking information on $\mathscr{Z}_t^{**}$ **then**
    if one component $\mathscr{Z}_t^c$ of $\mathscr{Z}_t^{**}$ is critically sub-visited and $S_t \notin \mathscr{S}(\mathscr{Z}_t^c)$ **then**
      Explore with π_t^- one step; ▷ *Exploration triggered by co-exploration*
    else
      CoExplore with π_t^+ until regeneration; ▷ *Exploitation triggered by co-exploration*
    end if
  else
    Exploit with π_t^+ until regeneration; ▷ *Vanilla exploitation*
  end if
end for

All the components are intentionally evasive and will be made precise in time. What Algorithm III.1 does is the following. First, it estimates the optimal pairs to compute good exploitation and exploration policies by approximating the solution of (III.7). Then, it goes through a few tests to decide whether it should explore, co-explore or exploit. It tests first if exploration is mandatory, in which case it uses the exploration policy obtained by estimating the regret lower bound (Theorem III.5). Otherwise, if co-exploration is mandatory, the algorithm behaves differently depending on whether information is lacking on the current recurrent class of optimal pairs or on another. If it is another recurrent class, it tries to travel to it using the exploration policy; otherwise it exploits using the exploitation policy. Otherwise, enough information has been gathered and the algorithm uses the exploitation policy. Remark that the blue block could be removed, because the algorithm does the same thing when it coexplores and when it exploits. In fact, co-exploration and exploitation are similar from a behavioral viewpoint, but are motivated by very different reasons. The distinction is crucial to the analysis of Algorithm III.1.

We now detail the components of Algorithm III.1.

10.2 Discontinuity of the lower bound

At the beginning of every episode, ECoE estimates the optimal pairs and the exploration policy by approximating the lower bound (III.7). But there is a major issue.

The lower bound $K(M; \mathscr{M})$ is intrinsically discontinuous in M.

$$\inf_{\mu \in \mathrm{Inv}(M/\mathscr{Z}^{**}(M))} \sum_{z \in \mathscr{Z}} \mu(z)\Delta^*(z) \quad \text{s.t.} \quad \forall M^\dagger \in \mathrm{Cnf}(M), \ \sum_{z \in \mathscr{Z}} \mu(z)\mathrm{KL}_z(M||M^\dagger) \geq 1. \tag{III.6}$$

The first discontinuity comes from the objective function $\mu \mapsto \sum_z \mu(z)\Delta^*(z, M)$, because the coefficients $\Delta^*(z, M)$ are not everywhere continuous with respect to M. The second and third discontinuities come from the discontinuity of $\mathscr{Z}^{**}(M)$ with respect to M, making $\mathrm{Inv}(M/\mathscr{Z}^{**}(M))$

and $\mathrm{Cnf}(M)$ discontinuous. In other words, the objective function, the navigation and information constraints are all discontinuous in M. Thankfully, the representation result of Proposition III.6, stating that elements $\mathrm{Inv}(M/\mathscr{Z}^{\star\star}(M))$ can be represented by elements of $\mathrm{Inv}(M)$ removes the difficulties due to navigation constraints. The other discontinuities cannot be avoided similarly and are addressed by relaxing the notion of optimality held by $\mathscr{Z}^{\star\star}(M)$ to $\mathscr{Z}^{\epsilon}_{\star\star}(M)$ (Section 10.3.1), force the uniformization of exploration measures with $\mathrm{Inv}_{\eta}(M/\mathscr{Z}^{\epsilon}_{\star\star}(M)$ (Section 10.3.2), and adapt the notions of confusing models and Bellman gaps to epsilonized variants $\mathrm{Cnf}^{\epsilon}(M)$ and $\Delta^{\epsilon}_{\star}(M)$ (Section 10.3.1).

We start by explaining the nature of these discontinuities on an example.

Important notations. In the following, we use the notations:

$$\|M'-M\| := \max_{z\in\mathscr{Z}}\big(\|r'(z)-r(z)\|_\infty + \|p'(z)-p(z)\|_1\big)$$
$$\mathrm{KL}(M\|M') := \max_{z\in\mathscr{Z}}\mathrm{KL}(M(z)\|M'(z)).$$

We further introduce $\|M'-M\|^* := \|M'-M\| + \mathbf{1}(M' \not\sim M)\cdot\infty$ and $\mathrm{KL}^*(M\|M') := \mathrm{KL}(M\|M')+\mathbf{1}(M'\not\sim M)\cdot\infty$, where $M'\sim M$ means that M and M' are mutually absolutely continuous with respect to each other, i.e., $\forall z\in\mathscr{Z}$, $\forall q\in\{r,p\}$, $q(z)\ll q'(z)\ll q(z)$.

So, the regret lower bound presented in Theorem III.5 and reported in (III.6) and (III.7) present discontinuities, making the estimation of $K(M;\mathscr{M})$ difficult if the learner has only access to a noisy version of M. Although these discontinuities are already present in the multi-armed bandit setting, they are not much of a big deal in those simpler settings and asymptotically optimal algorithms such as IMED, KLUCB or TS automatically deal with those discontinuities without any particular concern. Morally, this is because in bandits and more generally when the underlying model is recurrent, there isn't much of a difference between pulling an optimal pair and pulling a nearly optimal one. Specifically, the Bellman gaps are continuous at such models. When the underlying model is no more recurrent, the played pairs determine which states are visited, and discontinuities of $\mathscr{Z}^{\star\star}(M)$ induce discontinuities in the navigation constraints, i.e., of $\mathrm{Inv}(M/\mathscr{Z}^{\star\star}(M))$. Even though navigation constraints can be changed to $\mathrm{Inv}(M)$ via the representation result of Lemma III.7, it does not change the fact that an efficient planner is supposed to use $\mathscr{Z}^{\star\star}(M)$ to navigate the model more easily, hence has to tackle the discontinuities of $\mathscr{Z}^{\star\star}(M)$. An example is provided with Figure 10.2.1.

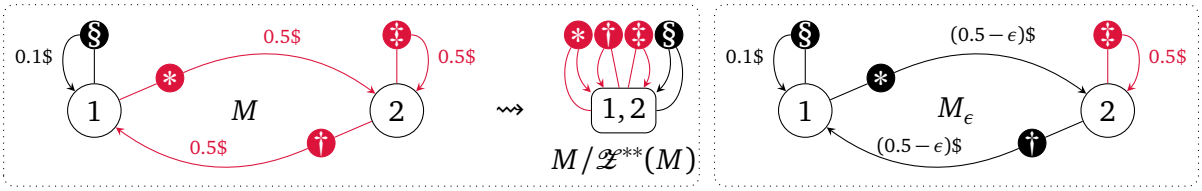


Figure 10.2.1: An example of discontinuity of the regret lower bound. The displayed transitions are deterministic and the labels represent the rewards' means. Optimal pairs are colored in red.

In Figure 10.2.1, we present two models, M and M_ϵ, that are statistically indistinguishable when $\epsilon\to 0$. However, the two models exhibit incompatible navigation constraints. On M, there is a recurrent optimal policy and $M/\mathscr{Z}^{\star\star}(M)$ is single state, hence navigation constraints are trivial. For M_ϵ however, we have $M_\epsilon/\mathscr{Z}^{\star\star}(M_\epsilon)\cong M_\epsilon$ and the navigation constraints are different. Moreover, the confusing set show discontinuities at M, since $M_\epsilon\notin\mathrm{Cnf}(M)$ but $M\in\mathrm{Cnf}(M_\epsilon)$. One can show that, for $\mathscr{M} := \{M_\alpha : \alpha\in(-0.5,0.5)\}$ where the reward are Bernoulli, the regret

lower bounds are:

$$K(M) = \frac{4}{10\ \mathrm{kl}(\frac{1}{10}, \frac{1}{2})} \quad \text{and} \quad K(M_\epsilon) \underset{\epsilon \to 0}{\sim} \frac{\epsilon}{\mathrm{kl}(0.5-\epsilon, 0.5)} \sim \frac{1}{2\epsilon} \text{ for } \epsilon > 0 \tag{III.1}$$

so that K is indeed discontinuous at $M \equiv M_0$. Because M and M_ϵ are indistinguishable, when the underlying model is M, no learner is capable of claiming that the current model is M with overwhelming confidence. Yet, when one is doubting about whether the model is M_ϵ or M, then the pairs $*, \dagger$ are nearly optimal, hence can arguably be used to navigate the model more easily while still scoring near optimally. After all, if the true model is a M_ϵ rather than M_0, then at some point $*$ and $\dagger$ will be rejected. From this discussion, we can raise the following observations.

(1) If the optimality of a pair is uncertain, one should consider it optimal; It makes navigation easier and if the pair is in fact sub-optimal, it will be rejected at some point.

A second observation adds to the list. As mentioned above, on the example provided by Figure 10.2.1, the space of confusing models displays a discontinuity at $\epsilon = 0$. This slightly refines observation (2).

(2) If one keeps track of the lower bound to determine which actions are to be played, one must address the discontinuities of the space of confusing models (information constraints).

Motivated by this example, we introduce a continuous regularized version of (III.6).

10.3 A continuous regularization of the lower bound

In this section, we overview the construction of the regularized lower bound.

10.3.1 Definitions: near optimal pairs, ϵ-confusing models and near optimal gaps

The set of optimal pairs $\mathcal{Z}^{**}(M)$ is not continuous in M hence $\Delta^*(M)$, $\mathrm{Inv}(M/\mathcal{Z}^{**}(M))$ and $\mathrm{Cnf}(M)$ may be ill-behaved in a neighborhood of M. From a learning viewpoint, this is at first sight an issue because even if $\hat{M}$ is arbitrarily close to M, one is not guaranteed to be able to track the solutions of the optimization problem (III.6) from $\hat{M}$ alone. To overcome this problem, we relax (III.6) by changing $\mathcal{Z}^{**}(M)$ for a more stable optimality notion: near-optimal pairs. A pair is **ϵ-optimal** ($z \in \mathcal{Z}^{\epsilon}_{**}(M)$) if it may be played infinitely often by some ϵ-optimal policy, see (III.2). Based on $\mathcal{Z}^{\epsilon}_{**}(M)$, we further epsilonize the notion of confusing model, by changing the role played by $\mathcal{Z}^{**}(M)$ with $\mathcal{Z}^{\epsilon}_{**}(M)$. Following the idea behind confusing models, an **ϵ-confusing model** is $M^\dagger \gg M$ such that $M^\dagger$ and M coincide on $\mathcal{Z}^{\epsilon}_{**}(M)$ and the optimal pairs of $M^\dagger$ are not all ϵ-optimal in M, see (III.3). Lastly, we define the **ϵ-gaps** as follows. Let $h^{\epsilon}_{*}(M)$ the supremum bias vector (for the product order) of all the $h_\pi(M)$ for $\pi \in \Pi$ satisfying $g^\pi(M) > g^*(M) - \epsilon e$, see (III.4). The ϵ-gaps are then defined using the Bellman equation associated to $(g^*(M), h^{\epsilon}_{*}(M))$, see (III.5).

$$\mathcal{Z}^{\epsilon}_{**}(M) := \{z \in \mathcal{Z} : \exists \pi \in \Pi, \exists \mu \in \mathrm{Inv}(\pi, M) : g^\pi(M) > g^*(M) - \epsilon e, \mu(z) > 0\} \tag{III.2}$$

$$\mathrm{Cnf}^{\epsilon}(M) := \left\{M^\dagger \gg M : M^\dagger = M \text{ on } \mathcal{Z}^{\epsilon}_{**}(M) \text{ and } g^*(M^\dagger) > g^*(M)\right\} \tag{III.3}$$

$$h^{\epsilon}_{*}(s, M) := \max\{h_\pi(s, M) : \pi \in \Pi \text{ such that } \forall s, g^\pi(s; M) \geq g^*(s; M) - \epsilon\} \tag{III.4}$$

$$\Delta_*^\epsilon(s,a,M) := \left[g^*(s,M) + h_*^\epsilon(s,M) - r(s,a,M) - p(s,a,M)h_*^\epsilon(M)\right]_+ \mathbf{1}((s,a) \notin \mathscr{Z}_{**}^\epsilon(M)) \tag{III.5}$$

The epsilonized notions $\mathscr{Z}_{**}^{\epsilon\epsilon}$, Cnf^ϵ and Δ_*^ϵ satisfy many interesting properties. Among many others, they properly address the discontinuity issues that their analogues $\mathscr{Z}^{**}$, Cnf and Δ^* suffer from, as specified by the result below.

Proposition III.12. *Let $M \in \mathscr{M}$ a communicating model. There exists $C_1, C_2, \epsilon_0 > 0$ such that, for all $\epsilon \in (0, \epsilon_0]$, if $C_1\|M' - M\|^* \leq \epsilon$, then*

(1) $\mathscr{Z}_{**}^{\epsilon\epsilon}(M') = \mathscr{Z}_{**}^{\epsilon\epsilon}(M)$;
(2) $\|\Delta_*^\epsilon(M') - \Delta^*(M)\|_\infty \leq C_2\|M' - M\|^*$;
(3) $\mathrm{Cnf}^\epsilon(M') = \{M^\dagger \ll M' : M^\dagger = M' \text{ on } \mathscr{Z}^{**}(M) \text{ and } g^*(M^\dagger) > g^*(M)\}$.

Remark that in Proposition III.12, a precision $\delta > 0$ is achieved on a neighborhood of M that does not depend on ϵ, provided that $\epsilon < \epsilon_0$. This ϵ_0 is moreover independent of the desired precision $\delta > 0$, meaning that the continuity of $\mathscr{Z}_{**}^{\epsilon\epsilon}$ and Δ_*^ϵ satisfies a cutoff phenomenon and continuity kicks in independently of ϵ provided that ϵ is not too large. The same property will hold for the uniformized regret lower bound (see Theorem III.13).

10.3.2 Uniformized exploration measures

For navigation constraints, we introduce $\mathrm{Inv}_\eta(M)$ the space of invariant measure of η-uniform policies, i.e., policies satisfying $\pi(a|s) \geq \eta$ for all $(s,a) \in \mathscr{Z}$, where $\eta > 0$ is a constant. That is,

$$\mathrm{Inv}_\eta(M) := \bigcup \{\mathrm{Inv}(\pi, M) : \pi \in \Pi \text{ and } \forall s, \forall a \in \mathscr{A}(s), \pi(a|s) \geq \eta\}. \tag{III.6}$$

The purpose of this uniformization is to force the optimal exploration μ to be fully supported, and the associated exploration policy to be η-uniform. By communicativity of the underlying model, any algorithm exploring with η-uniform policies is guaranteed to cover every pair during exploration, and this prevents Algorithm III.2 to overfit its exploration policy to the empirical data, that has a non-negligible probability to be off.

10.3.3 Regularized regret lower bound

The regularized regret lower bound combines a relaxation of the degree of optimality of pairs by dropping $\mathscr{Z}^{**}(M)$ to $\mathscr{Z}_{**}^{\epsilon\epsilon}(M)$ (Section 10.3.1) with the adapted notions of optimality gaps and confusing models, to the uniformization of exploration measures (Section 10.3.2) and adds a strongly convex regularization term to the objective function to make the minimizer unique.

Definition III.7 (Regularized regret lower bound)**.** *Fix $(\epsilon, \eta) > 0$. The **(ϵ, η)-regularized regret lower bound** for $M \in \mathscr{M}$ is the solution $K_\eta^\epsilon(M) \in [0, \infty]$ of the optimization problem:*

$$\inf_{\mu \in \mathrm{Inv}_\eta(M)} \sum_{z \in \mathscr{Z}} \mu(z)\Delta_*^\epsilon(z,M) + \eta\|\mu\|_2^2 \quad s.t. \quad \inf_{M^\dagger \in \mathrm{Cnf}^\epsilon(M)} \sum_{z \in \mathscr{Z}} \mu(z)\mathrm{KL}_z(M\|M^\dagger) \geq 1. \tag{III.7}$$

The regularized optimization problem (III.7) has (1) a unique solution, (2) is eventually a good proxy for the real optimization problem (III.6) and (3) has interesting continuity properties, see Theorem III.13 below.

Theorem III.13 (Properties of the regularized lower bound)**.** *Assume that the model space $\mathcal{M}$ is in product form $\mathcal{M} = \prod_{z\in\mathcal{Z}}[0,1]\times\mathcal{P}(z)$. The regularized regret lower bound has the following properties.*

(1) *The optimization problem (III.7) has a unique solution $\mu_\eta^\epsilon(M) \in \mathrm{Inv}_\eta(M)$;*
(2) *$\epsilon \mapsto K_\eta^\epsilon(M)$ is non-decreasing, and $K_\eta^\epsilon(M) \to K(M)$ when $\epsilon, \eta \to 0$ simultaneously;*
(3) *Assume that $M = (r,p)$ satisfies $0 < r(z) < 1$. There exists $\epsilon_0 > 0$ such that, for all $(M_n') \in \mathcal{M}^{\mathbf{N}}$ and $(\epsilon_n) \in (\mathbf{R}_+^*)^{\mathbf{N}}$ with $\epsilon_n \le \epsilon_0$, we have:*

$$K_\eta^{\epsilon_n}(M_n') \to K_\eta^0(M) \text{ and } \mu_\eta^{\epsilon_n}(M_n') \to \mu_\eta^0(M) \text{ when } \mathrm{KL}^*(M_n' \| M) + \tfrac{1}{\epsilon_n}\|M_n' - M\|^* \to 0.$$

An important take-away of this result is that the level of optimality relaxation given by ϵ_n must be large in front of $\|M_n' - M\|^*$, meaning that the estimation precision $\|M_n' - M\|^*$ and the suboptimality tolerance ϵ_n must vanish at different rates, roughly $\|M_n' - M\|^* = o(\epsilon_n)$.

Important remark. During the proof of Theorem III.13, it is shown that $K_\eta^\epsilon(M;\mathcal{M}) < \infty$ for all communicating M and model space $\mathcal{M}$, showing on the way that $K(M;\mathcal{M}) < \infty$.

10.4 Asymptotic regret guarantees of ECoE

With the regularized lower bound (Definition III.7) in hand, we can finally provide a complete description of ECoE. The algorithms works by approximating $\mathcal{Z}^{**}(M)$ by $\mathcal{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ where $\hat{M}_t$ is the model of empirical observations[1] and $\epsilon(t)$ is a vanishing sub-optimality tolerance. The exploitation policy is π_t^+ chosen as the uniform policy on $\mathcal{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$, and the exploration policy π_t^- as the policy induced by the optimal (regularized) exploration measure reaching $K_\eta^{\epsilon(t)}(\hat{M}_t)$ (see Theorem III.13), for some exploration parameter $\eta > 0$. The exploration and co-exploration are done by **generalized log-likelihood ratio tests** (similarly to Marjani et al. (2021); Marjani and Proutiere (2021)) that are carefully truncated to subsets of $\mathcal{Z}$.

The complete pseudo-code is provided with Algorithm III.2.

We have made the use of color to better isolate the three main parts of the algorithm. The first green block is relative to the algorithm's exploration and eventually corresponds to the dominant part of the regret. The second blue block corresponds to co-exyploration, at the interplay of exploration and exploitation, because the algorithm is iterating the empirically optimal policy for information purposes. The first and last red block corresponds to exploitation. A skeleton $\mathcal{Z}_t$ is used to truncate information tests, and although it is the same skeleton than in Burnetas and Katehakis (1997); Pesquerel and Maillard (2022), it plays a very different role in the analysis.

High level overview of the algorithm. The algorithm works by phases $k = 1, 2, 3, \ldots$ that are morally very short. At the beginning of a phase, it computes the nearly optimal pairs $\mathcal{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ and deduces an exploitation policy π_t^+ and an exploration policy π_t^-. For various reasons, the

[1]The **model of empirical observations** is the model $\hat{M}_t = (\mathcal{Z}, \hat{r}_t, \hat{p}_t)$ where $\hat{r}_t(z) := N_t(z)^{-1}\sum_{i=0}^{t-1}\mathbf{1}(Z_i = z)R_i$ is the average reward at z and $\hat{p}_t(s|z) := N_t(z)^{-1}\sum_{i=0}^{t-1}\mathbf{1}(Z_i = z, S_{i+1} = s)$ is the average number of observed transitions to s upon playing z.

Algorithm III.2 The ECoE algorithm.

Parameters: Exploration uniformization $\eta > 0$, ambient space $\mathscr{M}$.

Use

$$\mathrm{Alt}^{\epsilon(t)}(\hat{M}_t) := \left\{\hat{M}^\dagger \gg \hat{M}_t : \mathscr{Z}^{**}(\hat{M}^\dagger) \not\subseteq \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)\right\}. \tag{III.8}$$

Use near-optimality threshold $\epsilon(t) = \frac{1}{\log\log(t)}$. Use GLR overshoot $\delta(t) := \frac{1}{\log\log(t)} = \omega\left(\frac{\log\log(t)}{\log(t)}\right)$.

for episodes $k = 1, 2, \ldots$ **do**
  Set $t_k \leftarrow t$;
  Update exploitation policy $\pi^+_{t_k}$: uniform on $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ from $\mathscr{S}(\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t))$ and uniform elsewhere;
  Update exploration measure $\mu_{t_k} \leftarrow \mu_\eta^{\epsilon(t)}(\hat{M}_t)$, exploration policy $\pi^-_{t_k}(a|s) \propto \mu_t(s,a)$;
  Update skeleton $\mathscr{Z}_t \leftarrow \{z \in \mathscr{Z} : N_t(z) \geq \log^2(t)\}$;
  Update extended skeleton $\mathscr{Y}_t \leftarrow \mathscr{Z}_t \cup \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$;
  if S_t is not recurrent under π^+_t **then**
    Play A_t according to $\pi^+_t(S_t, -)$; EXPLORATION ($t \in \mathscr{T}^-$)
    $t \leftarrow t+1$, add $t-1$ in $\mathscr{T}^-$;
  else if $\exists M^\dagger \in \mathrm{Alt}^{\epsilon(t)}(\hat{M}_t)$ s.t. $M^\dagger|_{\mathscr{Y}_t} = \hat{M}_t|_{\mathscr{Y}_t}$ and $\sum_z N_t(z)\mathrm{KL}_z(\hat{M}_t||M^\dagger) \leq (1+\delta(t))\log(t)$ **then**
    Play A_t according to $\pi^-_t(S_t, -)$;
    $t \leftarrow t+1$, add $t-1$ in $\mathscr{T}^-$;
  else if $\exists M^\dagger \in \mathrm{Alt}^{\epsilon(t)}(\hat{M}_t)$ s.t. $M^\dagger|_{\mathscr{Z}_t} = \hat{M}_t|_{\mathscr{Z}_t}$ and $\sum_z N_t(z)\mathrm{KL}_z(\hat{M}_t||M^\dagger) \leq (1+\delta(t))\log(t)$ **then** COEXPLORATION ($t \in \mathscr{T}^\pm$)
    Split $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ into communicating components $\mathscr{Z}_t^1, \ldots, \mathscr{Z}_t^{m(t)}$;
    Let $\mathscr{Z}_t^{i(t)}$ the current component containing S_t;
    if $\log\min\{N_t(z) : z \in \mathscr{Z}_t^{i(t)}\} < 2\log\min\{N_t(z) : z \in \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)\}$ **then**
      Add $t \in \mathscr{T}_0^\pm$;
      repeat
        Play A_t according to $\pi^+_t(-|S_t)$;
        $t \leftarrow t+1$;
        if $S_t \notin \mathscr{S}(\mathscr{Z}_t^{i(t)})$ **then** add $t-1$ in $\mathscr{T}^!$ and **break**; ▷ transition discovery
        else add $t-1$ in $\mathscr{T}^\pm$;
      until $S_t = S_{t_k}$; ▷ regeneration
    else
      Play A_t according to $\pi^-_t(S_t, -)$;
      $t \leftarrow t+1$, add $t-1$ in $\mathscr{T}^-$;
    end if
  else
    Add t in $\mathscr{T}_0^+$; EXPLOITATION ($t \in \mathscr{T}^+$)
    repeat
      Play A_t according to $\pi^+_t(-|S_t)$;
      $t \leftarrow t+1$;
      if $S_t \notin \mathscr{S}(\mathscr{Z}_t^{i(t)})$ **then** add $t-1$ in $\mathscr{T}^!$ and **break**; ▷ transition discovery
      else add $t-1$ in $\mathscr{T}^+$;
    until $S_t = S_{t_k}$; ▷ regeneration
  end if
end for

exploration policy is chosen η-uniform. Then, the algorithm must decides whether (1) it explores, i.e., plays π_t^- once to gather information on presumed sub-optimal pairs; (2) it co-explores, i.e., plays π_t^+ to gain information on a component of the presumed nearly optimal pairs; (3) it exploits, i.e., plays π_t^+ to score as much as possible.

(0) DEFAULT EXPLORATION: By default, if the algorithm is not on a recurrent component of $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$, it iterates π_t^- and ends the phase. Otherwise, it goes through a bunch of tests to decide what to do.

(1) EXPLORATION: It first checks if there is a lack of information on sub-optimal pairs by running a generalized log-likelihood ratio test (GLR) that is **truncated** to $\mathscr{Z} \setminus \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$. Morally, it makes the assumption that the information on nearly optimal pairs is perfect, then does a classical GLR test. If the test is positive, it iterates π_t^- once and the phase ends. Otherwise, it deals with co-exploration.

(2) COEXPLORATION: The idea is to check if there is a lack of information on nearly optimal pairs. It is done by running a non-truncated GLR test, i.e., the assumption on the perfectness of information on $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ is dropped and a classical GLR test is ran. If the test is positive, the algorithm concludes that information is missing on $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ and that these pairs should be visited more. However, $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ may be split into several communicating components and the lacking information may be in a component to which the current state does not belong to. The algorithm does a test based on visit counts to decide whether the current component is worth co-exploring or if the lack of information is much more likely to come from another component. In the first case, it iterates π_t^+ until regeneration and ends the phase. In the second case, it wants to reach a more informationally critical component, and engages a travel by exploring: It iterates π_t^- once and the phase ends. The idea is to explore by the meantime, and reconsider once another component of $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ is reached. If the prior GLR test is negative, the algorithm deals with exploitation.

(3) EXPLOITATION: The algorithm iterates π_t^+ until regeneration.

(!) PANIC: When the algorithm iterates a policy until regeneration, this regeneration may never come if the algorithm is wrong about the support of transition kernels. In that case, at some point, the algorithm will observe a transition (S_t, A_t, S_{t+1}) it has never seen before. We call these time-instant **panic times**, and the algorithm deals with them by killing the current phase right away.

Theorem III.14. *Assume that the model space $\mathscr{M}$ is in product form $\mathscr{M} = \prod_{z \in \mathscr{Z}} [0,1] \times \mathscr{P}(z)$. Fix the parameters of Algorithm III.2 to $\eta > 0$. For all $M \in \mathscr{M}$ such that $0 < r(z) < 1$, the regret of Algorithm III.2 is asymptotically bounded by:*

$$\limsup_{T \to \infty} \frac{\mathbf{E}^M[\mathrm{Reg}(T)]}{\log(T)} \leq K_\eta^0(M). \tag{III.9}$$

For every fixed $\eta > 0$, Algorithm III.2 is consistent (because $K_\eta^0(M) < \infty$), see the remark following Theorem III.13). Since $\inf_{\eta > 0} K_\eta^0(M) = K(M)$ by Theorem III.13, this provides a family of consistent planners $((\pi_t^\eta)_t)$ indexed by $\eta > 0$ such that:

$$\forall M \in \mathscr{M}, \quad \inf_{\eta > 0} \limsup_{T \to \infty} \frac{\mathbf{E}^{(\pi_t^\eta), M}[\mathrm{Reg}(T)]}{\log(T)} = K(M). \tag{III.10}$$

Accordingly, the regret lower bound of Theorem III.5 is tight.

The proof of Theorem III.14 is difficult and requires many non-standard techniques.

It is deferred to the appendix.

10.5 Instantiations of ECoE

10.5.1 ECoE for multi-armed bandits

Although Algorithm III.2 is complex, it is rather similar to a blended version of two well-known methods when instantiated to bandit settings: MED Honda and Takemura (2010) and IMED Honda and Takemura (2015). First, if $\mathscr{M}$ is a class of bandits, then the co-exploration test may be ignored because the environment is stateless. In Section 8.5.1, we have shown that in the bandit setting, we have:

$$K(M) = \sum_{z \notin \mathscr{Z}^{**}} \frac{\Delta^*(z)}{\mathrm{kl}(r(z)||\max(r))}.$$

So, if $\epsilon, \eta > 0$ are small enough, $\mu_\eta^\epsilon(z; \hat{M}_t) \approx \mathrm{kl}(\hat{r}_t(z)||\max(\hat{r}_t))^{-1}$. Therefore, sub-optimal pairs are sampled proportionally to $1/\mathrm{kl}(\hat{r}_t(z)||\max(\hat{r}_t))$, and the visit counts of ECoE track the lower bound explicitly, which is in spirit similar to what MED does. To decide whether it should explore or not, the exploration GLR test of ECoE is morally asymptotically equivalent to:

$$\exists ? z \in \mathscr{Z} : \hat{r}_t(z) < \max(\hat{r}_t) \text{ and } N_t(z) < \log^2(t) \text{ and } N_t(z)\mathrm{kl}(\hat{r}_t(z)||\max(\hat{r}_t)) \leq \log(t)$$

Straight forward computations show that this rule is asymptotically equivalent to checking if there exists a suboptimal pair such that $N_t(z)\mathrm{kl}(\hat{r}_t||\max(\hat{r}_t)) + \log(N_t(z)) < \log(t)$, which is reminiscent of IMED's index. A version of ECoE specialized for bandits is reported with Algorithm III.3.

Algorithm III.3 ECoE for bandits.

for episodes $k = 1, 2, \ldots$ **do**
  Update $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) := \{z \in \mathscr{Z} : \hat{r}_t(z) \geq \max(\hat{r}_t) - \epsilon(t)\}$;
  Update exploration policy $\mu_\eta^{\epsilon(t)}(z; \hat{M}_t) \approx \frac{\mathbf{1}(z \notin \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t))}{\mathrm{kl}(\hat{r}_t(z)||\max(\hat{r}_t))}$ (projected to η-uniform measures);
  Update skeleton $\mathscr{Z}_t := \{z \notin \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) : N_t(z) < \log^2(t)\}$;
  if $\exists z \in \mathscr{Z}_t : N_t(z)\mathrm{kl}(\hat{r}_t(z)||\max(\hat{r}_t)) < (1 + \delta(t))\log(t)$ **then**
    Explore by playing $A_t \sim \mu_\eta^{\epsilon(t)}(-; \hat{M}_t)$;
  else
    Exploit by playing A_t uniformly in $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$;
  end if
end for

10.5.2 ECoE for recurrent models

Because recurrent models are very similar to bandits, ECoE can be simplified to something similar to Algorithm III.3. In Section 8.5.2, we have shown that for recurrent models, we have:

$$K(M) = \sum_{z \notin \mathscr{Z}^{**}(M)} \frac{\Delta^*(z)}{C(M, z)}$$

where $C(M, z)$ is defined as in (III.12). Similarly to the bandit setting, co-exploration can be ignored because the algorithm never has to change of nearly optimal component, hence co-exploration always plays the exploitation policy until regeneration. So, for recurrent models, ECoE is similar to Algorithm III.3 with $\mathrm{kl}(\hat{r}(z)||\max(\hat{r}_t))$ changed to $C(M, z)$. The obtained algorithm is very close to an index policy with an index similar in spirit to IMED-RL Pesquerel and Maillard (2022). The use of the skeleton $\{z : N_t(z) > \log^2(t)\}$ is however very different from

Burnetas and Katehakis (1997); Pesquerel and Maillard (2022). In these works, the skeleton is used as a region of the Markov decision process that is considered learned and where the gain and bias functions can be safely estimated. This fact is crucial in the analysis of Burnetas and Katehakis (1997); Pesquerel and Maillard (2022) because by computing the index out of quantities with strong approximation properties, the so-called index is guaranteed to be well-behaved. ECoE has no such safeguard against ill-behaved indexes, because the skeleton is used to truncate GLR tests rather than guaranteeing strong approximation properties. The possible ill-behavior of the index of ECoE and induced misplays are balanced by the η-uniformization of the exploration policy.

Algorithm III.4 ECoE for recurrent models.

for episodes $k = 1, 2, \dots$ **do**
  Update $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$;
  Update exploitation policy π_t^+ like in Algorithm III.2;
  Update exploration policy $\mu_\eta^{\epsilon(t)}(z; \hat{M}_t) \approx \frac{\mathbf{1}(z \notin \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t))}{C(\hat{M}_t, z)}$ (projected to η-uniform measures);
  Update skeleton $\mathscr{Z}_t := \{z \notin \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) : N_t(z) < \log^2(t)\}$;
  if $\exists z \in \mathscr{Z}_t : N_t(z) C(\hat{M}_t, z) < (1 + \delta(t)) \log(t)$ **then**
    Explore by playing $A_t \sim \mu_\eta^{\epsilon(t)}(-; \hat{M}_t)$;
  else
    Exploit by playing $A_t \sim \pi_t^+(-|S_t)$ until regeneration;
  end if
end for

Remark that we do not provide a closed form for $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ nor a way to compute it.

10.6 Future directions

So the lower bound of Theorem III.5 is optimal, meaning that asymptotically optimal learners shall have total reward scaling as

$$\mathbf{E}_s^M\left[\sum_{t=0}^{T-1} R_t\right] = T g^*(s) - K(M)\log(T) + \mathrm{o}(\log(T)).$$

Many directions are still to investigate.

First, ECoE only works if the hyper parameter η responsible for the uniformization of exploration is constant. If it is enough to show that the regret lower bound is optimal, we should make η a time-dependent vanishing quantity so that ECoE is truly asymptotically optimal. This would require to study the sensibility of $K_\eta^\epsilon(M)$ to η, hence to refine the study of the regularized lower bound further. I conjecture that taking $\eta(t) = 1/\log\log\log(t)$ is enough but this decay is extremely slow, and a precise analysis would help in that matter.

Second, ECoE cannot run in reasonable time. With the current definition of $\mathscr{Z}_{**}^{\epsilon}(M)$, it can be proved that the computation of $\mathscr{Z}_{**}^{\epsilon}(M)$ is linked to a NP-hard problem. The issue is thankfully secondary, and there are other ways to relax the notion of optimal pair, e.g., $\Delta^*(z) > -\epsilon$. It seems that this alternative definition makes the computation of Δ_*^ϵ polynomial. This being said, there is still the issue that the GLR tests performed by ECoE are coNP-hard. Indeed, the GLR tests are directly linked to the computation of critical models hence to the CRITICAL-MODEL problem (see Theorem III.10). To avoid these tractability issues, these information tests should be performed by a mechanism of another kind instead. Posterior sampling and bootstrap

methods are probably a fruitful direction in that matter. After that, we would have to tackle the computation of nearly-optimal exploration policies.

Yet, the lower bound is intractable in general, so is it only possible to make ECoE run in reasonable time? Overall, the question of whether there simply exists a tractable method that is asymptotically optimal is sound. However, the lower bound is asymptotic, so the existence of such an algorithm is not incompatible with the lower bound's intractability. Looking at the pseudo-code of ECoE, we observe that many hyper parameters are slowly vanishing functions. This is imposed by the various discontinuities of the lower bound. In practice, it means that T must be greater than a tower of exponentials of the diameter (at least $\exp(\exp(D))$) before the regret upper bound of Theorem III.14 gets close to the lower bound of Theorem III.5. So, by then, even if the algorithms makes $\Theta(1)$ per-step operations, its capability to approach the regret lower bound $K(M)\log(T)$ would not contradict the intractability of $K(M)$.

Talking about towers of exponentials raises an interesting question. Is this tower of exponentials specific to ECoE or to the lower bound? In other words, is the regret lower bound of Theorem III.5 "too" asymptotic in nature? This kind of concern leads to the study of regret guarantees for medium time ranges, that are usually done within the model independent framework. This is the "best of both worlds" of Bubeck and Slivkins (2012) that requires both minimax optimal and instance dependent optimal regret guarantees, see also Garivier et al. (2022). Regarding multi-armed bandits, the question of tower of exponentials is more deeply discussed in Belomestny et al. (2023), with its relation to the second order error in Sanov's theorem and its role in the regret analysis of Thompson Sampling. Because of the discontinuities of the lower bound for Markov decision processes however, it is not clear that the best of both worlds is even achievable for general communicating models.

Although medium time-ranges regret guarantees are usually done within the model independent framework, I am not completely convinced that this interpretation of minimax guarantees is pertinent. The asymptotic nature of Theorem III.5 is rather the product of an asymptotic assumption, which is strong consistency. One way of circumventing the issue is to refine the consistency assumption. This is done for example by Garivier et al. (2018), where the learner is assumed to have regret bounded by a function of the form $C(M)\log(T)$ where $C(M)$ has a structure inspired from UCB's regret upper bound (see Auer (2002)). They use finer information theoretic results to show that, for multi-armed bandits, the optimal regret is of the form $K(M)\log(T) - \Omega(\log\log(T))$. The correct multiplicative coefficient in front of the second order term $\log\log(T)$ is unknown to this day, and regret rates of such kinds have been achieved by IMED Honda and Takemura (2015) for reward distributions of bounded support. It is not clear that for communicating Markov decision processes, the second order term is indeed a negative $\log\log(T)$, because multi-armed bandits do not require co-exploration. Now, the direction opened by Garivier et al. (2018) is not the only possibility. Another way, that I believe is fruitful, is to utterly revoke the asymptotic nature of the consistency assumption and to opt for a non-asymptotic consistency assumption. When motivated by this idea, I see only one way to escape a minimax formulation: Bayesian priors and Bayesian formulations of the learning task.

Beyond this, there is the question of generalizing the model dependent lower bound beyond the communicating setting. The lower bound technique could probably be used to provide a lower bound for weakly communicating models without modifications. But like I said earlier, a lower bound without a proof of tightness as much less value. Regarding weakly communicating models, the issue is precisely that the design of an asymptotically optimal algorithm would have to tackle the transient nature of a part of the environment. It is not clear that ECoE will automatically work for weakly communicating Markov decision processes. Beyond weakly communicating models, the regret lower bound is subjected to depend on the initial state, hence the technique would have to take this into account.

Appendix of Chapter 10

10.A Technical results for the proof of Theorem III.5

In this section, we list a few technical results on which Theorem III.5 relies.

Lemma III.15 (Unavoidable sub-optimal pairs)**.** *Let $M \in \mathscr{M}$. There is $c > 0$ such that for all $\mu \in \mathrm{Inv}(M/\mathscr{Z}_{**}(M))$, if $\mu(z) = 0$ for all $z \in \mathscr{Z}_{**}(M)$, then $\sum_{z \notin \mathscr{Z}_*(M)} \mu(z) \geq c\|\mu\|_1$.*

Proof of Lemma III.15. Consider the reward vector $f(z) := \mathbf{1}(z \in \mathscr{Z}_*(M))$, then the model M' obtained by copying $M/\mathscr{Z}_{**}(M)$, removing the state-action pairs $\mathscr{Z}_{**}(M)$ and setting the reward function to f. The model M' is communicating because it is obtained by removing loops from $M/\mathscr{Z}_{**}(M)$, which is communicating as a contraction of a communicating model. The remaining pairs $\mathscr{Z}_*(M)$ are not forward closed in M', hence no policy of M' can have a recurrent component contained in $\mathscr{Z}_*(M) \cap \mathscr{Z}(M')$. Accordingly, the optimal gain of M' satisfies $\max(g^*(M')) < 1$. Set $c := 1 - \max(g^*(M'))$. □

Lemma III.16. *Consider $X_n \in [0,1]$ a family of r.v. with $\mathbf{E}[X_n|\mathscr{F}_n] = \mu$. Let $\hat{\mu}_n := \frac{1}{n}\sum_{k=1}^n X_k$ their empirical mean. Let N a random variable of support $\{0,1,\dots,T\}$ where $T \geq 1$ is a fixed scalar. Then, for all $\epsilon > 0$,*

$$\mathbf{E}[N(\hat{\mu}_N - \mu)] \leq \epsilon(\mathbf{E}[N] + \log(1+T)) + \sqrt{\mathbf{E}\Big[N \log\Big(2 \vee \tfrac{3\sqrt{1+N}\log(1+T)}{2\epsilon^3 \mathbf{E}[N]}\Big)\Big]} + 1.$$

Proof. Let $\delta > 0$ that shall be tuned later. By Hoeffding's Lemma, X_n is conditionally σ-subgaussian for $\sigma = \frac{1}{2}$. By a time-uniform Azuma-Hoeffding's inequality (Lemma I.22), for all $m \geq 1$,

$$\mathbf{P}\left(\exists m \leq n \leq T, |\hat{\mu}_n - \mu| \geq 2\sigma\sqrt{\frac{\log(T\sqrt{1+T})}{m}}\right) \leq \frac{1}{T}. \tag{III.11}$$

Setting $m = m_\epsilon := (\frac{2\sigma}{\epsilon})^2 \log(T\sqrt{1+T})$, we have $\mathbf{P}(\exists m_\epsilon \leq n \leq T, |\hat{\mu}_n - \mu| \geq \epsilon) \leq \frac{1}{T}$. The target expectation is split into two. Denoting $f(N) := N(\hat{\mu}_N - \mu)$, we write:

$$\mathbf{E}[N(\hat{\mu}_N - \mu)] = \mathbf{E}[f(N)\mathbf{1}(N \geq m_\epsilon)] + \mathbf{E}[f(N)\mathbf{1}(N < m_\epsilon)]. \tag{III.12}$$

We start by controlling $\mathbf{E}[f(N)\mathbf{1}(N \geq m_\epsilon)]$. By construction of m_ϵ, we have:

$$\begin{aligned}\mathbf{E}[f(N)\mathbf{1}(N \geq m_\epsilon)] &\leq \mathbf{E}[f(N)\mathbf{1}(N \geq m_\epsilon)\mathbf{1}(|\hat{\mu}_N - \mu| < \epsilon)] + \mathbf{E}[f(N)\mathbf{1}(N \geq m_\epsilon)\mathbf{1}(|\hat{\mu}_N - \mu| \geq \epsilon)] \\ &\leq \epsilon\mathbf{E}[N] + T\mathbf{P}(\exists m \leq n \leq T, |\hat{\mu}_n - \mu| \geq \epsilon) \\ \text{(by (III.11))} &\leq \epsilon\mathbf{E}[N] + 1.\end{aligned}$$

We continue with the other term $\mathbf{E}[f(N)\mathbf{1}(N < m_\epsilon)]$. Denote $\mathscr{E}_\delta := (\forall n \geq 1, n(\hat{\mu}_n - \mu)^2 \leq 4\sigma^2 \log(\sqrt{1+n}/\delta))$. By a time-uniform Azuma-Hoeffding's inequality again (Lemma I.22), this good event has probability at least $1 - \delta$. We obtain:

$$\mathbf{E}[f(N)\mathbf{1}(N < m_\epsilon)] = \mathbf{E}[f(N)\mathbf{1}(N < m_\epsilon)\mathbf{1}(\mathscr{E}_\delta)] + \mathbf{E}[f(N)\mathbf{1}(N < m_\epsilon)\mathbf{1}(\mathscr{E}_\delta^c)]$$

$$\begin{aligned}
&\leq 2\sigma\mathbf{E}\left[\sqrt{N\log\left(\tfrac{\sqrt{1+N}}{\delta}\right)}\right] + \delta m_\epsilon \\
&\overset{(*)}{\leq} 2\sigma\sqrt{\mathbf{E}\left[N\log\left(\tfrac{\sqrt{1+N}}{\delta}\right)\right]} + \delta m_\epsilon \\
&\equiv 2\sigma\sqrt{\mathbf{E}\left[N\log\left(\tfrac{\sqrt{1+N}}{\delta}\right)\right]} + \delta(\tfrac{2\sigma}{\epsilon})^2\log\left(T\sqrt{1+T}\right)
\end{aligned}$$

where $(*)$ follows from Jensen's inequality. Set $\delta := \frac{\epsilon^3}{6\sigma^2}\frac{\mathbf{E}[N]}{\log(1+T)}$ and plug everything together. □

Applied to sequential control, where N_T is the number of triggers up to time T, we see that when $\mathbf{E}[N_T] + \mathbf{E}[\log(N_T)] \gg \log\log(T)$, then $\mathbf{E}[N_T(\hat{\mu}_{N_T} - \mu)] = o(\mathbf{E}[N])$.

10.B A continuous regularization of the regret lower bound

To ease the proof of Theorem III.13, we introduce a few notions and notations. For $M \in \mathcal{M}$, we introduce the *candidate exploration measures* as the η-uniform measures satisfying the epsilonized exploration and information constraints:

$$\mathcal{I}_\eta^\epsilon(M) := \left\{\mu \in \mathrm{Inv}_\eta(M/\mathcal{Z}_{**}^\epsilon(M)) : \forall M^\dagger \in \mathrm{Cnf}^\epsilon(M),\ \sum_z \mu(z)\mathrm{KL}_z(M||M^\dagger) \geq 1\right\} \tag{III.13}$$

As the condition "$\mathrm{KL}^*(M_n'||M) + \frac{1}{\epsilon_n}\|M_n' - M\|^* \to 0$" will appear repeatedly in the following analysis, we introduce the shorthand

$$d_{\epsilon_n}(M_n'||M) := \mathrm{KL}^*(M_n'||M) + \tfrac{1}{\epsilon_n}\|M_n' - M\|^*. \tag{III.14}$$

10.B.1 Proof of Theorem III.13: Uniqueness of the optimal uniformized exploration measure

For a fixed $M^\dagger \in \mathrm{Cnf}^\epsilon(M)$, the set of $\mu \in \mathbf{R}^{\mathcal{Z}}$ satisfying $\sum_z \mu(z)\mathrm{KL}_z(M||M^\dagger) \geq 1$ is a closed half-space of $\mathbf{R}^{\mathcal{Z}}$. It follows that:

$$\left\{\mu \in \mathbf{R}^{\mathcal{Z}} : \forall M^\dagger \in \mathrm{Cnf}^\epsilon(M),\ \sum_z \mu(z)\mathrm{KL}_z(M||M^\dagger) \geq 1\right\} \tag{III.15}$$

is closed and convex. Meanwhile, $\mathrm{Inv}_\eta(M/\mathcal{Z}^{**}(M))$ is also closed and convex, because it is given as the intersection of linear constraints:

$$\mu \in \mathrm{Inv}_\eta(M/\mathcal{Z}^{**}(M)) \iff \begin{cases} \forall s,a,s', & \sum_{a'}\mu(s',a') = p(s'|s,a)\mu(s,a); \\ \forall s,a, & \mu(s,a) \geq \eta\sum_{a'}\mu(s,a'). \end{cases} \tag{III.16}$$

It follows that $\mathcal{I}_\eta^\epsilon(M)$ is closed and convex. So the solution of (III.7) is the minimum of a η-strongly convex function (for the ℓ_2-norm) over a closed convex set. It is therefore well-defined and unique. ■

10.B.2 Proof of Theorem III.13: Approximation properties of the uniformized lower bound

In this section, we prove the statement (2) of Theorem III.13: Given $M \in \mathcal{M}$ we have $K_\eta^\epsilon(M) \to K(M)$ when $\epsilon, \eta \to 0$. For comparison, we report the regret lower bound and its regularized version side by side:

$$\inf_{\mu \in \mathrm{Inv}(M/\mathcal{Z}^{**}(M))} \sum_{z\in\mathcal{Z}} \mu(z)\Delta^*(z,M) \quad \text{s.t.} \quad \forall M^\dagger \in \mathrm{Cnf}(M),\ \sum_{z\in\mathcal{Z}} \mu(z)\mathrm{KL}_z(M||M^\dagger) \geq 1 \tag{III.6}$$

$$\inf_{\mu\in \mathrm{Inv}_\eta(M/\mathscr{Z}^\epsilon_{**}(M))} \sum_{z\in\mathscr{Z}} \mu(z)\Delta^\epsilon_*(z,M) + \eta\|\mu\|_2^2 \quad \text{s.t.} \quad \forall M^\dagger \in \mathrm{Cnf}^\epsilon(M),\ \sum_{z\in\mathscr{Z}} \mu(z)\mathrm{KL}_z(M||M^\dagger) \geq 1 \tag{III.7}$$

By Proposition III.48, we know that for ϵ small enough, we have $\mathscr{Z}^\epsilon_{**}(M) = \mathscr{Z}^{**}(M)$. It means that $M/\mathscr{Z}^\epsilon_{**}(M) = M/\mathscr{Z}^{**}(M)$ and that $\mathrm{Cnf}^\epsilon(M) = \mathrm{Cnf}(M)$ for such ϵ. Furthermore, by Proposition III.49, we also have $\Delta^\epsilon_*(M) = \Delta^*(M)$ for ϵ small enough. All together, we see that if $\epsilon > 0$ is small enough, then for all $\eta > 0$, (III.19) is equivalent to the following problem:

$$\inf_{\mu\in \mathrm{Inv}_\eta(M/\mathscr{Z}^{**}(M))} \sum_{z\in\mathscr{Z}} \mu(z)\Delta^*(z,M) + \eta\|\mu\|_2^2 \quad \text{s.t.} \quad \forall M^\dagger \in \mathrm{Cnf}(M),\ \sum_{z\in\mathscr{Z}} \mu(z)\mathrm{KL}_z(M||M^\dagger) \geq 1. \tag{III.17}$$

Accordingly, it doesn't depend on ϵ anymore and is the same optimization problem as (III.30), except that the condition $\mu \in \mathrm{Inv}(M/\mathscr{Z}^{**}(M))$ is changed to the stronger $\mu \in \mathrm{Inv}_\eta(M/\mathscr{Z}^{**}(M))$ and that a quadratic regularization term $\eta\|\mu\|_2^2$ has been added to the objective function. It is therefore clear that $K(M) \leq K^\epsilon_\eta(M)$ thus we only need to prove that $\limsup K^\epsilon_\eta(M) \leq K(M)$ as $\epsilon, \eta \to 0$. To that extent, we show that invariant measures of $M/\mathscr{Z}^{**}(M)$ can be approximated by η-uniform invariant measures of $M/\mathscr{Z}^{**}(M)$ provided that η is small enough with Lemma III.17 below.

Lemma III.17 (Uniformization of invariant measures)**.** *Let M a communicating model and pick $\mu \in \mathrm{Inv}(M)$. There exists a family $(\mu_\eta : \eta > 0)$ with $\mu_\eta \in \mathrm{Inv}_\eta(M)$ such that $\mu_\eta \to \mu$ for the ℓ_1-norm when $\eta \to 0$.*

Proof. Let $\mu \in \mathrm{Inv}(M)$ and denote $\mathscr{Z}_1, \dots, \mathscr{Z}_k$ its recurrent components. Without loss of generality, we assume that μ is a probability measure, i.e., $\|\mu\|_1 = 1$. For $\mathscr{Z}' \subseteq \mathscr{Z}$, we denote $\mathscr{S}(\mathscr{Z}') := \{s : \exists a, (s,a) \in \mathscr{Z}\}$ the states where elements of $\mathscr{Z}'$ are rooted. To every component $\mathscr{Z}_i$ corresponds a set of **exit pairs** $\mathscr{Z}_i^-$ that are rooted in $\mathscr{S}(\mathscr{Z}_i)$ but are not supported in μ, i.e., $\mathscr{Z}_i^- := \{(s,a) : s \in \mathscr{S}(\mathscr{Z}_i)\} \setminus \mathscr{Z}_i$. For $\lambda \equiv (\lambda_1, \dots, \lambda_k) \in (\mathbf{R}_+^*)^k$, consider the following randomized policy:

$$\pi^\lambda_\eta(a|s) := \begin{cases} \frac{1}{|\mathscr{A}(s)|} & \text{if } s \notin \mathscr{S}(\mathscr{Z}_1 \cup \dots \mathscr{Z}_k); \\ \lambda_i\eta + (1 - \lambda_i\eta|\mathscr{A}(s)|)\frac{\mu(s,a)}{\sum_{a'\in\mathscr{A}(s)}\mu(s,a')} & \text{if } s \in \mathscr{S}(\mathscr{Z}_i). \end{cases}$$

Observe that π^λ_η is $\eta\min(\lambda)$-uniform so that it has a unique invariant probability measure denoted μ^λ_η. We now provide an equivalent of μ^λ_η when $\eta \to 0$.

Consider the minor $[M] := M/(\mathscr{Z}_1 \cup \dots \cup \mathscr{Z}_k) = M/\mathscr{Z}_1/\cdots/\mathscr{Z}_k$ and denote $\mathscr{S}_i := \mathscr{S}[\mathscr{Z}_i]$ for simplicity. In $[M]$, $\mathscr{S}_i$ becomes a single state denoted $[\mathscr{S}_i]$. We consider the modification $[M]'$ of $[M]$ where the playable actions from each $[\mathscr{S}_i]$ are modified as follows. There are two actions, i^+ and i^- with kernels given by:

$$\begin{aligned} [p]'([\mathscr{S}_i]|[\mathscr{S}_i],[i^+]) &:= 1; \\ [p]'([s']|[\mathscr{S}_i],[i^-]) &:= \sum_{(s,a)\in\mathscr{Z}_i^-} \tfrac{\mu(s)}{\|\mu|_{\mathscr{Z}_i}\|_1}[p]([s']|[s,a]). \end{aligned}$$

Morally, i^+ is a looping action, while i^- is an exit action that plays an exit pair (s,a) pondered by the probability of being in s when navigating with μ on $\mathscr{Z}_i$. Let $[\pi]'$ the policy of $[M]'$ that plays i^- from every $[s_i]$ and is uniform from other states. This policy is recurrent and has a unique invariant probability measure $[\mu]'$. Remark that, for all $(s,a) \in \mathscr{Z}_i$, we have:

$$\mu^\lambda_\eta(s,a) \underset{\eta\to 0}{\sim} \frac{\lambda_i}{\|\lambda\|_1} \cdot \frac{[\mu]'[\mathscr{S}_i]}{\sum_j [\mu]'[\mathscr{S}_j]} \cdot \frac{\mu(s,a)}{\|\mu|_{\mathscr{Z}_i}\|_1}. \tag{III.18}$$

Choose $\lambda_i = \|\mu|_{\mathscr{Z}_i}\|_1/[\mu]'[\mathscr{S}_i]$. This concludes the proof. □

We now finish the proof. Fix $\delta > 0$ and let $\mu \in \mathrm{Inv}(M/\mathscr{Z}^{**}(M))$ achieving $K(M)$ in (III.6) within δ range, i.e., $\sum_z \mu(z)\Delta^*(z,M) \le K(M) + \delta$. By Lemma III.17, provided that $\eta > 0$ is small enough, there exists $\mu_\eta \in \mathrm{Inv}(M/\mathscr{Z}^{**}(M))$ such that $\|\mu_\eta - \mu\|_1 \le \delta$. So:

$$\begin{aligned} K^\epsilon_\eta(M) &\le \sum_{z\in\mathscr{Z}} \mu_\eta(z,M)\Delta^*(z) + \eta\|\mu_\eta\|_2^2 \\ &\le \sum_{z\in\mathscr{Z}} \mu(z,M)\Delta^*(z) + \delta\|\Delta^*(M)\|_\infty + \eta\|\mu_\eta\|_1^2 \\ &\le K(M) + \delta + \delta\|\Delta^*(M)\|_\infty + 2\eta\big(\|\mu\|_1^2 + \delta^2\big). \end{aligned}$$

We conclude that, for all $\delta > 0$

$$\limsup_{\epsilon,\eta\to 0} K^\epsilon_\eta(M) \le K(M) + \delta(1 + \|\Delta^*(M)\|_\infty)$$

so $\limsup K^\epsilon_\eta(M) \le K(M)$ when $\epsilon, \eta \to 0$. ∎

10.B.3 Proof of Theorem III.13: Continuity properties of the uniformized lower bound

Fix (ϵ_n) the sequence of optimality relaxation levels and assume that $\epsilon_n \le \epsilon_0$ as given by Proposition III.12. In this section, we show the following sub-statement of Theorem III.13:

Assume that $M = (r,p)$ satisfies $0 < r(z) < 1$. For all $(M'_n) \in \mathscr{M}^{\mathbf{N}}$, we have:

$$K^{\epsilon_n}_\eta(M'_n) \to K^0_\eta(M) \quad \textit{and} \quad \mu^{\epsilon_n}_\eta(M'_n) \to \mu^0_\eta(M) \quad \textit{when} \quad \mathrm{KL}^*(M'_n||M) + \frac{\|M'_n - M\|^*}{\epsilon_n} \to 0. \tag{III.19}$$

In general, the continuity of a maximizer with respect to the parameters is dealt with the Maximum Theorem Berge (1957). A few subtleties make the result difficult to apply in our setting however. There is first the double regime requirement on the convergence speeds of (M'_n) and (ϵ_n) that would require an artificial space transformation, and then we would have to show that the constraints embodied by $\mathscr{I}^\epsilon_\eta(M)$ are lower and upper semi-continuous. This is however not the case at infinity. In the end, it seems to easier to go with an *ad hoc* argument than to try to apply the Maximum Theorem directly.

The result (III.19) is established by showing the twin results below. In spirit, they are closed to semicontinuity results.

Lemma III.18. *Assume that $M \in \mathscr{M}$ is such that, for all $z \in \mathscr{Z}$, $0 < r(z) < 1$. Fix $\eta > 0$ and let $(M_n) \in \mathscr{M}^{\mathbf{N}}$.*

(1) ("Lower semicontinuity") *At infinity, $\mu^0_\eta(M)$ is near $\mathscr{I}^{\epsilon_n}_\eta(M'_n)$ in the following sense:*

$$\inf_{\mu' \in \mathscr{I}^{\epsilon_n}_\eta(M'_n)} \left\| \mu' - \mu^0_\eta(M) \right\|_1 = \mathrm{o}(1) \quad \textit{when} \quad d_{\epsilon_n}(M'_n||M) \to 0 \tag{III.20}$$

(2) ("Upper semicontinuity") *At infinity, $\mu^{\epsilon_n}_\eta(M'_n)$ is near $\mathscr{I}^0_\eta(M)$ in the following sense:*

$$\inf_{\mu \in \mathscr{I}^0_\eta(M)} \left\| \mu - \mu^{\epsilon_n}_\eta(M'_n) \right\|_1 = \mathrm{o}(1) \quad \textit{when} \quad d_{\epsilon_n}(M'_n||M) \to 0. \tag{III.21}$$

Outline of the next sections. In Section 10.B.3.1, we show that Lemma III.18 is enough to conclude. In Section 10.B.5, we introduce various preliminary material about the properties of invariant measures, of candidate measures and confusing models, preparing Section 10.B.6 and Section 10.B.7. The Section 10.B.6 is dedicated to the proof of (III.20), and Section 10.B.7 to the proof of (III.21). Although (III.20) and (III.21) are very different (yet dual) results, their proofs mostly identical up to a few minor differences.

10.B.3.1 Lemma III.18 is enough to conclude

Recall that $\epsilon_n \le \epsilon_0$. Pick $(M'_n) \in \mathscr{M}^{\mathbb{N}}$. Remark that, by Proposition III.12, we have

$$\Delta_*^{\epsilon_n}(M'_n) = \Delta^*(M) + \mathrm{o}(1) \quad \text{when} \quad \mathrm{KL}^*(M'_n||M) + \frac{\|M'_n - M\|^*}{\epsilon_n} \to 0. \tag{III.22}$$

(**STEP 1**) *The regret lower bound $K_\eta^{\epsilon_n}(M'_n)$ is asymptotically bounded with $K_\eta^0(M)$ as:*

$$K_\eta^{\epsilon_n}(M'_n) \le K_\eta^0(M) + \mathrm{o}(1) \quad \textit{when} \quad \mathrm{KL}^*(M'_n||M) + \tfrac{1}{\epsilon_n}\|M'_n - M\|^* \to 0 \tag{III.23}$$

Proof. Denote P'_n the ℓ_2-projection operation onto the closed convex set $\mathscr{I}_\eta^{\epsilon_n}(M'_n)$ (see Section 10.B.1). By Lemma III.18, we know that $\|P'_n(\mu_\eta^0(M)) - \mu_\eta^0(M)\|_1 = \mathrm{o}(1)$ when $\mathrm{KL}^*(M'_n||M) + \frac{1}{\epsilon_n}\|M'_n - M\|^* \to 0$. We have:

$$\begin{aligned}
K_\eta^{\epsilon_n}(M'_n) &\le \sum_{z\in\mathscr{Z}} \Big(P'_n(\mu_\eta^0(M))\Big)(z)\Delta_*^{\epsilon}(z,M') + \eta\left\|P'_n(\mu_\eta^0(M))\right\|_2^2 \\
&\overset{(\dagger)}{\le} \sum_{z\in\mathscr{Z}} \Big(P'_n(\mu_\eta^0(M))\Big)(z)\Delta^*(z,M) + \eta\left\|P'_n(\mu_\eta^0(M))\right\|_2^2 + \mathrm{o}\Big(\left\|P'_n(\mu_\eta^0(M))\right\|\Big) \\
&\overset{(\ddagger)}{\le} \sum_{z\in\mathscr{Z}} \mu_\eta^0(z,M)\Delta^*(z,M) + \eta\left\|\mu_\eta^0(M)\right\|_2^2 + \mathrm{o}\Big(1 + \left\|P'_n(\mu_\eta^0(M))\right\| + \left\|\mu_\eta^0(M)\right\|\Big) \\
&= K_\eta^0(M) + \mathrm{o}(1)
\end{aligned}$$

where (†) is an application of Proposition III.12 in the form of (III.22) and (†) follows from Lemma III.18, and all the o(−) hold when $\mathrm{KL}^*(M'_n||M) + \frac{1}{\epsilon_n}\|M'_n - M\|^* \to 0$. □

(**STEP 2**) *The regret lower bound $K_\eta^{\epsilon_n}(M'_n)$ is asymptotically bounded with $K_\eta^0(M)$ as:*

$$K_\eta^{\epsilon_n}(M'_n) \ge K_\eta^0(M) + \mathrm{o}(1) \quad \textit{when} \quad \mathrm{KL}^*(M'_n||M) + \tfrac{1}{\epsilon_n}\|M'_n - M\|^* \to 0. \tag{III.24}$$

Proof. Let P_0 the ℓ_2-projection operation onto the closed convex set $\mathscr{I}_\eta^0(M)$. By Lemma III.18, we have $\|P_0(\mu_\eta^{\epsilon_n}(M'_n)) - \mu_\eta^{\epsilon_n}(M'_n)\| = \mathrm{o}(1)$ when $\mathrm{KL}^*(M'_n||M) + \frac{1}{\epsilon_n}\|M'_n - M\|^* \to 0$. With similar computations than in (**STEP 1**), we show:

$$K_\eta^0(M) \le K_\eta^{\epsilon_n}(M'_n) + \mathrm{o}\Big(1 + \|\mu_\eta^{\epsilon_n}(M'_n)\|\Big) \tag{III.25}$$

when $\mathrm{KL}^*(M'_n||M) + \frac{1}{\epsilon_n}\|M'_n - M\|^* \to 0$. To conclude that the RHS is $K_\eta^{\epsilon_n}(M'_n) + \mathrm{o}(1)$, it is enough to show $\|\mu_\eta^{\epsilon_n}(M'_n)\|$ is bounded. By definition $\mu_\eta^{\epsilon_n}(M'_n)$ achieves $K_\eta^{\epsilon_n}(M'_n)$, so we see that:

$$\left\|\mu_\eta^{\epsilon_n}(M'_n)\right\|_2^2 \le \tfrac{1}{\eta}K_\eta^{\epsilon_n}(M'_n) \overset{(*)}{\le} \tfrac{1}{\eta}\Big(K_\eta^0(M) + \mathrm{o}(1)\Big) \tag{III.26}$$

when $\mathrm{KL}^*(M'_n||M) + \frac{1}{\epsilon_n}\|M'_n - M\|^* \to 0$, where $(*)$ follows from the previous bound on $K_\eta^{\epsilon_n}(M'_n)$, see (III.23). Combining (III.25) and (III.26) together, we get $K_\eta^0(M) \le K_\eta^{\epsilon_n}(M'_n) + \mathrm{o}(1)$. We therefore have shown

$$K_\eta^{\epsilon_n}(M'_n) = K_\eta^0(M) + \mathrm{o}(1) \quad \text{when} \quad \mathrm{KL}^*(M'_n||M) + \tfrac{1}{\epsilon_n}\|M'_n - M\|^* \to 0.$$

This proves the claim. □

(**STEP 3**) *The optimal exploration measures are asymptotically related as:*

$$\mu_\eta^{\epsilon_n}(M'_n) = \mu_\eta^0(M) + \mathrm{o}(1) \quad \textit{when} \quad \mathrm{KL}^*(M'_n||M) + \tfrac{1}{\epsilon_n}\|M'_n - M\|^* \to 0. \tag{III.27}$$

Proof. Assume, ad absurdum, that there exists $\delta > 0$ such that, for all $n \geq 1$, there exists $M'_n \in \mathscr{M}$ such that $\mathrm{KL}^*(M'_n||M) + \epsilon_n^{-1}\|M'_n - M\|^* \leq \frac{1}{n}$ satisfying $\|\mu_\eta^{\epsilon_n}(M'_n) - \mu_\eta^0(M)\|_1 > \delta$. By (III.26), the sequence $\mu_\eta^{\epsilon_n}(M'_n)$ is bounded hence, by compactness, we can assume that it is converging to some μ' satisfying $\|\mu' - \mu_\eta^0(M)\| > \delta$ without loss of generality. We obtain:

$$K_\eta^0(M) \overset{(\dagger)}{=} \lim_{n\to\infty} K_\eta^{\epsilon_n}(M'_n) = \sum_{z\in\mathscr{Z}} \mu'(z) \lim_{n\to\infty} \Delta_*^{\epsilon_n}(z, M'_n) + \eta\|\mu'\|_2^2 \overset{(\ddagger)}{=} \sum_{z\in\mathscr{Z}} \mu'(z)\Delta^*(z, M) + \eta\|\mu'\|_2^2$$

where (†) follows from (III.24) and (‡) follows from Proposition III.12 in the form of (III.22). By Lemma III.18 and because $\mathscr{I}_\eta^0(M)$ is closed (Section 10.B.1), we have $\mu' \in \mathscr{I}_\eta^0(M)$. But $\mu_\eta^0(M)$ is the unique member of $\mathscr{I}_\eta^0(M)$ achieving $K_\eta^0(M)$, hence $\mu' = \mu_\eta^0(M)$; a contradiction. □

All together, (III.23), (III.24) and (III.27) prove Theorem III.13 as stated in (III.19). ■

10.B.4 Preliminary results on the set of confusing models

Lemma III.19 (Reduced confusing models)**.** *Assume that the underlying model space $\mathscr{M}$ is in product form, i.e., that $\mathscr{M} = \prod_{z\in\mathscr{Z}}[0,1] \times \mathscr{P}(\mathscr{Z})$. Let M a communicating model, let $\epsilon > 0$ and $M^\dagger \in \mathrm{Cnf}^\epsilon(M)$. There exists $M^\ddagger \in \mathrm{Cnf}^\epsilon(M)$ such that:*

(1) For all $z \in \mathscr{Z}$ and $q \in \{r, p\}$, $\mathrm{KL}(q(z)||q^\ddagger(z)) \leq \mathrm{KL}(q(z)||q^\dagger(z))$;
(2) $M^\ddagger$ has small bias span: $\mathrm{sp}(h^*(M^\ddagger)) \leq D(M)$;
(3) $M^\ddagger(z) = M(z)$ for all $z \notin \mathscr{Z}^{**}(M^\ddagger) \setminus \mathscr{Z}_{**}^\epsilon(M)$ and gain-optimal policies of $M^\ddagger$ are unichain.

Proof. Without loss of generality, we can assume that $M^\dagger(z) = M(z)$ outside of $\mathscr{Z}^{**}(M^\dagger) \setminus \mathscr{Z}_{**}^\epsilon(M)$ and that every optimal policy of $M^\dagger$ is unichain. Consider the extended model M' where, from state $s \in \mathscr{S}$, the choice of an action consists in choosing whether the transition is done in M or in $M^\dagger$, hence choosing $a \in \mathscr{A}(s)$, a reward among $\{r(s,a), r^\dagger(s,a)\}$ and a transition among $\{p(s,a), p^\dagger(s,a)\}$. The model M' is communicating because it contains M with diameter bounded by $D(M') \leq D(M)$. We further have:

$$g^*(M') \geq g^*(M^\dagger) \quad \text{and} \quad g^*(M') \geq g^*(M).$$

Any (extended) policy achieving optimal bias in M' defines a policy π' on M and a model $M^\ddagger$ such that $g^*(M^\ddagger) = g^*(M') = g_{\pi'}(M^\ddagger)$ and $\mathrm{sp}(h^*(M^\ddagger)) = \mathrm{sp}(h^*(M')) \leq D(M') \leq D(M)$ (see Proposition II.2) making $M^\ddagger$ satisfy (2). Moreover, (1) is automatically satisfied.

We now justify that $M^\ddagger \in \mathrm{Cnf}^\epsilon(M)$. Because $M^\dagger \in \mathrm{Cnf}^\epsilon(M)$, $\mathscr{Z}_{**}^\epsilon(M)$ remains untouched in $M^\dagger$ and $M^\ddagger$. Accordingly, every ϵ-gain optimal policy of M has all its recurrent pairs within $\mathscr{Z}_{**}^\epsilon(M)$ and has gain at most $g^*(M)$. But since $M^\dagger \in \mathrm{Cnf}^\epsilon(M)$, we have $g^*(M^\dagger) > g^*(M)$, so $g^*(M^\ddagger) > g^*(M)$ and since $M^\ddagger(z) = M(z)$ for $z \in \mathscr{Z}_{**}^\epsilon(M)$ by construction, we have $M^\ddagger \in \mathrm{Cnf}^\epsilon(M)$.

If (3) isn't satisfied, then either π' picks transients rewards or transitions from $M^\dagger$ or is multi-chain. In the first case, remove the transient transitions and in the second case, remove a recurrent component; Then start the construction over. Repeat until (3) is met. □

Lemma III.20. *If M is communicating, then* $\inf_{M^\dagger\in\mathrm{Cnf}(M)} \mathrm{KL}(M||M^\dagger) > 0$.

Proof. Let $M^\dagger \in \mathrm{Cnf}(M)$. By assumption, we have $p^\dagger \gg p$ so $M^\dagger$ is communicating as well, so the uniform policy π_u is recurrent under $p^\dagger$ and p. By Lemma III.40, if $\|M^\dagger\|M\|$ is smaller than some $\epsilon(M) > 0$, then $D(p^\dagger_{\pi_u}) \leq 2D(p_{\pi_u})$, where $D(p^\pi)$ is the **policy-diameter** of π (Definition III.8).

Let $\pi, \pi^\dagger$ bias optimal policies of M and $M^\dagger$ respectively. By adapting the rationale of Proposition II.2, we see that $\mathrm{sp}(h^*(M)) \leq D(p_{\pi_u})$ and $\mathrm{sp}(h^*(M)) \leq D(p^\dagger_{\pi_u})$. Both $M, M^\dagger$ are communicating, so we have $\mathrm{sp}(g^*(M)) = 0$ and $\mathrm{sp}(g^*(M^\dagger)) = 0$, so by Theorem II.1,

$$\begin{aligned} \|g^\pi(M^\dagger) - g^\pi(M)\|_\infty &\leq \|r^\dagger - r\|_\infty + \frac{1}{2} D(p_{\pi_u}) \|p^\dagger - p\|_1, \\ \|g^{\pi^\dagger}(M^\dagger) - g^{\pi^\dagger}(M)\|_\infty &\leq \|r^\dagger - r\|_\infty + \frac{1}{2} D(p^\dagger_{\pi_u}) \|p^\dagger - p\|_1. \end{aligned} \tag{III.28}$$

Let $\Delta_g(M) := \min\{\|g^\pi(M) - g^*(M)\|_\infty : \pi \notin \Pi^*(M)\} > 0$ the gain-gap of M. Since $\pi^\dagger \notin \Pi^*(M)$,

$$\begin{aligned} \Delta_g(M) &\leq \|g^{\pi^\dagger}(M) - g^\pi(M)\|_\infty = \max\Big(g^\pi(M) - g^{\pi^\dagger}(M)\Big) \\ &\leq \max\Big(g^\pi(M) - g^\pi(M^\dagger)\Big) + \max\Big(g^\pi(M^\dagger) - g^{\pi^\dagger}(M^\dagger)\Big) + \max\Big(g^{\pi^\dagger}(M^\dagger) - g^{\pi^\dagger}(M)\Big) \\ &\leq \max\Big(g^\pi(M) - g^\pi(M^\dagger)\Big) + \max\Big(g^{\pi^\dagger}(M^\dagger) - g^{\pi^\dagger}(M)\Big) \\ &\leq 2\Big(\|r^\dagger - r\|_\infty + D(p_{\pi_u})\|p^\dagger - p\|_1\Big). \end{aligned}$$

We obtain that:

$$\|M^\dagger - M\| \geq \frac{\Delta_g(M)}{4(D(p_{\pi_u}) \vee 1)} \wedge \epsilon(M). \tag{III.29}$$

Conclude using Pinsker's inequality. □

10.B.5 Preliminary results on uniform invariants and candidate measures

We denote $\Pi_\eta(M)$ the space of η-uniform policies, i.e., randomized policies such that $\pi(a|s) \geq \eta$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$. For p_π a policy kernel, we denote $D(p_\pi)$ its diameter (Definition III.8).

Lemma III.21 (Continuity of uniform invariant measures)**.** *In this assertion, we say that a function $f : \mathcal{M} \to \mathbf{R}^d$ is* KL*-continuous *if for all $M \in \mathcal{M}$ and $\epsilon > 0$, $\exists \delta > 0, \mathrm{KL}^*(M'||M) < \delta \Rightarrow \|f(M') - f(M)\| < \epsilon$.*

(1) If M is communicating, then $D_\eta(M) := \sup_{\pi \in \Pi_\eta(M)} D(p_\pi) < \infty$;

(2) D_η is continuous at communicating models: if M is communicating, then

$$\forall \epsilon > 0, \exists \delta > 0, \mathrm{KL}^*(M'||M) < \epsilon \Rightarrow |D_\eta(M') - D_\eta(M)| + \delta;$$

(3) There exists $c(M) > 0$ such that, for all $\mu \in \mathrm{Inv}_\eta(M), \min(\mu) \geq \|\mu\|_1 c(M)$ and $c(-)$ is KL**-continuous at communicating models;*

(4) $\mathcal{M} \mapsto \mathrm{Inv}_\eta(M) \cap \mathcal{P}(\mathcal{Z})$ is KL**-continuous at M if M is communicating for the Hausdorff distance.*

Proof. We prove all the claims in order. We consider first the model M_η constructed from M as follows. The state-action space is still $\mathcal{Z}$, but the kernels and rewards are given by:

$$\begin{aligned} p_\eta(s,a) &:= (1 - \eta|\mathcal{A}(s)|)p(s,a) + \eta \sum_{a' \in \mathcal{A}(s)} p(s,a'); \\ r_\eta(s,a) &:= (1 - \eta|\mathcal{A}(s)|)r(s,a) + \eta \sum_{a' \in \mathcal{A}(s)} r(s,a'). \end{aligned}$$

The obtained model M_η is communicating, because the execution of the fully uniform policy on M_η is equivalent to the execution of the uniform policy on M. By construction, the execution of any η-uniform policy π of M can be seen as the execution of a randomized policy π_η of M_η.

(**STEP 1**) *If M is communicating, then $D_\eta(M) := \sup_{\pi\in\Pi_\eta(M)} D(p_\pi) < \infty$.*

Proof. Fix $s \in \mathscr{S}$ and consider M_η^s the transform of M_η making s absorbing. Consider the reward function $\mathbf{f}(s',a') := \mathbf{1}(s=s')$ and consider the deterministic policy $\pi_\eta^{\mathbf{f}}$ with maximal bias for $\mathbf{f}$ as a cost function, in particular, it is solving the Bellman equations associated to the minimization of the total f-reward. Because M is communicating and $\pi_\eta^{\mathbf{f}}$ is fully supported, it eventually reaches s and has gain $g^{\mathbf{f}} = 0$. By construction, $\pi_\eta^{\mathbf{f}}$ has optimal $\mathbf{f}$-bias $h^{\mathbf{f}}(s') := \mathbf{E}_{s'}[\inf\{t \geq 1 : S_t = s\}]$ which is here maximized because $\pi_\eta^{\mathbf{f}}$ aims at minimizing the $\mathbf{f}$-reward. Accordingly, letting $D_\eta^s < \infty$ the bias span of π_η, we see that every policy of $\Pi_\eta(M)$ has reaching time to s upper-bounded by D_η^s in M.

Set $D_\eta := \max_s D_\eta^s < \infty$. Observe that D_η corresponds to the diameter of a policy $\pi \in \Pi_\eta(M)$ and that it upper-bounds the diameter of every other policy of $\Pi_\eta(M)$. Accordingly, $D_\eta = \sup_{\pi\in\Pi_\eta(M)} D(p_\pi)$. □

(**STEP 2**) *D_η is continuous at communicating models: if M is communicating, then*

$$\forall \epsilon > 0, \exists \delta > 0, \mathrm{KL}^*(M'||M) < \epsilon \Rightarrow |D_\eta(M') - D_\eta(M)| + \delta.$$

Proof. This is mostly a consequence of Lemma III.39 in the form of (III.125). By construction, every $\pi \in \Pi_\eta(M)$ has diameter bounded by $D_\eta(M)$. If $\mathrm{KL}^*(M'||M) < \infty$, then by (III.125), we have

$$\left|D(p'_\pi) - D(p_\pi)\right| \leq \tfrac{1}{2} D(p_\pi) D(p'_\pi) \|M' - M\|^* \leq \tfrac{1}{2} D_\eta(M) D(p'_\pi) \|M' - M\|^*.$$

To obtain an upper-bound of $D(p'_\pi)$, see that $D(p'_\pi)(1 - \frac{1}{2} D_\eta(M))\|M' - M\|^* \leq D(p_\pi)$. So, when $\|M' - M\|^*$ is small enough, we have:

$$D(p'_\pi) \leq \frac{D(p_\pi)}{1 - \frac{1}{2} D_\eta(M)\|M' - M\|^*} \leq D(p_\pi) + D_\eta(M)^2 \|M' - M\|^*.$$

Similarly for the lower bound, when $\|M' - M\|^*$ is small enough, we have

$$D(p'_\pi) \leq D(p_\pi) - D_\eta(M)^2 \|M' - M\|^*.$$

According, every $D(p_\pi)$ is locally $D_\eta(M)^2$-Lipschitz continuous at M for $\|-\|^*$. So D_η is locally $D_\eta(M)^2$-Lipschitz continuous at M as a supremum of Lipschitz continuous functions. □

(**STEP 3**) *There is $c(M) > 0$ such that, for all $\mu \in \mathrm{Inv}_\eta(M), \min(\mu) \geq \|\mu\|_1 c(M)$ and $c(-)$ is KL^*-continuous at communicating models.*

Proof. Fix $s \in \mathscr{S}$ and let $\mathbf{f} \equiv \mathbf{f}_s := \mathbf{1}(s = s')$. Seeing $\mathbf{f}$ as a reward function on M_η, we consider a bias-optimal deterministic policy $\pi_\eta^{\mathbf{f}}$ for $\mathbf{f}$ as a cost function. It has $\mathbf{f}$-gain $g^{\mathbf{f}} \in \mathbf{R}e$ and $\mathbf{f}$-bias $h^{\mathbf{f}}$ satisfying:

$$\forall (s',a') \in \mathscr{X}, \quad \Delta^{\mathbf{f}}(s',a') := -\mathbf{f}(s',a') + g^{\mathbf{f}}(s') + \left(e_{s'} - p_\eta(s',a')\right) h^{\mathbf{f}} \leq 0.$$

This deterministic policy $\pi_\eta^{\mathbf{f}}$ of M_η corresponds to a η-uniform policy $\pi^{\mathbf{f}}$ of M. Because M is communicating and $\pi^{\mathbf{f}}$ η-uniform, $\pi^{\mathbf{f}}$ is recurrent on M hence $g^{\mathbf{f}} > 0$. Now, every policy $\pi' \in \Pi_\eta(M)$ corresponds to a policy π'_η of M_η that satisfies:

$$\mathbf{f}(s', \pi'_\eta(s')) \geq g^{\mathbf{f}}(s') + \left(e_{s'} - p_\eta(s', \pi'_\eta(s'))\right) h^{\mathbf{f}}.$$

So $g^{\mathbf{f}}(\pi'_\eta) \geq g^{\mathbf{f}}$.

Now, because every $\pi \in \Pi_\eta(M)$ is recurrent, the first return time to s is $\frac{1}{\mu_\pi(s)}$ where μ_π is the unique probability invariant measure of π (Levin and Peres, 2017, Proposition 1.19). It follows that for all $\mu \in \mathrm{Inv}_\eta(M)$ and $a \in \mathscr{A}(s)$, we have $\mu(s,a) \geq \eta g^{\mathbf{f}_s}(s)\|\mu\|_1$. The lower-bound $c(M) := \eta \min_s g^{\mathbf{f}_s}(s) > 0$ is continuous as the minimum of finitely many gains (here $g^{\mathbf{f}_s}$, all continuous by Lemma III.41) of a continuous communicating transform (here $M \mapsto M_\eta$) of M. □

(**STEP 4**) *$\mathscr{M} \mapsto \mathrm{Inv}_\eta(M) \cap \mathscr{P}(\mathscr{Z})$ is KL^*-continuous at M if M is communicating for the Hausdorff distance.*

Proof. Every measure of $\mathrm{Inv}_\eta(M)$ corresponds to some η-uniform policy π whose diameter is upper-bounded by $D_\eta(M)$ by (**STEP 2**). Provided that $\mathrm{KL}^*(M'||M) < \infty$, we have $p_\pi \sim p'_\pi$ and by Lemma III.42, the unique probability measures $\mu'_\pi \in \mathrm{Inv}(\pi, M')$ and $\mu_\pi \in \mathrm{Inv}(\pi, M)$ satisfy:

$$\|\mu'_\pi - \mu_\pi\|_\infty \leq D_\eta(M)\|p'_\pi - p_\pi\|_1 \leq D_\eta(M)\|M' - M\|^*.$$

It follows that the Hausdorff distance in ℓ_∞-norm between $\mathrm{Inv}_\eta(M') \cap \mathscr{P}(\mathscr{Z})$ and $\mathrm{Inv}_\eta(M) \cap \mathscr{P}(\mathscr{Z})$ is bounded by $D_\eta(M)\|M' - M\|^*$, which vanishes with $\mathrm{KL}^*(M'||M)$ by Pinsker's inequality. □

This concludes the proof of Lemma III.21. ■

Lemma III.22 (Candidate measures are bounded away from 0)**.** *There exists a constant $\gamma(M) > 0$ such that,*

$$\inf\left\{\|\mu\|_\infty : \mu \in \mathscr{I}_\eta^{\epsilon_n}(M'_n)\right\} \geq \gamma(M) \quad \textit{when} \quad d_{\epsilon_n}(M'_n||M) \to 0. \tag{III.30}$$

Proof. Pick $M^\dagger \in \mathrm{Cnf}(M)$. For $n \geq 1$, denote $M'^\dagger_n$ the model obtained by setting $M'^\dagger_n(z) = M^\dagger(z)$ if $z \notin \mathscr{Z}^{**}(M)$ and $M'^\dagger(z) = M'_n(z)$ if $z \in \mathscr{Z}^{**}(M)$. We show that

$$M'^\dagger_n \in \mathrm{Cnf}^{\epsilon_n}(M'_n) \quad \text{when} \quad d_{\epsilon_n}(M'_n||M) \to 0. \tag{III.31}$$

When $d_{\epsilon_n}(M'_n||M)$ is small enough, we know that $\mathscr{Z}_{**}^{\epsilon_n}(M'_n) = \mathscr{Z}^{**}(M)$, see (III.33) and Proposition III.12. Since $M'_n \sim M$, a policy has the same recurrent classes on M'_n and M, so that ϵ_n-gain optimal policies of M'_n are exactly gain optimal policies of M. Pick $\pi \in \Pi^*(M)$ and $\pi^\dagger \in \Pi^*(M^\dagger)$. Denote $C := 1 + \max\{\mathrm{sp}(h_{\pi^\dagger}(M^\dagger)), \mathrm{sp}(h_\pi(M))\}$ and remark that by communicativity, we have $g_{\pi^\dagger}(M^\dagger) \in \mathbf{R}e$ and $g_\pi(M) \in \mathbf{R}e$. When $d_{\epsilon_n}(M'_n||M) \to 0$, we have

$$\begin{aligned}
g_{\pi^\dagger}(M'^\dagger_n) - g_\pi(M'^\dagger_n) &\overset{(\dagger)}{=} g_{\pi^\dagger}(M'^\dagger_n) - g_\pi(M'_n) \\
&\overset{(\ddagger)}{\geq} g_{\pi^\dagger}(M^\dagger) - g_\pi(M'_n) - C\|M'^\dagger_n - M^\dagger\|e \\
&\overset{(\ddagger)}{\geq} g_{\pi^\dagger}(M^\dagger) - g_\pi(M) - 2C\|M'^\dagger_n - M^\dagger\|e \\
&= g_{\pi^\dagger}(M^\dagger) - g_\pi(M^\dagger) - 2C\|M'^\dagger_n - M^\dagger\|e
\end{aligned} \tag{III.32}$$

where (†) holds when $d_{\epsilon_n}(M'_n||M) \to 0$, and (‡) both follow from Lemma III.41. Because $M^\dagger \in \mathrm{Cnf}(M)$, the quantity $g_{\pi^\dagger}(M^\dagger) - g_\pi(M^\dagger)$ is positive and independent from the choice of (M'_n). Let $c := \min_{\pi \in \Pi^*(M)}(g_{\pi^\dagger}(M^\dagger) - g_\pi(M^\dagger)) > 0$. We deduce that for $d_{\epsilon_n}(M'_n||M)$ small enough, if in addition $\|M'_n - M\| \leq \frac{c}{2C}$, then every optimal policy of M has lesser gain than $\pi^\dagger$ in $M'^\dagger_n$. As $\mathscr{Z}_{**}^{\epsilon_n}(M'_n) = \mathscr{Z}^{**}(M)$ when $d_{\epsilon_n}(M'_n||M)$ is small enough, and knowing that $M'^\dagger_n$ is a copy of M'_n on $\mathscr{Z}^{**}(M)$ by construction, it follows the recurrent pairs of gain optimal policies of M'_n

cannot be all contained in $\mathscr{Z}_{**}^{\epsilon_n}(M_n'^{\dagger})$. Accordingly, we have $M_n'^{\dagger} \in \mathrm{Cnf}^{\epsilon_n}(M_n')$, showing (III.31). Moreover, we have:

$$\forall z \in \mathscr{Z}, \quad \mathrm{KL}_z(M_n'||M_n'^{\dagger}) = \mathrm{KL}_z(M||M^{\dagger}) + \mathrm{o}(1)$$

when $\mathrm{KL}^*(M_n'||M)$ is small enough, hence $\mathrm{KL}_z(M_n'||M_n'^{\dagger}) \leq 2\mathrm{KL}_z(M||M^{\dagger})$ when $d_{\epsilon_n}(M_n'||M) \to 0$. Pick $\mu \in \mathscr{I}_{\eta}^{\epsilon_n}(M_n')$. Because $\sum_z \mu(z)\mathrm{KL}_z(M_n'||M_n'^{\dagger}) \geq 1$ when $d_{\epsilon_n}(M_n'||M) \to 0$, there exists (z_n) such that:

$$\frac{1}{|\mathscr{Z}|} \leq \mu(z_n)\mathrm{KL}_{z_n}(M_n'||M_n'^{\dagger}) \leq \mu(z_n) \cdot 2\mathrm{KL}_{z_n}(M||M^{\dagger}) \quad \text{when} \quad d_{\epsilon_n}(M_n'||M) \to 0$$

so $\|\mu\|_\infty \geq \gamma(M) := \frac{1}{2|\mathscr{Z}|}\mathrm{KL}(M||M^{\dagger})$, and $\mathrm{KL}(M||M^{\dagger})$ is bounded away from 0 by Lemma III.20. □

10.B.6 Proof of Lemma III.18: "lower" semicontinuity

We throughout fix $M \in \mathscr{M}$ with $0 < r(z) < 1$ and assume that $\mathrm{Cnf}(M) \neq \varnothing$. Let $(M_n') \in \mathscr{M}^{\mathbf{N}}$ arbitrary such that $\|M_n' - M\|^* = \mathrm{o}(\epsilon_n)$. Recall that by Proposition III.12, we have

$$\mathscr{Z}_{**}^{\epsilon_n}(M_n') \to \mathscr{Z}^{**}(M) \quad \text{and} \quad \Delta_*^{\epsilon_n}(M_n') = \Delta^*(M) + \mathrm{o}(1) \quad \text{when} \quad d_{\epsilon_n}(M_n'||M) \to 0. \tag{III.33}$$

Specifically, it means for all $\delta > 0$, provided that $d_{\epsilon_n}(M_n'||M)$ is small enough relatively to δ, we have $\mathscr{Z}_{**}^{\epsilon_n}(M_n') = \mathscr{Z}^{**}(M)$ and $\|\Delta_*^{\epsilon_n}(M_n') - \Delta^*(M)_\infty\| \leq \delta$.

Idea of the proof. We want to show that $\mu_\eta^0(M)$ is close to $\mathscr{I}_\eta^{\epsilon_n}(M_n')$ provided that $d_{\epsilon_n}(M_n'||M)$ is small enough. The first goal of the proof is to show that $\mu_\eta^0(M)$ is close to "rejecting" all elements of $\mathrm{Cnf}_n^{\epsilon}(M_n')$ in the sense that:

$$\inf_{M_n' \in \mathrm{Cnf}^{\epsilon_n}(M_n')} \sum_{z \in \mathscr{Z}} \mu_\eta^0(z, M)\mathrm{KL}_z(M_n'||M_n'^{\dagger}) \geq 1 - \mathrm{o}(1) \quad \text{when} \quad d_{\epsilon_n}(M_n'||M) \to 0. \tag{III.34}$$

Then, $\mu_\eta^0(M)$ will be "corrected" into an element of $\mathscr{I}_\eta^{\epsilon_n}(M_n')$ by invoking properties of $\mathrm{Inv}_\eta(M_n')$. In (**STEP 1**), we show that one can restrict the analysis to confusing models $M_n'^{\dagger}$ with rewards uniformly bounded away form the boundary. Then, we construct from $M_n'^{\dagger}$ a model $M_n^{\dagger}$ that is almost a confusing model for M, that we repair into a confusing model with $M_n^{\ddagger\delta_n}$ where δ_n is a parameter. (**STEP 2**) introduce technical results that are used in (**STEP 3**) to tune δ_n, making sure that $M_n^{\ddagger\delta_n}$ is very close to $M_n^{\dagger}$. In (**STEP 4**), we relate the sums $\sum_{z\in\mathscr{Z}} \mu_\eta^0(z, M)\mathrm{KL}_z(M_n'||M_n'^{\dagger})$ and $\sum_{z\in\mathscr{Z}} \mu_\eta^0(z, M)\mathrm{KL}_z(M_n||M_n^{\ddagger\delta_n})$ to establish (III.34). We conclude by projecting $\mu_\eta^0(M)$ onto $\mathrm{Inv}_\eta^{\epsilon_n}(M_n')$ with Lemma III.21.

(**STEP 1**) *There exists a constant $\epsilon > 0$ such that, if $d_{\epsilon_n}(M_n'||M)$ is small enough, then every $M_n'^{\dagger} \in \mathrm{Cnf}^{\epsilon_n}(M_n')$ such that $\exists z \in \mathscr{Z}, r_n'^{\dagger}(z) < \epsilon$ or $r_n'^{\dagger}(z) > 1 - \epsilon$, satisfies:*

$$\sum_{z\in\mathscr{Z}} \mu_\eta^0(z, M)\mathrm{KL}_z(M_n'||M_n'^{\dagger}) \geq 1. \tag{III.35}$$

In other words, we can focus on $M_n'^{\dagger} \in \mathrm{Cnf}^{\epsilon_n}(M_n')$ with rewards within $[\epsilon, 1-\epsilon]$.

Proof. We denote $\mu \equiv \mu_\eta^0(M)$ for short. By (III.30), we have $\|\mu\| \geq \gamma(M) \equiv \gamma > 0$. By Lemma III.21, because μ is η-uniform we have $\min(\mu) \geq c\|\mu\|_\infty$ where $c > 0$ is a constant, hence $\mu(z) \geq c\gamma > 0$ for all $z \in \mathscr{Z}$. So:

$$\sum_{z\in\mathscr{Z}} \mu(z)\mathrm{KL}_z(M_n'||M_n'^{\dagger}) \geq c\gamma \max_{z\in\mathscr{Z}} \mathrm{KL}_z(M_n'||M_n'^{\dagger}).$$

If $\max_z \mathrm{KL}_z(M'_n||M'^{\dagger}_n) \geq \frac{1}{c\gamma}$ then the above is already greater that one and (III.35) is satisfied.

We focus on the case where $\max_z \mathrm{KL}_z(M'_n||M'^{\dagger}_n) < \frac{1}{c\gamma}$. In particular, we have $\mathrm{KL}(r'_n(z)||r'^{\dagger}_n(z)) \leq \frac{1}{c\gamma}$ for all $z \in \mathscr{Z}$. Since $r'_n(z) < 1$ for all $z \in \mathscr{Z}$ provided that $d_{\epsilon_n}(M'_n||M)$ is small enough, and knowing that the rewards are Bernoulli, we get:

$$r'^{\dagger}_n(z) \leq 1 - \exp\left(-\frac{\frac{1}{c\gamma} + \mathrm{Ent}(r'_n(z))}{1 - r'_n(z)}\right) \leq 1 - \frac{1}{2}\exp\left(-\frac{\frac{1}{c\gamma} + \mathrm{Ent}(r(z))}{1 - r(z)}\right)$$

where the second inequality holds when $d_{\epsilon_n}(M'_n||M)$ is small enough. Introduce $1-\epsilon$ has the supremum of the RHS for $z \in \mathscr{Z}$, satisfying $1-\epsilon < 1$ since $r(z) < 1$ for all $z \in \mathscr{Z}$. Similarly, using $r(z) > 0$, we show that $r'^{\dagger}_n(z)$ is bounded from below, say $r'^{\dagger}_n(z) > \epsilon$ for all $z \in \mathscr{Z}$. □

Definitions of $M^{\dagger}_n$ and $M^{\ddagger\delta}_n$. Following (III.35), we assume that $M'^{\dagger}_n \in \mathrm{Cnf}^{\epsilon_n}(M'_n)$ has rewards with $[\epsilon, 1-\epsilon]$. Without loss of generality, we can consider $M'^{\dagger}_n$ reduced according to Lemma III.19. In particular, we have $\mathrm{sp}(h^*(M'^{\dagger}_n)) \leq D(M'_n)$. Since $M'^{\dagger}_n \in \mathrm{Cnf}^{\epsilon_n}(M'_n)$, its rewards and kernels are the same than M'_n on $\mathscr{Z}^{\epsilon_n}_{**}(M)$ (i.e., on $\mathscr{Z}^{**}(M)$ once $d_{\epsilon_n}(M'_n||M)$ is small enough) on which the gain is $g^*(M'_n)$. Yet, $M'^{\dagger}_n$ has greater gain than M'_n, hence there exists another recurrent class of $M'^{\dagger}_n$ with greater gain which is unique because $M'^{\dagger}_n$ is reduced, see Lemma III.19. Picking a bias optimal policy $\pi'^{\dagger}_n$ of $M'^{\dagger}_n$, we see that some of the recurrent pairs of $\pi'^{\dagger}_n$ are not within $\mathscr{Z}^{**}(M)$ for $d_{\epsilon_n}(M'_n||M)$ small enough.

Denote $\mathscr{Z}'^{\dagger}_n$ the recurrent pairs of $\pi'^{\dagger}_n$ that are not members of $\mathscr{Z}^{**}(M)$. Consider $M^{\dagger}_n$ given as follows:

$$M^{\dagger}_n(z) := \begin{cases} M(z) & \text{if } x \in \mathscr{Z}^{**}(M) \\ M'^{\dagger}_n(z) & \text{if } x \notin \mathscr{Z}^{**}(M) \end{cases} \tag{III.36}$$

Morally, $M^{\dagger}_n$ is almost an alternative model of M, but $\pi'^{\dagger}_n$ may have lost its status of optimal policy from $M'^{\dagger}_n$ to $M^{\dagger}_n$. To counterbalance this, we slightly increase the asymptotic reward of $\pi'^{\dagger}_n$ by considering $M^{\ddagger}_n \equiv M^{\ddagger\delta}_n$ obtained by changing $r^{\dagger}_n$ to $r^{\dagger}_n + \delta \mathbf{1}(\mathscr{Z}'^{\dagger}_n)$, that simply adds $\delta > 0$ to pairs of $\mathscr{Z}'^{\dagger}_n$ only.

We now search for conditions on $\delta > 0$ making $M^{\ddagger\delta}_n \in \mathrm{Cnf}(M)$.

(**STEP 2**) *Recall that $\overline{p}(\pi)$ denotes the asymptotic kernel of π under p, i.e., is given by $\overline{p}(s,s';\pi) := \lim \frac{1}{T}\mathbf{E}^{\pi}_s[\sum_{t=1}^{T} \mathbf{1}(S_t = s')]$. Let $\pi \in \Pi$ with recurrent components within $\mathscr{Z}^{\epsilon_n}_{**}(M)$ in M. We have:*

$$g(\pi'^{\dagger}_n, M^{\ddagger\delta}_n) \geq g(\pi'^{\dagger}_n, M'^{\dagger}_n) - (1 + 2D(M))\|M'_n - M\|e - \delta \overline{p}'^{\dagger}_n(\pi'^{\dagger}_n) e_{\mathscr{Z}'^{\dagger}_n}; \tag{III.37}$$

$$g(\pi, M^{\ddagger\delta}_n) \leq g^*(M'^{\dagger}_n|_{\mathscr{Z}^{**}(M)}) + (1 + 2D(M))\|M'_n - M\|^* e \tag{III.38}$$

when $d_{\epsilon_n}(M'_n||M)$ is small enough.

Proof. We start with (III.37). To lighten up notations, we write $\pi \equiv \pi'^{\dagger}_n$ and $u'^{\dagger}_{\pi}, u^{\ddagger\delta}_{\pi}, u^{\dagger}_{\pi}$ for $u(\pi, M'^{\dagger}_n), u(\pi, M^{\ddagger\delta}_n)$ and $u(\pi, M^{\dagger}_n)$. We have:

$$\begin{aligned}
g^{\ddagger\delta}_{\pi} = \overline{p}^{\ddagger\delta}_{\pi} r^{\ddagger\delta}_{\pi} &= \overline{p}^{\dagger}_{\pi}\left(r^{\dagger}_{\pi} + \delta e_{\mathscr{Z}'^{\dagger}_n}\right) \\
&= \overline{p}^{\dagger}_{\pi}\left(r'^{\dagger}_{\pi} + \left(r^{\dagger}_{\pi} + \delta e_{\mathscr{Z}'^{\dagger}_n} - r'^{\dagger}_{\pi}\right)\right) \\
&\overset{(*)}{=} \overline{p}^{\dagger}_{\pi}\left(g'^{\dagger}_{\pi} + \left(\mathbf{I} - p'^{\dagger}_{\pi}\right)h'^{\dagger}_{\pi} + \left(r^{\dagger}_{\pi} + \delta e_{\mathscr{Z}'^{\dagger}_n} - r'^{\dagger}_{\pi}\right)\right) \\
&\overset{(\S)}{=} g'^{\dagger}_{\pi} + \overline{p}^{\dagger}_{\pi}\left(\left(\mathbf{I} - p^{\dagger}_{\pi}\right)h'^{\dagger}_{\pi} + \left(p^{\dagger}_{\pi} - p'^{\dagger}_{\pi}\right)h'^{\dagger}_{\pi} + \left(r^{\dagger}_{\pi} + \delta e_{\mathscr{Z}'^{\dagger}_n} - r'^{\dagger}_{\pi}\right)\right) \\
&\overset{(\$)}{=} g'^{\dagger}_{\pi} + \overline{p}^{\dagger}_{\pi}\left(\left(p^{\dagger}_{\pi} - p'^{\dagger}_{\pi}\right)h'^{\dagger}_{\pi} + \left(r^{\dagger}_{\pi} - r'^{\dagger}_{\pi}\right)\right) + \delta \overline{p}^{\dagger}_{\pi} e_{\mathscr{Z}'^{\dagger}_n}
\end{aligned}$$

where $(*)$ follows from the Poisson equation $r'^{\dagger}_{\pi} = g'^{\dagger}_{\pi} + (\mathbf{I} - p'^{\dagger}_{\pi})h'^{\dagger}_{\pi}$, (§) from $\mathrm{sp}(g'^{\dagger}_{\pi}) = 0$ and ($) from $\overline{p}^{\dagger}_{\pi}(\mathbf{I} - p^{\dagger}_{\pi}) = 0$. By Proposition II.2, we have $\mathrm{sp}(h'^{\dagger}_{\pi}) = \mathrm{sp}(h^*(M'^{\dagger}_n)) \le D(M'^{\dagger}_n)$ and because $M'^{\dagger}_n$ is reduced, we have $D(M'^{\dagger}_n) \le D(M'_n)$ by Lemma III.19. Moreover, $D(M'_n) \le D_*(M'_n)$ and when $d_{\epsilon_n}(M'_n||M_n)$ is small enough, we have $D(M'_n) \le 2D(M_n)$, see Lemma III.39 and (III.125). We obtain

$$g(\pi'^{\dagger}_n, M^{\ddagger\delta}_n) \ge g(\pi'^{\dagger}_n, M'^{\dagger}_n) - (1 + 2D(M))\|M'_n - M\|e - \delta\overline{p}'^{\dagger}_n(\pi'^{\dagger}_n)e_{\mathscr{Z}'^{\dagger}_n}$$

when $d_{\epsilon_n}(M'_n||M) \to 0$, establishing (III.37).

We proceed with (III.38). Assume that $d_{\epsilon_n}(M'_n||M)$ is small enough so that $\mathscr{Z}^{\epsilon_n}_{**}(M) = \mathscr{Z}^{**}(M)$. If $\pi \in \Pi$ has recurrent components within $\mathscr{Z}^{\epsilon_n}_{**}(M)$, then since $M'^{\dagger}_n \gg M'_n$ and that M'_n and M are mutually absolutely continuous, it follows that the recurrent class of π has to be a subset of $\mathscr{Z}^{**}(M)$. We have:

$$g(\pi, M^{\ddagger\delta}_n) = g(\pi, M) \le g^*(M) \overset{(*)}{\le} g^*(M'_n|_{\mathscr{Z}^{**}(M)}) + D(M)\|M'_n - M\|e$$

where $(*)$ follows from Lemma III.41. □

(**STEP 3**) *There exists a constant $c(M) > 0$ such that, setting*

$$\delta_n := \frac{2(1 + 2D(M))\|M'_n - M\|}{c(M)} \wedge (1 - \max(r)) \tag{III.39}$$

we have $M^{\ddagger}_n \equiv M^{\ddagger\delta_n}_n \in \mathrm{Cnf}(M)$ provided that $d_{\epsilon_n}(M'_n||M)$ is small enough.

Proof. Assume that $d_{\epsilon_n}(M'_n||M)$ is small enough so that (III.37) and (III.38) hold. Let $\pi \in \Pi^*(M)$. We have:

$$g(\pi'^{\dagger}_n, M^{\ddagger\delta}_n) \ge g(\pi, M^{\dagger\delta}_n) - 2(1 + 2D(M))\|M'_n - M\|e + \delta\overline{p}(\pi'^{\dagger}_n, M^{\dagger}_n)e_{\mathscr{Z}'^{\dagger}_n}.$$

We claim that $\overline{p}(\pi^{\dagger}_n, M'^{\dagger}_n)e_{\mathscr{Z}'^{\dagger}_n} \ge c(M)e$ where $c(M) := |\mathscr{S}|^{-1}(\min_{z\in\mathscr{Z}}\min\{p(s|z) > 0 : s \in \mathscr{S}\})^{|\mathscr{S}|-1}$.

To see this, consider the recurrent component $\mathscr{Z}''_n$ of $\pi'^{\dagger}_n$ in $M'^{\dagger}_n$, that remains a recurrent component in $M^{\dagger}_n$. Let μ''_n the unique invariant probability measure of $\mathscr{Z}''_n$ on $M^{\dagger}_n$ (which is the line $\overline{p}(s, -; \pi'^{\dagger}_n)$ for $(s, \pi'^{\dagger}_n(s)) \in \mathscr{Z}''_n$). Consider $(s''_n, a''_n) \in \mathscr{Z}''_n$ such that that $\mu''_n(s''_n, a''_n) \ge |\mathscr{S}|^{-1}$. If it belongs to $\mathscr{Z}'^{\dagger}_n$ (the recurrent pairs of $\pi'^{\dagger}_n$ that are not members of $\mathscr{Z}^{**}(M)$), the claim is established. Otherwise, because $\mathscr{Z}''_n$ is recurrent and that $\mathscr{Z}''_n \cap \mathscr{Z}'^{\dagger}_n \neq \varnothing$, there is a path from (s''_n, a''_n) to $(s'''_n, a'''_n) \in \mathscr{Z}''_n \cap \mathscr{Z}'^{\dagger}_n$ of length most $|\mathscr{S}| - 1$ taking only pairs from $\mathscr{Z}^{**}(M)$. This path has probability $|\mathscr{S}|c(M)$ at least because M and $M^{\dagger}_n$ coincide on $\mathscr{Z}^{**}(M)$, hence $\mu''_n(s'''_n, a'''_n) \ge c(M)$, proving the claim.

Overall, we obtain:

$$g(\pi'^{\dagger}_n, M^{\ddagger\delta}_n) \ge g(\pi, M^{\dagger\delta}_n) + 2\big(\delta c(M) - (1 + 2D(M))\|M'_n - M\|\big)e.$$

Solving $\delta c(M) - 2(1 + D(M))\|M'_n - M\| \ge 0$ in δ provides the value of δ_n. Remark that when $d_{\epsilon_n}(M'_n||M) \to 0$, we have $\|M'_n - M\| \to 0$ hence $\delta_n \to 0$. Accordingly, we eventually have $\max(r) + \delta_n < 1$. □

(**STEP 4**) *Establishing* (III.34)*: $\mu^0_\eta(M)$ nearly rejects $\mathrm{Cnf}^{\epsilon_n}(M'_n)$ provided that $d_{\epsilon_n}(M'_n||M)$ is small enough:*

$$\inf_{M'^{\dagger}_n \in \mathrm{Cnf}^{\epsilon_n}(M'_n)} \sum_{z\in\mathscr{Z}} \mu^0_\eta(z, M)\mathrm{KL}_z(M'_n||M'^{\dagger}_n) \ge 1 - o(1) \quad \textit{when} \quad d_{\epsilon_n}(M'_n||M) \to 0. \tag{III.34}$$

Proof. We denote $\mu \equiv \mu_\eta^0(M)$ for short. Provided that $\mathrm{KL}^*(M_n'||M)$ is small enough, then for $z \in \mathscr{Z}$ and $q \in \{r,p\}$, on the support of $q(z)$ we have $(1-\rho_n)q(z;i) \leq q_n'(z;i)$ where $\rho_n \to 0$ when $d_{\epsilon_n}(M_n'||M_n) \to 0$. We have:

$$\begin{aligned}
(-) &:= \sum_{z\in\mathscr{Z}} \mu(z)\mathrm{KL}_z(M_n'||M_n'^\dagger) \\
&= \sum_{z\in\mathscr{Z}} \mu(z) \sum_q \sum_i q_n'(z;i) \log\left(\frac{q_n'(z;i)}{q_n'^\dagger(z;i)}\right) \\
&= \sum_{z\in\mathscr{Z}} \mu(z) \sum_q \left(\sum_i q_n'(z;i)\log\left(\frac{1}{q_n'^\dagger(z;i)}\right) - \mathrm{Ent}(q_n'(z))\right) \\
&\overset{(*)}{\geq} \sum_{z\in\mathscr{Z}} \mu(z) \sum_q \left(\sum_i (1-\rho_n)q(z;i)\log\left(\frac{1}{(1-\rho_n)q_n^\dagger(z;i)}\right) - (1-\rho_n)\mathrm{Ent}(q(z)) - \ell\rho_n\right) \\
&\geq \sum_{z\in\mathscr{Z}} \mu(z) \sum_q \Big((1-\rho_n)\mathrm{KL}(q(z)||q_n^\dagger(z)) - (\ell+1)\rho_n\Big) \\
&= (1-\rho_n)\sum_{z\in\mathscr{Z}} \mu(z)\mathrm{KL}_z(M||M_n^\dagger) - 2\rho_n(\ell+1)\|\mu\|_1
\end{aligned}$$

where $\ell = 1+2\log|\mathscr{S}|$ and $(*)$ relates $\mathrm{Ent}(q_n'(z))$ to $\mathrm{Ent}(q(z))$. Now, $\mathrm{KL}_z(M||M_n^\dagger) = \mathrm{KL}(r(z)||r_n^\dagger(z)) + \mathrm{KL}(p(z)||p_n^\dagger(z))$ and we know that $p_n^\dagger = p_n^\ddagger$. We proceed with:

$$\begin{aligned}
\sum_{z\in\mathscr{Z}} \mu(z)\mathrm{KL}(r(z)||r_n^\dagger(z)) &= \sum_{z\in\mathscr{Z}} \mu(z)\mathrm{kl}(r(z), r_n^\dagger(z)) \\
&= \sum_{z\in\mathscr{Z}} \mu(z)\mathrm{kl}\Big(r(z), r_n^\ddagger(z) - \delta_n \mathbf{1}(z \in \mathscr{Z}_n'^\dagger)\Big) \\
&\overset{(*)}{\geq} \sum_{z\in\mathscr{Z}} \mu(z)\left(\mathrm{kl}(r(z), r_n^\ddagger(z)) - \delta_n \mathbf{1}(z \in \mathscr{Z}_n'^\dagger)\tfrac{r(z)-r_n^\ddagger(z)}{r_n^\ddagger(z)(1-r_n^\ddagger(z))}\right) \\
&\geq \sum_{z\in\mathscr{Z}} \mu(z)\left(\mathrm{kl}(r(z), r_n^\ddagger(z)) - \delta_n \mathbf{1}(z \in \mathscr{Z}_n'^\dagger)\tfrac{1}{r_n^\ddagger(z)(1-r_n^\ddagger(z))}\right) \\
&\geq \sum_{z\in\mathscr{Z}} \mu(z)\left(\mathrm{kl}(r(z), r_n^\ddagger(z)) - \delta_n \mathbf{1}(z \in \mathscr{Z}_n'^\dagger)\tfrac{1}{\epsilon(1-\epsilon)}\right)
\end{aligned}$$

where $(*)$ is obtained by convexity of $\mathrm{kl}(r(z),-)$. Accordingly, we have:

$$\sum_{z\in\mathscr{Z}} \mu(z)\mathrm{KL}_z(M||M_n^\dagger) \geq \sum_{z\in\mathscr{Z}} \mu(z)\mathrm{KL}_z(M||M_n^\ddagger) - \tfrac{\delta_n\|\mu\|_1}{\epsilon(1-\epsilon)}$$

since $M^\ddagger \in \mathrm{Cnf}(M)$ and $\mu \in \mathscr{I}_\eta^0(M)$. Following (III.39), we obtain:

$$\sum_{z\in\mathscr{Z}} \mu_\eta^0(z,M)\mathrm{KL}_z(M_n'||M_n'^\dagger) \geq 1-\rho_n - \Big(\tfrac{\delta_n}{\epsilon(1-\epsilon)} + 2\rho_n(\ell+1)\Big)\cdot\|\mu_\eta^0(M)\|_1 = 1-\mathrm{o}(1)$$

when $d_{\epsilon_n}(M_n'||M) \to 0$. □

We can now conclude. Denote $\mu \equiv \mu_\eta^0(M)$ for short. From (III.34), for all $\epsilon > 0$, we have:

$$\inf_{M_n'^\dagger \in \mathrm{Cnf}^{\epsilon_n}(M_n')} (1+\epsilon)\sum_{z\in\mathscr{Z}} \mu(z)\mathrm{KL}_z(M_n'||M_n'^\dagger) \geq 1 \tag{III.40}$$

when $d_{\epsilon_n}(M_n'||M)$ is small relatively to ϵ and M. By Lemma III.21, if $d_{\epsilon_n}(M_n'||M)$ is small enough, there exists $\mu_n' \in \mathrm{Inv}_\eta^{\epsilon_n}(M_n')$ such that $(1+\epsilon)\mu \leq \mu_n' \leq (1+2\epsilon)\mu_n'$, so by (III.40), we have $\mu_n' \in \mathscr{I}_\eta^{\epsilon_n}(M_n')$ and by construction, $\|\mu_n' - \mu\| \leq \epsilon\|\mu\|$. This establishes (III.20). ■

10.B.7 Proof of Lemma III.18: "upper" semicontinuity

This is the same proof as Lemma III.18.

10.C Analysis of ECoE

In this part, we analyze the regret of Algorithm III.2, ECoE. For reference, the pseudo-code is reported on Algorithm III.5.

We prove the theorem below.

Theorem III.23. *Fix the parameters of Algorithm III.5 to $\eta, \delta > 0$. For all $M \in \mathscr{M}$ such that $0 < r(z) < 1$, the regret of Algorithm III.5 is asymptotically bounded by:*

$$\limsup_{T\to\infty} \frac{\mathbf{E}^M[\mathrm{Reg}(T)]}{\log(T)} \leq K^0_\eta(M). \tag{III.42}$$

10.C.1 High level architecture of the regret analysis

We are working with the version of the algorithm with additional flag variables, see Algorithm III.5. For instance, time instants are partitioned into four categories.

(1) $t \in \mathscr{T}^-$ is an *exploration time* where the algorithm isn't playing π^+_t and is fetching information;
(2) $t \in \mathscr{T}^\pm$ is a *co-exploration time*, where the algorithm is playing π^+_t for informational purposes;
(3) $t \in \mathscr{T}^+$ is an *exploitation time*, where the algorithm is playing π^+_t to score maximally;
(4) $t \in \mathscr{T}^!$ is a *panic time*, where an unknown transition has been seen and data needs to be updated.

The number of visits of a sub-optimal pair $z \in \mathscr{Z}^-(M)$ are decomposed as follows:

$$N_T(z) = \sum_{t=0}^{T-1} \mathbf{1}(Z_t = z, t \in \mathscr{T}^-) + \sum_{t=0}^{T-1} \mathbf{1}(Z_t = z, t \in \mathscr{T}^\pm) + \sum_{t=0}^{T-1} \mathbf{1}(Z_t = z, t \in \mathscr{T}^+) + \sum_{t=0}^{T-1} \mathbf{1}(Z_t = z, t \in \mathscr{T}^!). \tag{III.43}$$

That is, we distinguish four cases. Either $Z_t = z$ is played because (1) the algorithm is exploring to collect information on presumably sub-optimal pairs, or (2) because it is co-exploring to collect information on the (wrongly) presumed optimal policy, or (3) it is exploiting the (wrongly) presumed optimal policy, or (4) the algorithm just panicked with a transition that was never seen so far. Below is a short description of how every term behaves.

(1) $Z_t = z$ with $t \in \mathscr{T}^-$ accounts for the dominant part of $N_T(z)$ and is the hardest to analyze and is bounded using a $\log^2(T)$*-barrier technique*. Thanks to uniform exploration properties (Lemma III.29), every pair of M is visited at least $\Omega(|\mathscr{T}^-(t)|)$ where $\mathscr{T}^-(t)$ is the number of exploration times up to t. So, with moderately high probability $1 - \exp(-\sqrt{\log(T)})$, $\hat{M}_t$ is well concentrated around M once $|\mathscr{T}^-(t)| \geq \gamma \log(T)$ for $\gamma > 0$ a small constant, so the exploitation policy is correct, the exploration measure is correct and the algorithm will explore optimally. In case of concentration failure, thanks to the skeleton $\mathscr{Z}_t$, $\mathrm{Reg}(T)$ and $N_z(T)$ are shown to be $\mathrm{O}(\log^2(T))$ with overwhelming probability. This $\mathrm{O}(\log^2(T))$

Algorithm III.5 A near optimal algorithm (annotated version)

Parameters: Exploration uniformization $\eta > 0$, ambient space $\mathscr{M}$.

Use

$$\mathrm{Alt}^{\epsilon(t)}(\hat{M}_t) := \left\{\hat{M}^\dagger \gg \hat{M}_t : \mathscr{Z}^{**}(\hat{M}^\dagger) \not\subseteq \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)\right\}. \tag{III.41}$$

Use near-optimality threshold $\epsilon(t) = \frac{1}{\log\log(t)}$. Use GLR overshoot $\delta(t) := \frac{1}{\log\log(t)} = \omega\left(\frac{\log\log(t)}{\log(t)}\right)$.

for episodes $k = 1, 2, \ldots$ **do**
  Set $t_k \leftarrow t$;
  Update exploitation policy $\pi^+_{t_k}$: uniform on $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ from $\mathscr{S}(\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t))$ and uniform elsewhere;
  Update exploration measure $\mu_{t_k} \leftarrow \mu_\eta^{\epsilon(t)}(\hat{M}_t)$, deduce exploration policy $\pi^-_{t_k}(a|s) \propto \mu_t(s,a)$;
  Update skeleton $\mathscr{Z}_t \leftarrow \{z \in \mathscr{Z} : N_t(z) \geq \log^2(t)\}$;
  Update extended skeleton $\mathscr{Y}_t \leftarrow \mathscr{Z}_t \cup \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$;
  if S_t is not recurrent under π^+_t **then**
    Play A_t according to $\pi^+_t(S_t, -)$; ▷ EXPLORATION ($t \in \mathscr{T}^-$)
    $t \leftarrow t+1$, add $t-1$ in $\mathscr{T}^-$;
  else if $\exists M^\dagger \in \mathrm{Alt}^{\epsilon(t)}(\hat{M}_t)$ s.t. $M^\dagger|_{\mathscr{Y}_t} = \hat{M}_t|_{\mathscr{Y}_t}$ and $\sum_z N_t(z)\mathrm{KL}_z(\hat{M}_t||M^\dagger) \leq (1+\delta(t))\log(t)$ **then**
    Play A_t according to $\pi^-_t(S_t, -)$;
    $t \leftarrow t+1$, add $t-1$ in $\mathscr{T}^-$;
  else if $\exists M^\dagger \in \mathrm{Alt}^{\epsilon(t)}(\hat{M}_t)$ s.t. $M^\dagger|_{\mathscr{Z}_t} = \hat{M}_t|_{\mathscr{Z}_t}$ and $\sum_z N_t(z)\mathrm{KL}_z(\hat{M}_t||M^\dagger) \leq (1+\delta(t))\log(t)$ **then**
    ▷ COEXPLORATION ($t \in \mathscr{T}^\pm$)
    Split $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ into communicating components $\mathscr{Z}_t^1, \ldots, \mathscr{Z}_t^{m(t)}$;
    Let $\mathscr{Z}_t^{i(t)}$ the current component containing S_t;
    if $\log\min\{N_t(z) : z \in \mathscr{Z}_t^{i(t)}\} < 2\log\min\{N_t(z) : z \in \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)\}$ **then**
      Add $t \in \mathscr{T}_0^\pm$;
      repeat
        Play A_t according to $\pi^+_t(-|S_t)$;
        $t \leftarrow t+1$;
        if $S_t \notin \mathscr{S}(\mathscr{Z}_t^{i(t)})$ **then** add $t-1$ in $\mathscr{T}^!$ and **break**; ▷ transition discovery
        else add $t-1$ in $\mathscr{T}^\pm$;
      until $S_t = S_{t_k}$; ▷ regeneration
    else
      Play A_t according to $\pi^-_t(S_t, -)$;
      $t \leftarrow t+1$, add $t-1$ in $\mathscr{T}^-$;
    end if
  else
    Add t in $\mathscr{T}_0^+$; ▷ EXPLOITATION ($t \in \mathscr{T}^+$)
    repeat
      Play A_t according to $\pi^+_t(-|S_t)$;
      $t \leftarrow t+1$;
      if $S_t \notin \mathscr{S}(\mathscr{Z}_t^{i(t)})$ **then** add $t-1$ in $\mathscr{T}^!$ and **break**; ▷ transition discovery
      else add $t-1$ in $\mathscr{T}^+$;
    until $S_t = S_{t_k}$; ▷ regeneration
  end if
end for

is killed by the error probability $\exp(-\sqrt{\log(T)})$ when T is large enough, so that what happens in case of concentration failure can be neglected.

(2) $Z_t = z$ with $t \in \mathscr{T}^{\pm}$ means that the algorithm is tackling with a possible lack of information on the empirically optimal policy that is actually sub-optimal. The expected number of such time instants is directly linked to the correctness of the co-exploration test, and it is shown that there are $O(\log\log(T)^3)$ sub-optimal co-exploration times in expectation (Lemma III.24).

(3) $Z_t = z$ with $t \in \mathscr{T}^{+}$ means that the algorithm isn't explorating, yet the empirically optimal policy is wrong. The expected number of such time instant is directly linked to the correctness of the exploration test, and it shown that the terms $\mathbf{1}(Z_t = z, t \in \mathscr{T}^{+})$ account for $O(1)$ in expectation (Lemma III.25).

(4) Every time the algorithm the algorithm panics, a new transition has been observed. Because the total number of transitions is $|\mathscr{Z}| \times |\mathscr{S}|$, the cardinal of $|\mathscr{T}^{!}|$ is bounded and these terms can be neglected.

Lemma III.24 (Wrong co-exploration). *For all $z \notin \mathscr{Z}^{**}(M)$, we have:*

$$\mathbf{E}\left[\sum_{t=0}^{T-1} \mathbf{1}(Z_t = z, t \in \mathscr{T}^{\pm})\right] = O\big(\log\log(T)^3\big). \tag{III.44}$$

Lemma III.25 (Wrong exploitation). *For all $z \notin \mathscr{Z}^{**}(M)$, we have:*

$$\mathbf{E}\left[\sum_{t=0}^{\infty} \mathbf{1}(Z_t = z, t \in \mathscr{T}^{+})\right] < \infty. \tag{III.45}$$

Lemma III.26 (Exploration). *For all $z \notin \mathscr{Z}^{**}(M)$, we have:*

$$\mathbf{E}\left[\sum_{t=0}^{T-1} \mathbf{1}(Z_t = z, t \in \mathscr{T}^{-})\right] \leq \mu_{\eta}^{0}(z, M)\log(T) + o(\log(T)). \tag{III.46}$$

10.C.2 Proof of Lemma III.25: Amount of wrong exploitation

Fix $z \notin \mathscr{Z}^{**}(M)$. Because exploitation is always done until regeneration, there are two kinds of exploitation time instants. The first kind are *initial* exploitation times $t \in \mathscr{T}_0^{+}$ and the second kind are exploitation times $t \in \mathscr{T}^{+} \setminus \mathscr{T}_0^{+}$ awaiting for regeneration. The amount of the first bound the second and are time-instants when the co-exploration GLR test has been passed.

(**STEP 1**) *For $z \notin \mathscr{Z}^{**}(M)$, we have:*

$$\mathbf{E}\left[\sum_{t=0}^{T-1} \mathbf{1}(Z_t = z, t \in \mathscr{T}^{+})\right] = O\left(\mathbf{E}\left[\sum_{t=0}^{T-1} \mathbf{1}(\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) \neq \mathscr{Z}^{**}(M), t \in \mathscr{T}_0^{+})\right]\right) \tag{III.47}$$

Proof. Let (τ_i^0) the stopping time enumeration of initial exploitation times such that $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) \neq \mathscr{Z}^{**}(M)$. They corresponds to phases $t_k, \ldots, t_{k+1} - 1$. Construct inductively the sequence of stopping times (τ_i) and the index progression $j(-)$ such that $\tau_i^0 = \tau_{j(i)}$ and $\tau_i^0 + \ell = \tau_{j(i)+\ell}$ if $\tau_i^0 + \ell$ and τ_i^0 are within the same phase. Fix $z_0 \notin \mathscr{Z}^{**}(M)$. Observe that z_0 cannot be exploited

at time $\tau_{j(i)+\ell}$ unless at time $t = \tau_{j(i)}$, we have $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) \neq \mathscr{Z}^{**}(M)$. It follows that:

$$\sum_{t=0}^{T-1} \mathbf{1}(Z_t = z_0, t \in \mathscr{T}^+) = \sum_{i=1}^{\infty} \sum_{t=\tau_{j(i)}}^{\tau_{j(i+1)-1} \wedge (T-1)} \mathbf{1}(Z_t = z_0) \leq \sum_{i=1}^{\infty} \mathbf{1}(\tau_{j(i)} < T)\big(\tau_{j(i+1)-1} - \tau_{j(i)} + 1\big). \tag{III.48}$$

Because every exploitation phase is done until exploration (or panic), the aggregate duration of exploitation phases is bounded by the number of such phases. To see this, consider an exploitation phase $t \equiv t_k$. Let $\mathscr{Z}_0$ the component of $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ that the algorithm is intending to exploit. Consider the reward function $f(s,a) = \mathbf{1}(s = S_{t_k}) + \mathbf{1}(s \notin \mathscr{S}(\mathscr{Z}_0))$ which is marking the initial state of the current phase and the outside of the states spawned by $\mathscr{Z}_0$. By design, over the exploitation phase we have:

$$2 = \sum_{t=t_k}^{t_{k+1}} f(Z_t) = \sum_{t=t_k}^{t_{k+1}-1} f(Z_t) + 1..$$

Consider g^f, h^f the gain and bias functions associated to the reward function f obtained by iterating the exploitation policy of the current phase $\pi^+_{t_k}$. Remark that $\mathrm{sp}(g^f(s)) = 0$ and that $\min(g^f) > 0$ Let $0 < c, C < \infty$ the minimum (respectively maximum) value that $\min(g^f)$ (respectively $\mathrm{sp}(h^f)$) can take for all (finitely many) values of $\pi^+_{t_k}, \mathscr{Z}_0, S_{t_k}$. We have:

$$1 = \sum_{t=t_k}^{t_{k+1}-1} f(Z_t) \overset{(\dagger)}{=} \sum_{t=t_k}^{t_{k+1}-1} \big(g^f(S_t) + \big(e_{S_t} - p(S_t, A_t)\big)h^f\big) \geq (t_{k+1} - t_k)c - C + \sum_{t=t_k}^{t_{k+1}-1} \big(e_{S_{t+1}} - p(S_t, A_t)\big)h^f$$

where $(\dagger)$ uses the Poisson equation $g^f(s) + h^f(s) = f(s,a) + p(s,a)h^f$. Taking the expectation, we find that $\mathbf{E}[t_{k+1} - t_k] \leq \frac{1}{c} \cdot (1 + C) =: C_0$. Together with (III.48), we find:

$$\begin{aligned}
\mathbf{E}\left[\sum_{t=0}^{T-1} \mathbf{1}(Z_t = z_0, t \in \mathscr{T}^+)\right] &\leq \mathbf{E}\left[\sum_{i=1}^{\infty} \mathbf{1}(\tau_{j(i)} < T)\big(\tau_{j(i+1)-1} - \tau_{j(i)} + 1\big)\right] \\
&= \mathbf{E}\left[\sum_{i=1}^{\infty} \mathbf{1}(\tau_{j(i)} < T)\mathbf{E}\Big[\tau_{j(i+1)-1} - \tau_{j(i)} + 1 \;\Big|\; S_0, A_0, R_0, \ldots, S_{\tau_{j(i)}}\Big]\right] \\
&\leq C_0 \mathbf{E}\left[\sum_{i=1}^{\infty} \mathbf{1}(\tau_{j(i)} < T)\right] = C_0 \mathbf{E}\left[\sum_{t=0}^{T-1} \mathbf{1}(\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) \neq \mathscr{Z}^{**}(M), t \in \mathscr{T}_0^+)\right].
\end{aligned}$$

This proves the claim. □

(**STEP 2**) *The number of initial exploitation times such that $\mathscr{Z}^{**}(M)$ is miss-estimated are is bounded in expectation:*

$$\mathbf{E}\left[\sum_{t=0}^{\infty} \mathbf{1}(\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) \neq \mathscr{Z}^{**}(M), t \in \mathscr{T}_0^+)\right] < \infty. \tag{III.49}$$

Proof. By definition of initial exploitation times, if $t \in \mathscr{T}_0^+$ then the algorithm has passed the co-exploration GLR test, meaning that for all $\hat{M}^\dagger \in \mathrm{Alt}^{\epsilon(t)}(\hat{M}_t)$ that coincides with $\hat{M}_t$ on the current skeleton $\mathscr{Z}_t$, we have

$$\psi_{N_t}(\hat{M}_t || \hat{M}^\dagger) := \sum_{z \in \mathscr{Z}} N_t(z) \mathrm{KL}_z(\hat{M}_t || \hat{M}^\dagger) \geq (1 + \delta(t)) \log(t). \tag{III.50}$$

Further assume that $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) \neq \mathscr{Z}^{**}(M)$. Consider the model M'_t given by $M'_t(z) := \hat{M}_t(x)$ for $z \in \mathscr{Z}_t$ and $M'_t(z) := M(x)$ otherwise. Since $\hat{M}_t \ll M$, we have $\hat{M}_t \ll M'_t$ as well. Introduce the good event:

$$\mathscr{E}_t := \left(\forall z \in \mathscr{Z}, N_t(z) \geq \log^2(t) \Rightarrow d_{\epsilon(t)}(\hat{M}_t(z)||M(z)) < \tfrac{\epsilon_0}{|\mathscr{Z}|}\right) \tag{III.51}$$

where $\epsilon_0 > 0$ is chosen small enough such that, when we have $d_{\epsilon(t)}(M'_t||M) < \epsilon_0$, it follows that $\mathscr{Z}_{**}^{\epsilon(t)}(M'_t) = \mathscr{Z}^{**}(M) \neq \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ (see Proposition III.12). But recall that M'_t and $\hat{M}_t$ are copies of one another on $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$, hence of the gain of any *unichain* policy converging to $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ is the same in $\hat{M}_t$ and M'_t. Moreover, choosing this policy uniform outside of its unique recurrent component guarantees that its recurrent component is unique and is the same in $\hat{M}_t$ and M'_t. And yet $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) \setminus \mathscr{Z}_{**}^{\epsilon(t)}(M'_t) \neq \varnothing$ so necessarily, $\mathscr{Z}^{**}(M'_t) \setminus \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) \neq \varnothing$. In other words, $M'_t \in \mathrm{Alt}^{\epsilon(t)}(\hat{M}_t)$. Moreover, the models $\hat{M}_t$ and M'_t coincide on the skeleton by construction. By (III.50), we see that, on $\mathscr{E}_t$,

$$(1+\delta(t))\log(t) \leq \psi_{N_t}(\hat{M}_t||M'_t) = \sum_{z \notin \mathscr{Z}_t} N_t(z)\mathrm{KL}_x(\hat{M}_t||M) \equiv \psi_{N'_t}(\hat{M}_t||M) \tag{III.52}$$

where N'_t is copy of N_t which is reset to zero on pairs of the skeleton. We conclude the proof with a combinatorial argument invoking Sanov's theorem. For $n \in \mathbf{N}^{\mathscr{Z}}$, denote $\hat{M}^n$ the observed model under the deterministic $N_t = n$, hence $\hat{M}^{N_t} = \hat{M}_t$ by definition. Let $\mathscr{M}^n$ the (discrete) space of MDPs where rewards and kernels at z are all the possible empirical distributions obtained with $n(z)$ samples. Denote $[\log^2(t)] := \{0, \dots, \lfloor \log^2(t) \rfloor\}$. We have:

$$\begin{aligned}
(*) &:= \mathbf{E}\Big[\mathbf{1}(\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) \neq \mathscr{Z}^{**}(M), t \in \mathscr{T}_0^+)\Big] \\
&\leq \mathbf{E}\Big[\mathbf{1}(\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) \neq \mathscr{Z}^{**}(M), t \in \mathscr{T}_0^+, \mathscr{E}_t)\Big] + \mathbf{E}\Big[\mathbf{1}(\mathscr{E}_t^c)\Big] \\
&\overset{(\dagger)}{\leq} \mathbf{E}\left[\mathbf{1}\left(\psi_{N'_t}(\hat{M}_t||M) := \sum_{z \notin \mathscr{Z}_t} N_t(z)\mathrm{KL}(\hat{M}_t||M) \geq (1+\delta(t))\log(t)\right)\right] + \mathbf{P}\big(\mathscr{E}_t^c\big) \\
&\overset{(\ddagger)}{=} \sum_{n \in [\log^2(t)]^{\mathscr{Z}}} \mathbf{E}\Big[\mathbf{1}\big(\psi_{N'_t}(\hat{M}_t||M) \geq (1+\delta(t))\log(t), (\forall z \notin \mathscr{Z}_t, N_t(z) = n(z)), (\forall z \in \mathscr{Z}_t, n(z) = 0)\big)\Big] + \mathbf{P}\big(\mathscr{E}_t^c\big) \\
&= \sum_{n \in [\log^2(t)]^{\mathscr{Z}}} \sum_{M' \in \mathscr{M}^n} \mathbf{E}\Big[\mathbf{1}\big(\psi_n(M'||M) \geq (1+\delta(t))\log(t)\big)\mathbf{1}\big(\hat{M}^n = M'\big)\Big] + \mathbf{P}\big(\mathscr{E}_t^c\big) \\
&= \sum_{n \in [\log^2(t)]^{\mathscr{Z}}} \sum_{M' \in \mathscr{M}^n} \mathbf{1}\big(\psi_n(M'||M) \geq (1+\delta(t))\log(t)\big)\mathbf{P}(\hat{M}^n = M') + \mathbf{P}\big(\mathscr{E}_t^c\big) \\
&\overset{(\S)}{\leq} \sum_{n \in [\log^2(t)]^{\mathscr{Z}}} \sum_{M' \in \mathscr{M}^n} \mathbf{1}\left(\sum_{z \in \mathscr{Z}} n(z)\mathrm{KL}_z(M'||M) \geq (1+\delta(t))\log(t)\right)\exp\left(-\sum_{z \in \mathscr{Z}} n(z)\mathrm{KL}_z(M'||M)\right) + \mathbf{P}\big(\mathscr{E}_t^c\big) \\
&\leq \sum_{n \in [\log^2(t)]^{\mathscr{Z}}} \sum_{M' \in \mathscr{M}^n} \exp(-(1+\delta(t))\log(t)) + \mathbf{P}\big(\mathscr{E}_t^c\big) \\
&\overset{(\$)}{=} \big(1+\lfloor\log^2(t)\rfloor\big)^{|\mathscr{Z}|}\big(1+\lfloor\log^2(t)\rfloor\big)^{|\mathscr{Z}||\mathscr{S}|}\big(1+\lfloor\log^2(t)\rfloor\big)^{2|\mathscr{Z}|}\left(\frac{1}{t}\right)^{1+\delta(t)} + \mathbf{P}\big(\mathscr{E}_t^c\big) \\
&\overset{(\#)}{=} o\left(\frac{1}{t\log^2(t)}\right) + \mathbf{P}\big(\mathscr{E}_t^c\big).
\end{aligned}$$

In the above, (†) is obtained with (III.52), (‡) follows by construction of N'_t, (§) is a consequence of an all-time Sanov theorem (Lemma III.33), ($) follows from classical combinatorial bounds (Lemma III.34) and (#) is obtained by expanding the definition of $\delta(t)$ and using $\log\log(t) = o(\log(t)^{1/2})$. By Lemma III.32, we know that $\sum_{t=1}^{\infty} \mathbf{P}(\mathscr{E}_t^c) < \infty$. We conclude accordingly that $\sum_{t=0}^{\infty} \mathbf{E}[\mathbf{1}(\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) \neq \mathscr{Z}^{**}(M), t \in \mathscr{T}_0^+)] < \infty$. □

We conclude by combining (III.47) and (III.49). ■

10.C.3 Proof of Lemma III.24: Amount of wrong co-exploration

Fix $z \notin \mathscr{Z}^{**}(M)$.

Decomposition with good events. If the algorithm co-explores, then it has passed the exploration GLR test, meaning that for all $\hat{M}^\dagger \in \mathrm{Alt}^{\epsilon(t)}(\hat{M}_t)$ that coincides with $\hat{M}_t$ on the current skeleton $\mathscr{Z}_t$ and nearly optimal pairs $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$, we have:

$$\psi_{N_t}(\hat{M}_t||\hat{M}^\dagger) := \sum_{z\in\mathscr{Z}} N_t(z)\mathrm{KL}_z(\hat{M}_t||\hat{M}^\dagger) \geq (1+\delta)\log(t). \tag{III.53}$$

Consider the model M'_t given by $M'_t(z) := \hat{M}_t(z)$ for $z \in \mathscr{Z}_t \cup \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ and $M'_t(z) = M(z)$ otherwise. Introduce the events:

$$\begin{aligned}
\mathscr{E}_t &:= (\forall z \in \mathscr{Z},\ \ N_t(z) \geq \log^2(t) \Rightarrow d_{\epsilon(t)}(\hat{M}_t(z)||M(z) < \tfrac{\epsilon_0}{|\mathscr{Z}|}) \\
\mathscr{E}'_t &:= (\forall z \in \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t),\ \ d_{\epsilon(t)}(\hat{M}_t(z)||M(z)) < \tfrac{\epsilon_0}{|\mathscr{Z}|})
\end{aligned} \tag{III.54}$$

The first event $\mathscr{E}_t$ states that the data is approximately correct on the skeleton and the second $\mathscr{E}'_t$ that it is approximately correct on nearly optimal pairs. We decompose $\mathbf{1}(Z_t = z, t \in \mathscr{T}^\pm)$ as follows:

$$\mathbf{E}[\mathbf{1}(Z_t = z, t \in \mathscr{T}^\pm)] \leq \mathbf{E}[\mathbf{1}(Z_t = z, t \in \mathscr{T}^\pm, \mathscr{E}_t, \mathscr{E}'_t)] + \mathbf{E}[\mathbf{1}(Z_t = z, t \in \mathscr{T}^\pm, \mathscr{E}_t'^c)] + \mathbf{E}[\mathbf{1}(\mathscr{E}_t^c)]$$

The first term $\mathbf{E}[\mathbf{1}(Z_t = z, t \in \mathscr{T}^\pm, \mathscr{E}_t, \mathscr{E}'_t)]$ is handled similarly to the wrong exploitation term (Section 10.C.2) and is shown to be $o(t^{-1}\log^{-2}(t))$. The third term $\mathbf{E}[\mathbf{1}(\mathscr{E}_t^c)]$ is shown to be $o(t^{-2})$ with Lemma III.32. The dominant term is the second one, $\mathbf{E}[\mathbf{1}(Z_t = z, t \in \mathscr{T}^\pm, \mathscr{E}_t'^c)]$ and account for the speed at which co-exploration gather information on $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$. It is shown to be $O(\log\log(T)^3)$, see Lemmas III.27 and III.28.

Introducing the co-exploration structure $\mathfrak{Z}_t^\pm$. When a co-exploration phase begins, it last until regeneration or until regeneration is compromised because of the discovery of a new transition that provokes a panic time. We therefore take into account the current *co-exploration structure*. Co-exploration times $\mathscr{T}^\pm$ are further refined by taking account of the current nearly optimal pairs $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$. By design, the algorithm splits $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ into components $\mathscr{Z}_t^1, \dots, \mathscr{Z}_t^{m(t)}$ that are the communicating components of $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ under the empirically observed model $\hat{M}_t$.[2] The decomposition itself is denoted $\mathfrak{Z}_t^\pm := (\mathscr{Z}_t^1, \dots, \mathscr{Z}_t^{m(t)})$ and is called the *co-exploration structure* at time t. Given a co-exploration structure $\mathfrak{Z}_0$, we introduce $\mathscr{T}^\pm(\mathfrak{Z}_0)$ the co-exploration times where the co-exploration structure is $\mathfrak{Z}_0$ at the associated initial co-exploration time:

$$\mathscr{T}^\pm(\mathfrak{Z}_0) := \left\{ t \in \mathscr{T}^\pm : \mathfrak{Z}_{t_0}^\pm = \mathfrak{Z}_0 \text{ with } t_0 := \sup\left\{ t' : t' \in \mathscr{T}_0^\pm \right\} \right\} \tag{III.55}$$

Depending on whether the co-exploration structure $\mathfrak{Z}_0$ is closed in M (i.e., every $\mathscr{Z}_0 \in \mathfrak{Z}_0$ is communicating in M), the analysis is different. If $\mathfrak{Z}_0$ is not closed, it means that the support of the empirical kernel is off on one of the components $\mathscr{Z}_{t_0}^i \in \mathfrak{Z}_{t_0}^\pm$ of $\mathscr{Z}_{**}^{\epsilon(t_0)}(\hat{M}_{t_0})$ and the algorithm will quickly figure it out, hence the associated number of co-exploration times is

[2] A collection of pairs $\mathscr{Z}' \subseteq \mathscr{Z}$ is said *communicating* if $M|_{\mathscr{Z}'}$, obtained by restricting states to $\mathscr{S}(\mathscr{Z}')$ and playable actions from s to $\{s\} \times \mathscr{A}(s) \cap \mathscr{Z}'$, is well-defined and communicating. A *communicating component under* $\hat{M}$ of $\mathscr{Z}'$ is a maximal set $\mathscr{Z}'' \subseteq \mathscr{Z}'$ such that $\mathscr{Z}''$ is communicating for $\hat{M}$.

small (Section 10.C.3.2, see Lemma III.28). If it is closed, we show that co-exploration ensures uniform visit guarantees on $\mathscr{Z}_{**}^{\epsilon(t_0)}(\hat{M}_{t_0})$ and $\mathscr{E}_t'^c$ cannot hold for too long (Section 10.C.3.1, see Lemma III.27). We show that:

$$\begin{aligned}
(*) &:= \sum_{t=0}^{T-1} \mathbf{E}[\mathbf{1}(Z_t = z, t \in \mathscr{T}^{\pm})] \\
&\leq \sum_{t=0}^{\infty} \mathbf{E}[\mathbf{1}(Z_t = z, t \in \mathscr{T}^{\pm}, \mathscr{E}_t, \mathscr{E}_t')] + \sum_{t=0}^{T-1} \mathbf{E}[\mathbf{1}(Z_t = z, t \in \mathscr{T}^{\pm}, \mathscr{E}_t'^c)] + \sum_{t=0}^{\infty} \mathbf{E}[\mathbf{1}(\mathscr{E}_t^c)] \\
&\overset{(\dagger)}{=} \mathrm{O}(1) + \sum_{t=0}^{T-1} \mathbf{E}[\mathbf{1}(Z_t = z, t \in \mathscr{T}^{\pm}, \mathscr{E}_t'^c)] + \mathrm{O}(1) \\
&\leq \sum_{\mathfrak{Z}_0 \text{ closed}} \sum_{t=0}^{T-1} \mathbf{E}[\mathbf{1}(Z_t = z, t \in \mathscr{T}^{\pm}(\mathfrak{Z}_0), \mathscr{E}_t'^c)] + \sum_{\mathfrak{Z}_0 \text{ not closed}} \sum_{t=0}^{\infty} \mathbf{E}[\mathbf{1}(Z_t = z, t \in \mathscr{T}^{\pm}(\mathfrak{Z}_0))] + \mathrm{O}(1) \\
&\overset{(\ddagger)}{=} \mathrm{O}\big(\log\log(T)^3\big) + \mathrm{O}(1) + \mathrm{O}(1) = \mathrm{O}\big(\log\log(T)^3\big)
\end{aligned}$$

which is the desired result. In the above, (†) follows with the same technique than in the analysis of the amount of wrong exploitation (Section 10.C.2) and (‡) follows by Lemma III.28 and Lemma III.27. The remaining of the section is dedicated to a proof of Lemma III.28 and Lemma III.27.

10.C.3.1 Co-exploration with correct kernel supports

We start with the dominant term (Lemma III.27). The proof of Lemma III.28 will share many similarities.

Lemma III.27 (Co-exploration with correct supports). *Let $\mathfrak{Z}_0$ a closed co-exploration structure. We have:*

$$\mathbf{E}\left[\sum_{t=0}^{T-1} \mathbf{1}(t \in \mathscr{T}^{\pm}(\mathfrak{Z}_0), \mathscr{E}_t'^c)\right] = \mathrm{O}\big(\log\log(T)^3\big). \tag{III.56}$$

Proof of Lemma III.27. Fix $\mathfrak{Z}_0 = (\mathscr{Z}_0^1, \dots, \mathscr{Z}_0^m)$ a closed co-exploration structure and denote (τ_j) the stopping-time enumeration of $\mathscr{T}^{\pm}(\mathfrak{Z}_0)$. Let $\mathscr{T}_0^{\pm}(\mathfrak{Z}_0) := \{t \in \mathscr{T}_0^{\pm} : \mathfrak{Z}_t^{\pm} = \mathfrak{Z}_0\}$ the initial co-exploration associated to $\mathfrak{Z}_0$ and let $\tau_{j(i)}$ its i-th element. By definition, every $\mathscr{Z}_0^u$ is a communicating component of $\mathfrak{Z}_t^{\pm}$ under $\hat{M}_t$ when $\mathfrak{Z}_t^{\pm} = \mathfrak{Z}_0$; By assumption, $\mathscr{Z}_0^u$ is also communicating in M. Given a component u, we write $\mathscr{I}^u := \{i : S_{\tau_{j(i)}} \in \mathscr{S}(\mathscr{Z}_0^u)\}$ the index of initial co-exploration times starting in the component u.

(**STEP 1**) *Fix a component $u \in \{1, \dots, m\}$ and let $z_0 \in \mathscr{Z}_0^u$. There exists $\alpha > 0$ such that, for all $\ell \geq 1$,*

$$\mathbf{P}\Big(N_{\tau_{j(i+1)}}(z_0) < \alpha\ell \text{ and } \ell \leq |\mathscr{I}^u \cap \{1, \dots, i\}|\Big) = \mathrm{o}\big(\ell^{-2}\big). \tag{III.57}$$

Proof. Fix $z_0 \in \mathscr{Z}_0^u$ and introduce the reward function $f(z) := \mathbf{1}(z = z_0)$ and let g^f, h^f the associated gain and bias function obtained by iterating the uniform policy on $\mathscr{Z}_0^u$. We see that $\mathrm{sp}(g^f) = 0$ because $\mathscr{Z}_0^u$ is a communicating component. Denote $\beta := \mathrm{sp}(h^f)$ and $\alpha := \min(g^f) > 0$. Recall that $[i] := \{1, \dots, i\}$. We have:

$$N_{\tau_{j(i+1)}}(z_0) \geq \sum_{i' \in \mathscr{I}^u \cap [i]} \sum_{j'=j(i')}^{j(i'+1)-1} f(S_{\tau_{j'}}, A_{\tau_{j'}})$$

$$\overset{(\dagger)}{=} \sum_{i' \in \mathscr{I}^u \cap [i]} \sum_{j'=j(i')}^{j(i'+1)-1} \Big(g^f(S_{\tau_{j'}}) + \Big(e_{S_{\tau_{j'}}} - p(S_{\tau_{j'}}, A_{\tau_{j'}}) \Big) h^f \Big)$$

$$\overset{(\ddagger)}{=} \alpha \Bigg(\sum_{i' \in \mathscr{I}^u \cap [i]} (\tau_{j(i'+1)-1} - \tau_{j(i')} + 1) \Bigg) + \sum_{i' \in \mathscr{I}^u \cap [i]} \sum_{j'=j(i')}^{j(i'+1)-1} \Big(e_{S_{\tau_{j'+1}}} - p(S_{\tau_{j'}}, A_{\tau_{j'}}) \Big) h^f$$

$$\overset{(\S)}{\geq} \alpha T_i^u - \beta \sqrt{T_i^u \log\left(\frac{\sqrt{1+T_i^u}}{\delta} \right)}$$

where (†) invokes the Poisson equation $g^f(s,a) + h^f = f(s,a) + p(s,a)h^f$; (‡) uses the regeneration guarantees of co-exploration and (§) bounds the RHS martingale with a time-uniform Azuma-Hoeffding inequality (Lemma I.22) with probability $1-\delta$, and T_i^u is a short-hand for $\sum_{i' \in \mathscr{I}^u \cap [i]} (\tau_{j(i'+1)-1} - \tau_{j(i')} + 1)$. In particular, we have $T_i^u \geq |\mathscr{I}^u \cap [i]|$. Remark that:

$$\beta \sqrt{T_i^u \log\left(\frac{\sqrt{1+T_i^u}}{\delta} \right)} < \frac{1}{2} \alpha T_i^u \iff \delta > \exp\left(-\frac{\alpha \sqrt{T_i^u}}{2\beta} + \frac{1}{2} \log(1+T_i^u) \right) = o\big(|\mathscr{I}^u \cap [i]|^{-2} \big).$$

We conclude accordingly. □

(**STEP 2**) *There exists a constant $\alpha > 0$ such that:*

$$\forall i \geq 1, \forall z \in \bigcup \mathfrak{Z}_0, \quad \mathbf{P}\Big(N_{\tau_{j(i)}}(z) < \alpha \sqrt{i} \Big) = o\big(i^{-2}\big). \tag{III.58}$$

Proof. By the pigeon-hole principle, there exists a component u such that $|\mathscr{I}^u \cap [i]| \geq \lceil \frac{i}{m} \rceil$. So, for $i' < i$ the component u was chosen to start a co-exploration phase, i.e., there exists $i' < i$ such that:

$$\log \min \Big\{ N_{\tau_{j(i')}}(z) : z \in \mathscr{Z}_0^u \Big\} \leq 2 \log \min \Big\{ N_{\tau_{j(i')}}(z) : z \in \bigcup \mathfrak{Z}_0 \Big\} \tag{III.59}$$

and such that $|\mathscr{I}^u \cap [i']| \geq \lceil \frac{i}{m} \rceil - 1$. By (III.57), we have:

$$\mathbf{P}\left(\log \min \Big\{ N_{\tau_{j(i')}}(z) : z \in \mathscr{Z}_0^u \Big\} \geq \log\left(\alpha \cdot \frac{i-m}{m} \right) \right) = 1 - o\big(i^{-2}\big). \tag{III.60}$$

Because visit counts are monotone with respect to time, combining (III.59) and (III.60) we obtain that with probability $1 - o(i^{-2})$,

$$\min \Big\{ N_{\tau_{j(i')}}(z) : z \in \mathscr{Z}_0 \Big\} \geq \exp\left(\frac{1}{2} \log\left(\alpha \cdot \frac{i-m}{m} \right) \right) = \Omega\big(\sqrt{i}\big).$$

This proves the claim. □

Remark than in (III.58), the number of visits of z at the co-exploration time $\tau_{j(i)}$ is controlled relatively to the number of initial co-exploration times prior to $\tau_{j(i)}$, which is i. We know relate $j(i)$ and i.

(**STEP 3**) *There exists $\alpha, \beta > 0$ such that $\mathbf{P}(\alpha i < j(i) < \beta i) = 1 - o(i^{-2})$. In other words, the number of initial co-exploration times grows at the same speed than the number of co-exploration times.*

Proof. Consider a initial co-exploration $\tau_{j(i)}$ and denote $\mathscr{Z}_0^u$ the current co-explored component. Consider the reward function $f(s,a) := \mathbf{1}(s = S_{\tau_{j(i)}})$ and let g^f, h^f the associated gain and bias functions obtained by iterating the uniform policy on the current component $\mathscr{Z}_0^u$. We see that

$\mathrm{sp}(g^f)=0$ because $\mathscr{Z}_0^u$ is closed. Denote $\alpha_1 := \mathrm{sp}(h^f)$ and $\alpha_2 := \min(g^f) > 0$. Thanks to the regenerative design of co-exploration phases, we have:

$$\begin{aligned}
1 &= \sum_{t=\tau_{j(i)}}^{\tau_{j(i+1)-1}} f(S_t,A_t) \\
&\overset{(\dagger)}{=} \sum_{t=\tau_{j(i)}}^{\tau_{j(i+1)-1}} \big(g^f(S_t) + \big(e_{S_t} - p(S_t,A_t)\big)h^f\big) \\
&= \alpha_1(\tau_{j(i+1)-1} - \tau_{j(i)} + 1) + \sum_{t=\tau_{j(i)}}^{\tau_{j(i+1)-1}} \big(e_{S_{t+1}} - p(S_t,A_t)\big)h^f .
\end{aligned}$$

In the above, (†) follows by the Poisson equation $g^f(s,a) + h^f = f(s,a) + p(s,a)h^f$. Summing for i, we find:

$$\begin{aligned}
i &= \alpha_1 j(i+1) + \sum_{i'=1}^{i} \sum_{t=\tau_{j(i')}}^{\tau_{j(i'+1)-1}} \big(e_{S_{t+1}} - p(S_t,A_t)\big)h^f \\
&\overset{(\dagger)}{\geq} \alpha_1 j(i+1) - \alpha_2 \sqrt{j(i+1)\log\left(\frac{\sqrt{1+j(i+1)}}{\delta}\right)}
\end{aligned}$$

where (†) bounds the RHS martingale with a time-uniform Azuma-Hoeffding inequality (Lemma I.22) and holds with probability $1-\delta$. We get a similar lower-bound. We conclude similarly than in (**STEP 1**) by finding a condition on $\delta > 0$ such that the error term is lower than $\frac{1}{2}\alpha_1 j(i)$ and consider the worst α_1, α_2 possible (depending on the current component u and initial state S_{τ_i}). □

(**STEP 4**) *Recall that (τ_j) is the stopping-time enumeration of $\mathscr{T}^{\pm}(\mathfrak{Z}_0)$. We have*

$$\mathbf{E}\left[\sum_{t=0}^{T-1} \mathbf{1}(t \in \mathscr{T}^{\pm}(\mathfrak{Z}_0), \mathscr{E}_t'^c)\right] \leq \mathbf{E}\left[\sum_{j=1}^{\infty}\sum_{z\in\mathscr{Z}} \mathbf{1}\Big(\mathrm{KL}^*(\hat{M}_{\tau_j}(z)||M(z)) < \tfrac{\epsilon_0}{4|\mathscr{Z}|\log\log(T)} \text{ and } N_{\tau_j}(z) < \alpha\sqrt{j}\Big)\right] + \mathrm{O}(1). \tag{III.61}$$

Proof. By combining the results of (**STEP 2**) and (**STEP 3**), we find that there exists $\alpha > 0$ such that:

$$\forall i \geq 1, \forall z \in \bigcup \mathfrak{Z}_0, \quad \mathbf{P}\big(N_i(z) < \alpha\sqrt{i}\big) = \mathrm{o}\big(i^{-2}\big). \tag{III.62}$$

We write:

$$\begin{aligned}
(*) &:= \mathbf{E}\left[\sum_{t=0}^{T-1} \mathbf{1}(t \in \mathscr{T}^{\pm}(\mathfrak{Z}_0), \mathscr{E}_t'^c)\right] \\
&= \mathbf{E}\left[\sum_{j=1}^{\infty} \mathbf{1}(\tau_j < T, \mathscr{E}_{\tau_j}'^c)\right] \\
&\leq \mathbf{E}\left[\sum_{j=1}^{\infty} \mathbf{1}(\tau_j < T, \mathscr{E}_{\tau_j}'^c, (\forall z \in \mathfrak{Z}_0, N_{\tau_j}(z) \geq \alpha\sqrt{j}))\right] + \mathbf{E}\left[\sum_{j=1}^{\infty} \mathbf{1}(\exists z \in \mathfrak{Z}_0, N_{\tau_j}(z) < \alpha\sqrt{j})\right] \\
&\overset{(\dagger)}{=} \mathbf{E}\left[\sum_{j=1}^{\infty} \mathbf{1}\left(\tau_j < T, \begin{pmatrix} \exists z \in \bigcup \mathfrak{Z}_0, d_{\epsilon(\tau_j)}(\hat{M}_{\tau_j}(z)||M(z)) \geq \frac{\epsilon_0}{|\mathscr{Z}|} \big) \\ (\forall z \in \bigcup \mathfrak{Z}_0, N_{\tau_j}(z) \geq \alpha\sqrt{j}) \end{pmatrix}\right] + \sum_{j=1}^{\infty} \mathrm{o}\big(j^{-2}\big)
\end{aligned}$$

where (†) follows by (III.62). We focus on the second term. We simplify a little bit the event involving $d_{\epsilon(t)}(-)$. By expanding its definition and invoking Pinsker's inequality, we have:

$$d_\epsilon(M'(z)||M(z)) \leq \mathrm{KL}^*(M'(z)||M(z)) + \tfrac{1}{\epsilon}\sqrt{2\mathrm{KL}^*(M'(z)||M(z))}.$$

Provided that $\epsilon_0 < 1$, we can continue the previous computation with:

$$\begin{aligned}(*) &\leq \mathbf{E}\left[\sum_{j=1}^{\infty}\mathbf{1}\left(\tau_j < T, \begin{pmatrix}\exists z \in \bigcup \mathfrak{Z}_0, \frac{4}{\epsilon(\tau_j)}\mathrm{KL}^*(\hat{M}(\tau_j)(z)||M(z)) < \frac{\epsilon_0}{|\mathscr{Z}|}\big)\\ (\forall z \in \bigcup \mathfrak{Z}_0, N_{\tau_j}(z) \geq \alpha\sqrt{j})\end{pmatrix}\right)\right] + \mathrm{O}(1)\\ &\leq \mathbf{E}\left[\sum_{j=1}^{\infty}\mathbf{1}\left(\begin{pmatrix}\exists z \in \bigcup \mathfrak{Z}_0, \mathrm{KL}^*(\hat{M}_{\tau_j}(z)||M(z)) < \frac{\epsilon_0}{4|\mathscr{Z}|\log\log(T)}\big),\\ (\forall z \in \bigcup \mathfrak{Z}_0, N_{\tau_j}(z) \geq \alpha\sqrt{j})\end{pmatrix}\right)\right] + \mathrm{O}(1).\end{aligned}$$

This proves the claim. □

We finally prove the result. Starting by invoking (III.61), we have:

$$\begin{aligned}(*) &:= \mathbf{E}\left[\sum_{t=0}^{T-1}\mathbf{1}(t \in \mathscr{T}^{\pm}(\mathfrak{Z}_0), \mathscr{E}_t'^c)\right]\\ &\leq \mathbf{E}\left[\sum_{j=1}^{\infty}\sum_{z\in\mathscr{Z}}\mathbf{1}\Big(\mathrm{KL}^*(\hat{M}_{\tau_j}(z)||M(z)) < \tfrac{\epsilon_0}{4|\mathscr{Z}|\log\log(T)} \text{ and } N_{\tau_j}(z) < \alpha\sqrt{j}\Big)\right] + \mathrm{O}(1)\\ &\overset{(\dagger)}{=} \sum_{z\in\mathscr{Z}}\sum_{j=1}^{\infty}\mathbf{P}\Big(\mathrm{KL}(\hat{M}_{\tau_j}(z)||M(z)) < \tfrac{\epsilon_0}{4|\mathscr{Z}|\log\log(T)} \text{ and } N_{\tau_j}(z) < \alpha\sqrt{j}\Big) + \mathrm{O}(1)\\ &\overset{(\ddagger)}{\leq} 2\sum_{z\in\mathscr{Z}}\sum_{j=1}^{\infty}\left(\exp\left(-\frac{\epsilon_0\sqrt{j}}{4|\mathscr{S}||\mathscr{Z}|\log\log(T)} + \log\big(1+\sqrt{j}\big) + 1\right)\wedge 1\right) + \mathrm{O}(1)\end{aligned}$$

where (†) converts KL^* to KL using Lemma III.37 (which is applicable provided that $\epsilon_0/(4|\mathscr{Z}|\log\log(T))$ is small enough) and (‡) by Lemma III.36. Setting $C := \frac{\epsilon_0}{4|\mathscr{S}||\mathscr{Z}|}$, we have to upper-bound a term of the form:

$$\psi(T) := \sum_{n=1}^{\infty}\left(\big(1+\sqrt{n}\big)\exp\left(-\frac{C\sqrt{n}}{\log\log(T)}\right)\right)\wedge 1 \tag{III.63}$$

Set $\alpha := \frac{C}{\log\log(T)}$. We have:

$$\begin{aligned}\psi(T) &\leq 2\sum_{n=1}^{\infty}\big(\sqrt{n}\exp\big(-\alpha\sqrt{n}\big)\big)\wedge 1 = 2\left(\left\lceil\frac{1}{\alpha^2}\right\rceil + \sum_{n>\lceil\alpha^{-2}\rceil}^{\infty}\sqrt{n}\exp\big(-\alpha\sqrt{n}\big)\right)\\ &\overset{(\dagger)}{\leq} 2\left(\left\lceil\frac{1}{\alpha^2}\right\rceil + \int_0^{\infty}\sqrt{x}\exp\big(-\alpha\sqrt{x}\big)dx\right)\\ &= 2\left(\left\lceil\frac{1}{\alpha^2}\right\rceil + \frac{2}{\alpha^3}\int_0^{\infty}x^2\exp(-x)dx\right) = 2\left(\left\lceil\frac{1}{\alpha^2}\right\rceil + \frac{4}{\alpha^3}\right) = \mathrm{O}\big(\log\log(T)^3\big)\end{aligned}$$

where (†) is a sum-integral comparison, using the fact that $f(n) := \sqrt{n}\exp(-\alpha\sqrt{n})$ is shown to be decreasing on (α^{-2}, ∞). Accordingly, we end up with

$$\mathbf{E}\left[\sum_{t=0}^{T-1}\mathbf{1}(t \in \mathscr{T}^{\pm}(\mathfrak{Z}_0), \mathscr{E}_t'^u)\right] = \mathrm{O}\big(\log\log(T)^3\big). \tag{III.64}$$

This is the desired result. ■

10.C.3.2 Co-exploration with wrong kernel supports

The proof shares many similarities with the one of Lemma III.27, see Section 10.C.3.1.

Lemma III.28 (Co-exploration with wrong supports). *Let $\mathfrak{Z}_0$ a non-closed co-exploration structure. Then:*

$$\mathbf{E}\left[\sum_{t=0}^{\infty} \mathbf{1}(t \in \mathscr{T}^{\pm}(\mathfrak{Z}_0))\right] < \infty. \tag{III.65}$$

Proof of Lemma III.28. Fix $\mathfrak{Z}_0 = (\mathscr{Z}_0^1, \dots, \mathscr{Z}_0^m)$ a non closed set of pairs and denote (τ_j) the stopping-time enumeration of $\mathscr{T}^{\pm}(\mathfrak{Z}_0)$. Let $\mathscr{T}_0^{\pm}(\mathfrak{Z}_0) := \{t \in \mathscr{T}_0^{\pm} : \mathfrak{Z}_t^{\pm} = \mathfrak{Z}_0\}$ the initial co-exploration associated to $\mathscr{Z}_0$ and let $\tau_{j(i)}$ its i-th element. Given a component u, we write $\mathscr{I}^u := \{i : S_{\tau_{j(i)}} \in \mathscr{S}(\mathscr{Z}_0^u)\}$ the index of initial co-exploration times starting in the component u. Because $\mathfrak{Z}_0$ is not closed for M, some components $\mathscr{Z}_0^u$ are not forwardly closed in M.[3] Intuitively speaking, these components cannot be visited too much while keeping the co-exploration structure $\mathfrak{Z}_0$, because breaking transitions are found quickly. We start with a few facts that echo the proof of Lemma III.27.

(**STEP 1**) *Fix a component $u \in \{1, \dots, m\}$ and let $z_0 \in \mathscr{Z}_0^u$. There exists $\alpha > 0$ such that, for all $\ell \geq 1$,*

$$\mathbf{P}\Big(N_{\tau_{j(i+1)}}(z_0) < \alpha\ell \text{ and } \ell \leq |\mathscr{I}^u \cap \{1, \dots, i\}|\Big) = \mathrm{o}\big(\ell^{-2}\big). \tag{III.66}$$

Proof. The proof of (III.57) has to be adapted. Fix $z_0 \in \mathscr{Z}_0^u$ and introduce the reward function $f(s,a) := \mathbf{1}((s,a) = z_0 \text{ or } s \notin \mathscr{S}(\mathscr{Z}_0^u))$ tracking the visits of z_0 plus the possibility to exit the component $\mathscr{Z}_0^u$. Let g^f, h^f the associated gain and bias function obtained by iterating the uniform policy on $\mathscr{Z}_0^u$, extended to the uniform policy outside of $\mathscr{S}(\mathscr{Z}_0^u)$. We see that $\mathrm{sp}(g^f|_{\mathscr{S}(\mathscr{Z}_0^u)}) = 0$ because $\mathscr{Z}_0^u$ regardless of whether $\mathscr{Z}_0^u$ is communicating in M or not. Denote $\beta := \mathrm{sp}(h^f)$ and $\alpha := \min(g^f) > 0$.

By design of panic times, there can by at most one $i' \in \mathscr{I}^u$ such that the co-exploration episode starting at $t = \tau_{j(i')}$ is not regenerative, i.e., $S_{\tau_{j(i')}} \neq S_{\tau_{j(i'+1)-1}+1}$.

We have:

$$\begin{aligned}
N_{\tau_{j(i+1)}}(z_0) &\geq \sum_{i' \in \mathscr{I}^u \cap [i]} \sum_{j'=j(i')}^{j(i'+1)-1} f(S_{\tau_{j'}}, A_{\tau_{j'}}) \\
&\overset{(\dagger)}{=} \sum_{i' \in \mathscr{I}^u \cap [i]} \sum_{j'=j(i')}^{j(i'+1)-1} \Big(g^f(S_{\tau_{j'}}) + \Big(e_{S_{\tau_{j'}}} - p(S_{\tau_{j'}}, A_{\tau_{j'}})\Big) h^f\Big) \\
&\overset{(\ddagger)}{=} \alpha\left(\sum_{i' \in \mathscr{I}^u \cap [i]} (\tau_{j(i'+1)-1} - \tau_{j(i')} + 1)\right) - \beta + \sum_{i' \in \mathscr{I}^u \cap [i]} \sum_{j'=j(i')}^{j(i'+1)-1} \Big(e_{S_{\tau_{j'+1}}} - p(S_{\tau_{j'}}, A_{\tau_{j'}})\Big) h^f \\
&\overset{(\S)}{\geq} \alpha T_i^u - \beta\left(1 + \sqrt{T_i^u \log\left(\frac{\sqrt{1+T_i^u}}{\delta}\right)}\right)
\end{aligned}$$

where (†) invokes the Poisson equation $g^f(s,a) + h^f = f(s,a) + p(s,a)h^f$; (‡) uses that at most one episode isn't regenerative and (§) bounds the RHS martingale with a time-uniform

[3] Recall that $\mathscr{Z}' \subseteq \mathscr{Z}$ is *forward closed* in M if by starting in a state of $\mathscr{S}(\mathscr{Z}')$ and only playing pairs of $\mathscr{Z}'$, one remains in $\mathscr{Z}'$.

Azuma-Hoeffding inequality (Lemma I.22) with probability $1-\delta$, and T_i^u is a short-hand for $\sum_{i'\in\mathscr{I}^u\cap[i]}(\tau_{j(i'+1)-1}-\tau_{j(i')}+1)$. If T_i^u is large enough, we have $\frac{1}{3}\alpha T_i^u \geq \beta$ since $T_i^u \geq i$, and

$$\beta\sqrt{T_i^u \log\left(\frac{\sqrt{1+T_i^u}}{\delta}\right)} < \frac{1}{3}\alpha T_i^u \iff \delta > \exp\left(-\frac{\alpha\sqrt{T_i^u}}{3\beta}+\frac{1}{3}\log(1+T_i^u)\right) = \mathrm{o}\big(|\mathscr{I}^u\cap[i]|^{-2}\big).$$

We conclude accordingly. □

(**STEP 2**) *There exists a constant $\alpha>0$ such that:*

$$\forall i\geq 1, \forall z\in\bigcup\mathfrak{Z}_0,\quad \mathbf{P}\Big(N_{\tau_{j(i)}}(z)<\alpha\sqrt{i}\Big)=\mathrm{o}\big(i^{-2}\big). \tag{III.67}$$

Proof. Same proof than (III.58). □

(**STEP 3**) *There exists $\alpha,\beta>0$ such that $\mathbf{P}(\alpha i<j(i)<\beta i)=1-\mathrm{o}(i^{-2})$. In other words, the number of initial co-exploration times grows at the same speed than the number of co-exploration times.*

Proof. The proof is an adaptation of the (**STEP 3**) of Section 10.C.3.1 taking into account that component may not be communicating and are subject to panicking. Consider a initial co-exploration $\tau_{j(i)}$ and denote $\mathscr{Z}_0^u$ the current co-explored component. Consider the reward function $f(s,a):=\mathbf{1}(s=S_{\tau_{j(i)}}$ or $s\notin\mathscr{S}(\mathscr{Z}_0^u))$ and let g^f,h^f the associated gain and bias functions obtained by iterating the uniform policy on the current component $\mathscr{Z}_0^u$. That is, f tracks the return to the state from which the co-exploration was initiated together with exits of the current component $\mathscr{Z}_0^u$. We see that $\mathrm{sp}(g^f|_{\mathscr{S}(\mathscr{Z}_0^u)})=0$ because $\mathscr{Z}_0^u$ is closed. Denote $\alpha_1:=\mathrm{sp}(h^f)$ and $\alpha_2:=\min(g^f)>0$. By design of co-exploration phases, observe that f can be equal to 1 only once per phase (at initialization). We get:

$$\begin{aligned}
1&=\sum_{t=\tau_{j(i)}}^{\tau_{j(i+1)-1}} f(S_t,A_t)\\
&\overset{(\dagger)}{=}\sum_{t=\tau_{j(i)}}^{\tau_{j(i+1)-1}}\Big(g^f(S_t)+\big(e_{S_t}-p(S_t,A_t)\big)h^f\Big)\\
&=\alpha_1(\tau_{j(i+1)-1}-\tau_{j(i)}+1)+\sum_{t=\tau_{j(i)}}^{\tau_{j(i+1)-1}}\big(e_{S_{t+1}}-p(S_t,A_t)\big)h^f.
\end{aligned}$$

In the above, (†) follows by the Poisson equation $g^f(s,a)+h^f=f(s,a)+p(s,a)h^f$. Summing for i, we find:

$$\begin{aligned}
i&=\alpha_1 j(i+1)+\sum_{i'=1}^{i}\sum_{t=\tau_{j(i')}}^{\tau_{j(i'+1)-1}}\big(e_{S_{t+1}}-p(S_t,A_t)\big)h^f\\
&\overset{(\dagger)}{\geq}\alpha_1 j(i+1)-\alpha_2\sqrt{j(i+1)\log\left(\frac{\sqrt{1+j(i+1)}}{\delta}\right)}
\end{aligned}$$

where (†) bounds the RHS martingale with a time-uniform Azuma-Hoeffding inequality (Lemma I.22) and holds with probability $1-\delta$. We get a similar lower-bound. We conclude similarly than in (**STEP 1**) by finding a condition on $\delta>0$ such that the error term is lower than $\frac{1}{2}\alpha_1 j(i)$ and consider the worst α_1,α_2 possible (depending on the current component u and initial state S_{τ_i}). □

This is where the proof significatively deviates from the proof of Lemma III.27.

(**STEP 4**) *Let $\mathscr{Z}_0^u$ a component that is not communicating. There exists $\beta > 0$ such that*

$$\forall \delta > 0, \quad \mathbf{P}\left(\sum_{i \in \mathscr{I}^u} \left(\tau_{j(i+1)-1} - \tau_{j(i)} + 1 \right) \geq \beta\left(1 + \beta^3 + \beta^2 \log\left(\tfrac{1}{\delta}\right)\right) \right) \leq \delta. \tag{III.68}$$

Proof. Let $\mathscr{Z}_0^u$ a component that is not communicating. For simplicity, denote $T_i^u := \sum_{i' \in \mathscr{I}^u \cap [i]} (\tau_{j(i'+1)-1} - \tau_{j(i')} + 1)$. Consider the reward function $f(s,a) := \mathbf{1}(s \notin \mathscr{S}(\mathscr{Z}_0^u))$ and let g^f, h^f the associated gain and bias functions obtained by iterating the uniform policy on $\mathscr{Z}_0^u$, extended to the uniform policy outside of $\mathscr{S}(\mathscr{Z}_0^u)$. Observe that $g^f = e$. Let $\beta := \mathrm{sp}(h^f)$.

By design of panic times, for $t \in \{\tau_{j(i)}, \dots, \tau_{j(i+1)-1}\}$ we cannot have $f(Z_t) = 1$ if $i \in \mathscr{I}^u$. So, for all $i \geq 0$,

$$\begin{aligned}
0 &= \sum_{i' \in \mathscr{I}^u \cap [i]} \sum_{t=\tau_{j(i')}}^{\tau_{j(i'+1)-1}} f(Z_t) = \sum_{i' \in \mathscr{I}^u \cap [i]} \sum_{t=\tau_{j(i')}}^{\tau_{j(i'+1)-1}} \left(g^f(S_t) + \left(e_{S_t} - p(Z_t)\right) h^f \right) \\
&\geq T_i^u - \beta + \sum_{i' \in \mathscr{I}^u \cap [i]} \sum_{t=\tau_{j(i')}}^{\tau_{j(i'+1)-1}} \left(e_{S_{t+1}} - p(Z_t)\right) h^f \\
&\overset{(\dagger)}{\geq} T_i^u - \beta\left(1 + \sqrt{T_i^u \log\left(\tfrac{1+T_i^u}{\delta}\right)}\right)
\end{aligned}$$

where the last line (†) follows by a time-uniform Azuma-Hoeffding inequality (Lemma I.22) and holds with probability $1-\delta$. Making i go to infinity, by monotonicity, we see that $T^u := \sum_{i \in \mathscr{I}^u} (\tau_{j(i+1)-1} - \tau_{j(i)} + 1)$ satisfies:

$$\forall \delta > 0, \quad \mathbf{P}\left(T^u \leq \beta\left(1 + \sqrt{T^u \log\left(\tfrac{1+T^u}{\delta}\right)}\right)\right) \leq \delta \tag{III.69}$$

Conclude with straight forward algebra. □

We can now conclude. Recall that (τ_i) is the stopping-time enumeration of $\mathscr{T}^{\pm}(\mathfrak{Z}_0)$ and that $j(i)$ is the i-th initial stopping time in the sequence. Combining (**STEP 2**) and (**STEP 3**), we have:

$$\forall z \in \bigcup \mathfrak{Z}_0, \quad \mathbf{P}\left(N_{\tau_i}(z) < \alpha\sqrt{i}\right) = \mathrm{o}\left(i^{-2}\right) \tag{III.70}$$

for some $\alpha > 0$. Let $u \in \{1, \dots, m\}$ such that $\mathscr{Z}_0^m$ is not communicating. We have:

$$\begin{aligned}
\mathbf{E}\left[\sum_{t=0}^{\infty} \mathbf{1}(t \in \mathscr{T}^{\pm}(\mathfrak{Z}_0))\right] &= \mathbf{E}\left[\sum_{i=0}^{\infty} \mathbf{1}(\tau_i < \infty)\right] \\
&\overset{(\dagger)}{=} \mathbf{E}\left[\sum_{i=0}^{\infty} \mathbf{1}\left(\tau_i < \infty, \left(\forall z \in \bigcup \mathfrak{Z}_0, N_{\tau_i}(z) \geq \alpha\sqrt{i}\right)\right)\right] + \mathrm{O}(1) \\
&\leq \mathbf{E}\left[\sum_{i=0}^{\infty} \mathbf{1}\left(T^u \geq \alpha\sqrt{i}\right)\right] + \mathrm{O}(1) \\
&\overset{(\ddagger)}{\leq} \sum_{i=0}^{\infty} \exp\left(-\tfrac{\alpha}{\beta^2}\sqrt{i} + \beta(1+\beta^3)\right) + \mathrm{O}(1) < \infty
\end{aligned}$$

where (†) follows from (III.70) and (‡) from (III.68). This concludes the proof. ■

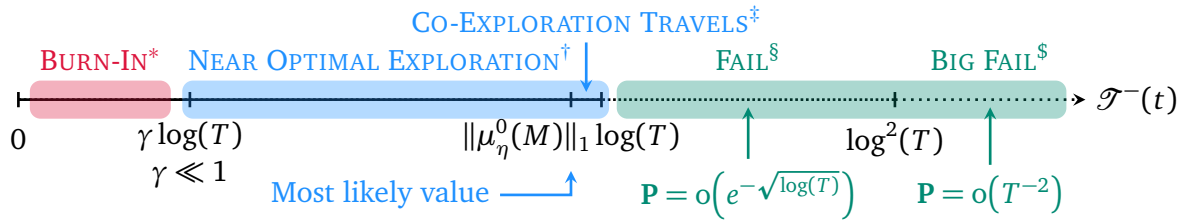


Figure 10.C.1: How the algorithm explores as $\mathscr{T}^-(t)$ grows.

10.C.4 Proof of Lemma III.26: Visits due to exploration

We start by providing a high level view of the analysis. We fix a precision parameter $\delta_0 > 0$ and some discard coefficient $\gamma > 0$.

Standard regime. We start by discarding what happens prior to time $\gamma \log(T)$, during the Burn-In period. Afterwards, right starting from time $\gamma \log(T)$, we invoke a uniform exploration argument (Lemma III.29) that all pairs have been visited at least $\alpha\gamma \log(T)$ times with reasonable probability. This amount of visits is enough to claim that the empirical data is concentrated enough so that (1) the near optimal pairs are correct with $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) = \mathscr{Z}^{**}(M)$, (2) we went beyond the possibility of other panic times with $\hat{M}_t \sim M$, (3) the exploration measures and policies are nearly $\mu_\eta^0(M)$ and (4) the proxy lower bound is nearly correct with $K_\eta^{\epsilon(t)}(\hat{M}_t) = K_\eta^0(M) \pm \delta_0$ (Lemma III.30). This event is referred to as the *trigger effect* and holds with reasonably high probability $1 - o(\exp(-\sqrt{\log(T)}))$. The algorithm therefore enters the Near Optimal Exploration period, during which exploration is (3) nearly optimal and (4) the exploration GLR test will nearly match the true regret lower bound. Once the time horizon $\mathscr{T}^-(t) \approx K_\eta^0(M) \log(T)$ is crossed, the exploration GLR test don't provoke exploration phases anymore and exploration is only triggered by co-exploration that starts exploration phases to switch of nearly optimal component. This is the Co-Exploration Travels phase. Thanks to the prior trigger effect (Lemma III.30), the co-exploration structure given by the nearly optimal pairs $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M})$ is invariant and equal to $\mathscr{Z}^{**}(M)$ during this phase and is shown to account for about $O(\log\log(T))$ exploration times, overall negligible. Therefore, the more likely values for $\mathscr{T}^-(t)$ are in the neighborhood of $K_\eta^0(M) \log(T)$.

Failures and the $\log^2(T)$ exploration barrier. If the number of exploration times goes beyond the threshold value $K_\eta^0 \log(T) + O(\log\log(T))$, it means that either uniform exploration or the trigger effect failed. We enter the Fail region. At horizon $O(\log^2(T))$, thanks to uniform exploration again (Lemma III.29), every pair is supposed to enter the skeleton, hence will be truncated during GLR tests. By design of the algorithm, if all pairs are truncated, then GLR tests will all pass and the algorithm will only exploit. This is referred to as the $\log^2(T)$-*exploration barrier* and is shown to hold with very high probability $1 - o(T^{-2})$. Because $\exp(-\sqrt{\log(T)}) \log^2(T) = o(1)$, this region is negligible in the expectation of $\mathscr{T}^-(T)$.

Total failures. In case exploration completely fails to be uniform and beyond the $\log^2(T)$-barrier, we enter the Big Fail region that has so little probability to happen that it can be straightforwardly neglected.

10.C.4.1 Main proof

Pick $z \notin \mathscr{Z}^{**}(M)$. Its *exploration visit count* at time t is given by

$$N_t^-(z) := \sum_{i=0}^{t-1} \mathbf{1}(Z_i = z, i \in \mathscr{T}^-) \tag{III.71}$$

In the sequel, we write $\mu_\eta^* := \|\mu_\eta^0(M)\|_1^{-1}\mu_\eta^0(M)$ the normalized optimal η-uniformized exploration measure of M (Theorem III.13) and π_η^* the associated optimal exploration policy, given by $\pi_\eta^*(a|s) \propto \mu_\eta^*(s,a)$. Introduce the following events:

$$\mathscr{E}_t := \left(|\mathscr{T}^-(t)| \geq \gamma \log(T) \Rightarrow \begin{pmatrix} \hat{M}_t \sim M \text{ and } \|\mu_\eta^{\epsilon(t)}(\hat{M}_t) - \mu_\eta^0(M)\|_\infty < \delta_0 \min(\mu_\eta^0(M)) \\ \text{and } \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) = \mathscr{Z}^{**}(M) \text{ and } |K_\eta^{\epsilon(t)}(\hat{M}_t) - K_\eta^0(M)| < \delta_0 \end{pmatrix} \right) \tag{III.72}$$

$$\mathscr{E}'_t := \left(\inf \left\{ \sum_{z \in \mathscr{Z}} N_t(z) \mathrm{KL}_z(\hat{M}_t || M^\dagger) : \begin{matrix} M^\dagger \in \mathrm{Alt}^{\epsilon(t)}(\hat{M}_t) \text{ and} \\ M^\dagger = \hat{M}_t \text{ on } \mathscr{Z}_t \cup \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) \end{matrix} \right\} \geq (1+\delta(t)) \log(t) \right) \tag{III.73}$$

We write $\mathscr{E} := \bigcap_{t=0}^{T-1} \mathscr{E}_t$ and $\mathscr{E}' := \bigcap_{t=0}^{T-1} \mathscr{E}'_t$. The first family of events $(\mathscr{E}_t)$ correspond to the motivated *trigger effect* and are used to manage during the Near Optimal Exploration period. The second family $(\mathscr{E}'_t)$ accounts tracks moments in time when the exploration GLR test is passed, when the algorithm move on to the co-exploration test, and will be of use to analyze the Co-Exploration Travels period.

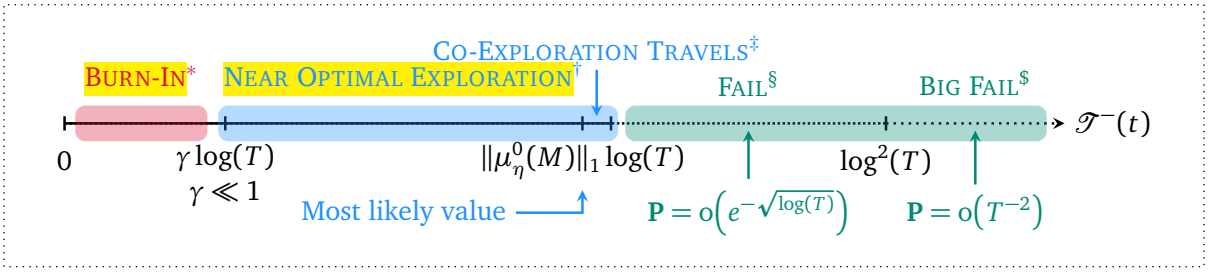
(**STEP 0**) *For $\mu \in (\mathbf{R}_+^*)^{\mathscr{Z}}$, denote π_μ the policy given by $\pi_\mu(a|s) = (\sum_{a' \in \mathscr{A}(s)} \mu(s,a'))^{-1}\mu(s,a)$. Let $\delta \in (0, \frac{1}{2})$. For $\mu, \mu' \in (\mathbf{R}_+^*)^{\mathscr{Z}}$, if $\|\mu' - \mu\|_\infty \leq \delta \min(\mu)$, then*

(1) for all $z \in \mathscr{Z}$, $(1-\delta)\mu(z) \leq \mu'(z) \leq (1-\delta)\mu(z)$ and $(1-\delta)\mu'(z) \leq \mu(z) \leq (1+2\delta)\mu'(z)$;
(2) for all $s \in \mathscr{S}$ and $a \in \mathscr{A}(s)$, $(1-2\delta)\pi_\mu(a|s) \leq \pi_{\mu'}(a|s) \leq (1-2\delta)^2\pi_\mu(a|s)$.

Proof. This is straight forward algebra. □

(**STEP 1**) *Let $z \notin \mathscr{Z}^{**}(M)$. There exists $C > 0$ independent of δ_0, γ, Tw such that, for all $\delta > 0$,*

$$\mathbf{P}\left(\exists t < T, \left| N_t^-(z) - |\mathscr{T}^-(t)| \mu_\eta^*(z) \right| \leq C \begin{pmatrix} \sqrt{|\mathscr{T}^-(t)| \log\left(\frac{2(1+|\mathscr{T}^-(t)|)}{\delta}\right)} + \mathbf{1}(\mathscr{E}^c)|\mathscr{T}^-(t)| \\ \gamma \log(T) + \delta_0 |\mathscr{T}^-(t)| + 2 \end{pmatrix} \right) \leq \delta. \tag{III.74}$$



Proof. We decompose the number of exploration visits as follows. Consider the reward function $f(s',a') := \mathbf{1}((s',a') = z)$ and let g_η^z, h_η^z the gain and bias functions associated to π_η^* with reward

function f. Remark that $g_\eta^z = \mu_\eta^*(z)e$. Further denote $\Delta_\eta^z(s',a') := g_\eta^z(s') + h_\eta^z(s') - f(s',a') - p(s',a')h_\eta^z$ the associated gap function. Let (τ_i) the stopping time enumeration of $\mathscr{T}^-$. We have:

$$N_t^-(z) = \sum_{i=1}^{|\mathscr{T}^-(t)|} \mathbf{1}(Z_{\tau_i} = z) \overset{(\dagger)}{=} |\mathscr{T}^-(t)|\mu_\eta^*(z) + \underbrace{\sum_{i=1}^{|\mathscr{T}^-(t)|} \Big(e_{S_{\tau_i}} - p(Z_{\tau_i})\Big)h_\eta^z}_{\mathrm{A}_t} - \underbrace{\sum_{i=1}^{|\mathscr{T}^-(t)|} \Delta_\eta^z(Z_{\tau_i})}_{\mathrm{B}_t}.$$

where $(\dagger)$ follows by the Poisson equation.

We start by dealing with the error term A_t. We expand this term as:

$$\mathrm{A}_t = \underbrace{\sum_{i=1}^{|\mathscr{T}^-(t)|} \Big(h_\eta^z(S_{\tau_i}) - h_\eta^z(S_{\tau_{i+1}})\Big)}_{\mathrm{A}_t^1} + \underbrace{\sum_{i=1}^{|\mathscr{T}^-(t)|} \Big(h_\eta^z(S_{\tau_{i+1}}) - h_\eta^z(S_{\tau_i+1})\Big)}_{\mathrm{A}_t^2} + \underbrace{\sum_{i=1}^{|\mathscr{T}^-(t)|} \Big(e_{S_{\tau_i+1}} - p(Z_{\tau_i})\Big)h_\eta^z}_{\mathrm{A}_t^3} \tag{III.75}$$

The term A_t^1 is a telescopic sum which is bounded by $\mathrm{sp}(h_\eta^z)$ in absolute value. The term A_t^2 is bounded by the number of panic times; Indeed, during $\{\tau_i + 1, \dots, \tau_{i+1} - 1\}$ the algorithm is either co-exploring or exploitating and both are done until regeneration, unless the algorithm has panicked. Because the number of panic times is at most $|\mathscr{Z}|$, we get $|\mathscr{T}_t^2| \le \mathrm{sp}(h_\eta^z)|\mathscr{Z}|$ a.s. The term A_t^3 is a martingale and by a time-uniform Azuma-Hoeffding inequality (Lemma I.22), we find:

$$|\mathrm{A}_t^3| \le \mathrm{sp}(h_\eta^z)\sqrt{|\mathscr{T}^-(t)|\log\Big(\tfrac{\sqrt{1+|\mathscr{T}^-(t)|}}{\delta}\Big)} \le \mathrm{sp}(h_\eta^z)\sqrt{|\mathscr{T}^-(t)|\log\Big(\tfrac{1+|\mathscr{T}^-(t)|}{\delta}\Big)} \tag{III.76}$$

with probability $1-\delta$, uniformly for $t \ge 0$. Combining (III.76) with the previous remarks and injecting it in (III.75), we obtain:

$$\mathbf{P}\Big(\exists t \ge 0, |\mathrm{A}_t| \ge \mathrm{sp}(h_\eta^z)\Big(\sqrt{|\mathscr{T}^-(t)|\log\Big(\tfrac{1+|\mathscr{T}^-(t)|}{\delta}\Big)} + 1 + |\mathscr{Z}|\Big)\Big) \le \delta. \tag{III.77}$$

We continue by dealing with the error term B_t. We expand it as:

$$\mathrm{B}_t = \sum_{i=1}^{\mathscr{T}^-(t)} \Big\langle e_{Z_{\tau_i}}, \Delta_\eta^z \Big\rangle = \underbrace{\sum_{i=1}^{\mathscr{T}^-(t)} \Big\langle e_{Z_{\tau_i}} - \pi_{\tau_i}^-(S_{\tau_i}), \Delta_\eta^z \Big\rangle}_{\mathrm{B}_t^1} + \underbrace{\sum_{i=1}^{\mathscr{T}^-(t)} \Big\langle \pi_{\tau_i}^-(S_{\tau_i}), \Delta_\eta^z \Big\rangle}_{\mathrm{B}_t^1} \tag{III.78}$$

We recognize a martingale like always, here B_t^1 that we bound using a time-uniform Azuma-Hoeffding inequality for a change (Lemma I.22). With probability $1-\delta$, we have:

$$|\mathrm{B}_t^1| \le \|\Delta_\eta^z\|_\infty \sqrt{|\mathscr{T}^-(t)|\log\Big(\tfrac{\sqrt{1+|\mathscr{T}^-(t)|}}{\delta}\Big)} \le \|\Delta_\eta^z\|_\infty \sqrt{|\mathscr{T}^-(t)|\log\Big(\tfrac{1+|\mathscr{T}^-(t)|}{\delta}\Big)}. \tag{III.79}$$

The other term B_t^2 is decomposed through the trigger effect event $\mathscr{E}$, introduced in (III.72). We write:

$$|\mathrm{B}_t^2| = \mathbf{1}(\mathscr{E})\left|\sum_{i=1}^{\lceil \gamma \log(T)\rceil \wedge |\mathscr{T}^-(t)|} \Big\langle \pi_{\tau_i}^-(S_{\tau_i}), \Delta_\eta^z \Big\rangle + \sum_{i=\lceil \gamma \log(T)\rceil}^{|\mathscr{T}^-(t)|} \Big\langle \pi_{\tau_i}^-(S_{\tau_i}), \Delta_\eta^z \Big\rangle\right| + \mathbf{1}(\mathscr{E}^c)\left|\sum_{i=1}^{|\mathscr{T}^-(t)|} \Big\langle \pi_{\tau_i}^-(S_{\tau_i}), \Delta_\eta^z \Big\rangle\right|$$

$$\le \mathbf{1}(\mathscr{E})\left(\|\Delta_\eta^z\|_\infty \gamma \log(T) + \left|\sum_{i=\lceil \gamma \log(T)\rceil}^{|\mathscr{T}^-(t)|} \Big\langle \pi_{\tau_i}^-(S_{\tau_i}), \Delta_\eta^z \Big\rangle\right|\right) + \mathbf{1}(\mathscr{E}^c)\|\Delta_\eta^z\|_\infty |\mathscr{T}^-(t)|$$

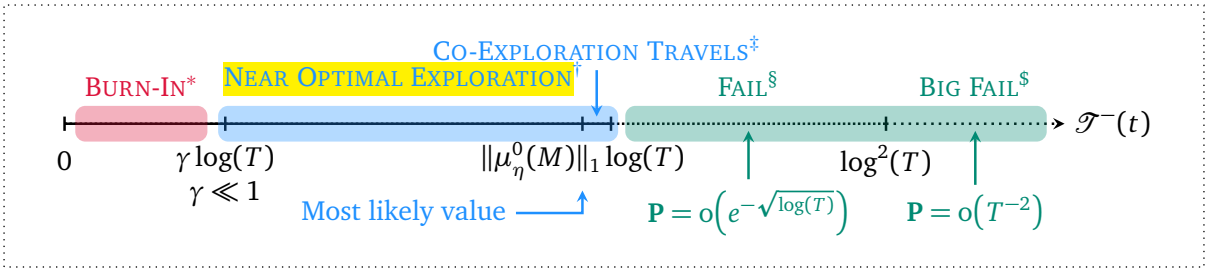
$$\overset{(\dagger)}{\leq} \mathbf{1}(\mathscr{E})\left|\sum_{i=\lceil\gamma\log(T)\rceil}^{|\mathscr{T}^-(t)|}\left\langle \pi^-_{\tau_i}(S_{\tau_i})-\pi^*_\eta(S_{\tau_i}),\Delta^z_\eta\right\rangle\right| + \|\Delta^z_\eta\|_\infty\big(\gamma\log(T)+\mathbf{1}(\mathscr{E}^c)|\mathscr{T}^-(t)|\big)$$
$$\overset{(\ddagger)}{\leq} \|\Delta^z_\eta\|_\infty\big(|\mathscr{T}^-(t)|(\mathbf{1}(\mathscr{E})\delta_0+\mathbf{1}(\mathscr{E}^c))+\gamma\log(T)\big). \tag{III.80}$$

In the above, (†) uses that $\langle\pi^*_\eta(S_{\tau_i}),\Delta^z_\eta\rangle = 0$ and (‡) unfolds the definition of $\mathscr{E}$, on which $\|\pi^-_{\tau_i}(s)-\pi^*_\eta(s)\|_1 \leq \delta_0$ for $i \geq \gamma\log(T)$ and $s\in\mathscr{S}$. Injecting (III.79) and (III.80) into (III.78) and combining with (III.77), we obtain the desired result by setting $C := \mathrm{sp}(h^z_\eta)\vee\|\Delta^z_\eta\|_\infty$. □

(**STEP 2**) *Assume that* $\ell \geq \delta_0^{-1}(\gamma\log(T)+2)+\delta_0^{-4}$. *Then, for all* $z\in\mathscr{Z}$,

$$\mathbf{P}\Big(\exists t\geq 0:\ N^-_t(z) < \ell\mu^*_\eta(z)\Big(1-\tfrac{4C\delta_0}{\mu^*_\eta(z)}\Big) \text{ and } |\mathscr{T}^-(t)|\geq\ell \text{ and } \mathscr{E}\Big) = \exp\big(-\delta_0^2\ell+\log(1+\ell)\big),$$
$$\mathbf{P}\Big(\exists t\geq 0:\ N^-_t(z) > \ell\mu^*_\eta(z)\Big(1+\tfrac{4C\delta_0}{\mu^*_\eta(z)}\Big) \text{ and } |\mathscr{T}^-(t)|\geq\ell \text{ and } \mathscr{E}\Big) = \exp\big(-\delta_0^2\ell+\log(1+\ell)\big) \tag{III.81}$$

where $C>0$ *is the constant given by* (***STEP 1***).



Proof. This is mostly about rewriting Equation (III.74) of (**STEP 1**). Fix $\delta>0$. On $\mathscr{E}$ and provided that $|\mathscr{T}^-(t)|\geq\ell$, by (III.74) is holds with probability $1-\delta$ that:

$$\begin{aligned}
N^-_t(z) &\geq |\mathscr{T}^-(t)|\mu^*_\eta(z) - C\left(\sqrt{|\mathscr{T}^-(t)|\log\left(\tfrac{1+|\mathscr{T}^-(t)|}{\delta}\right)}+\gamma\log(T)+\delta_0|\mathscr{T}^-(t)|+2\right)\\
&\geq |\mathscr{T}^-(t)|\mu^*_\eta(z)\left(1-\tfrac{C\delta_0}{\mu^*_\eta(z)}-\tfrac{C\gamma\log(T)}{\mu^*_\eta(z)|\mathscr{T}^-(t)|}-\tfrac{C}{\mu^*_\eta(z)}\sqrt{\tfrac{\log((1+|\mathscr{T}^-(t)|)/\delta)}{|\mathscr{T}^-(t)|}}\right)\\
&\geq |\mathscr{T}^-(t)|\mu^*_\eta(z)\left(1-\tfrac{C\delta_0}{\mu^*_\eta(z)}-\tfrac{C(\gamma\log(T)+2)}{\mu^*_\eta(z)|\mathscr{T}^-(t)|}-\tfrac{C}{\mu^*_\eta(z)}\sqrt{\tfrac{\log(1/\delta)}{|\mathscr{T}^-(t)|}}-\tfrac{C}{\mu^*_\eta(z)}\sqrt{\tfrac{\log(1+|\mathscr{T}^-(t)|)}{|\mathscr{T}^-(t)|}}\right)\\
&\overset{(\dagger)}{\geq} \ell\mu^*_\eta(z)\left(1-\tfrac{C\delta_0}{\mu^*_\eta(z)}-\tfrac{C(\gamma\log(T)+2)}{\mu^*_\eta(z)\ell}-\tfrac{C}{\mu^*_\eta(z)}\sqrt{\tfrac{\log(1/\delta)}{\ell}}-\tfrac{C}{\mu^*_\eta(z)}\sqrt{\tfrac{\log(1+\ell)}{\ell}}\right)\\
&\overset{(\ddagger)}{\geq} \ell\mu^*_\eta(z)\left(1-\tfrac{4C\delta_0}{\mu^*_\eta(z)}\right)
\end{aligned}$$

where (†) holds by monotonicity in ℓ and (†) holds provided that:

$$\max\left\{\frac{C(\gamma\log(T)+2)}{\mu^*_\eta(z)\ell},\frac{C}{\mu^*_\eta(z)}\sqrt{\frac{\log(1/\delta)}{\ell}},\frac{C}{\mu^*_\eta(z)}\sqrt{\frac{\log(1+\ell)}{\ell}}\right\}\leq\frac{C\delta_0}{\mu^*_\eta(z)}.$$

The above holds in particular when:

$$\ell\geq\max\left\{\frac{\gamma\log(T)+2}{\delta_0},\left(\frac{1}{\delta_0}\right)^4\right\}\quad\text{and}\quad\delta\geq\exp\big(-\delta_0^2\ell+\log(1+\ell)\big).$$

Pick $\delta := \exp(-\delta_0^2\ell+\log(1+\ell))$ to conclude the proof. The lower bound is obtained similarly. □

(**STEP 3**) *Introduce the exploration threshold and the GLR exploration function:*

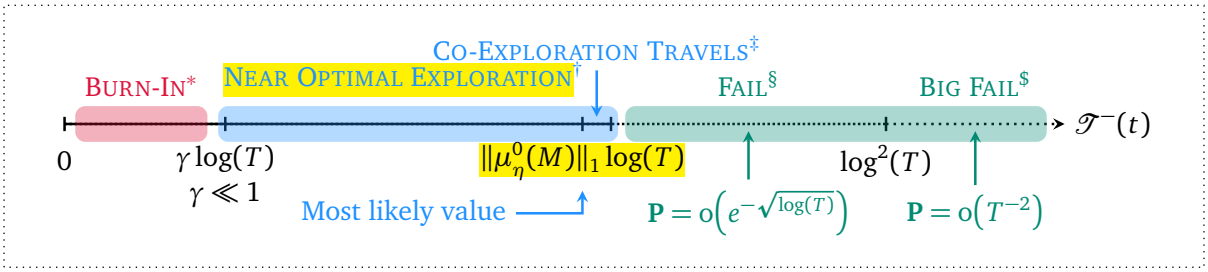
$$\ell_0 := \max\left\{\frac{(1+\delta(\gamma\log(T)))\|\mu_\eta^0(M)\|_1}{1-\delta_0\left(1+\frac{4C}{\mu_\eta^*(z)}\right)}\cdot\log(T), \frac{\gamma\log(T)+2}{\delta_0}, \left(\frac{1}{\delta_0}\right)^4\right\} \tag{III.82}$$

$$\mathrm{GLR}^-(t) := \inf\left\{\sum_{z\in\mathscr{Z}} N_t(z)\mathrm{KL}_z(\hat{M}_t||M^\dagger) : M^\dagger \in \mathrm{Alt}^{\epsilon(t)}(\hat{M}_t) \text{ and } \hat{M}_t = M^\dagger \text{ on } \mathscr{Z}_t \cup \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)\right\} \tag{III.83}$$

Given ℓ_0 and $\mathrm{GLR}^-(t)$ as defined in (III.82) and (III.83) respectively, for all $t \geq 0$, we have:

$$\left(N_t^-(z) \geq \ell_0\mu_\eta^*(z)\left(1-\tfrac{4C\delta_0}{\mu_\eta^*(z)}\right) \text{ and } |\mathscr{T}^-(t)| \geq \ell_0 \text{ and } \mathscr{E}\right) \subseteq .\left(\mathrm{GLR}^-(t) \geq (1+\delta(t))\log(T) \text{ and } |\mathscr{T}^-(t)| \geq \ell_0 \text{ and } \mathscr{E}\right) \tag{III.84}$$

In other words, passed the exploration threshold ℓ_0, the GLR tests "$\mathrm{GLR}^-(t) \geq (1+\delta(t))\log(t)$?" deciding exploration will all be passed and none will provoke an exploration phase.



Proof. In light of (III.81), assume that we are on the event:

$$\mathscr{E} \cap \left(\forall z \in \mathscr{Z} : N_t^-(z) > \ell\mu_\eta^*(z)\left(1-\tfrac{4C\delta_0}{\mu_\eta^*(z)}\right)\right). \tag{III.85}$$

Fix $t \geq 0$ and let $M^\dagger \in \mathrm{Alt}^{\epsilon(t)}(\hat{M}_t)$ such that $\hat{M}_t = M^\dagger$ on the extended skeleton $\mathscr{Z}_t \cup \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$. We have:

$$\begin{aligned}
\sum_{z\in\mathscr{Z}} N_t(z)\mathrm{KL}_z(\hat{M}_t||M^\dagger) &\geq \sum_{z\in\mathscr{Z}} \ell\mu_\eta^*(z)\left(1-\tfrac{4C\delta_0}{\mu_\eta^*(z)}\right)\mathrm{KL}_z(\hat{M}_t||M^\dagger) \\
&= \ell\left(1-\tfrac{4C\delta_0}{\mu_\eta^*(z)}\right)\sum_{z\in\mathscr{Z}} \tfrac{\mu_\eta^0(z,M)}{\|\mu_\eta^0(M)\|_1}\mathrm{KL}_z(\hat{M}_t||M^\dagger) \\
&\overset{(\dagger)}{\geq} \ell\left(1-\tfrac{4C\delta_0}{\mu_\eta^*(z)}\right)\sum_{z\in\mathscr{Z}} \tfrac{(1-\delta_0)\mu_\eta^{\epsilon(t)}(z,\hat{M}_t)}{\|\mu_\eta^0(M)\|_1}\mathrm{KL}_z(\hat{M}_t||M^\dagger) \\
&\overset{(\ddagger)}{\geq} \ell\cdot\left(1-\tfrac{4C\delta_0}{\mu_\eta^*(z)}\right)(1-\delta_0)\|\mu_\eta^0(M)\|_1^{-1} \geq \ell\left(1-\delta_0\left(1+\tfrac{4C}{\mu_\eta^*(z)}\right)\right)\|\mu_\eta^0(M)\|_1^{-1}.
\end{aligned} \tag{III.86}$$

In (†), relate $\mu_\eta^0(z,M)$ to $\mu_\eta^{\epsilon(t)}(z,\hat{M}_t)$ by using the definition of $\mathscr{E}$ and (**STEP 0**). In (‡), we use the fact $\mu_\eta^{\epsilon(t)}(\hat{M}_t)$ is an exploration measure and that $M^\dagger \in \mathrm{Cnf}^{\epsilon(t)}(\hat{M}_t)$ so that

$$\sum_{z\in\mathscr{Z}} \mu_\eta^{\epsilon(t)}(z,\hat{M}_t)\mathrm{KL}_z(\hat{M}_t||M^\dagger) \geq 1.$$

Continuing (III.86), we have:

$$\ell\left(1-\delta_0\left(1+\tfrac{4C}{\mu_\eta^*(z)}\right)\right)\|\mu_\eta^0(M)\|_1^{-1} \geq (1+\delta(t))\log(T) \iff \ell \geq \frac{(1+\delta(t))\|\mu_\eta^0(M)\|_1}{1-\delta_0\left(1+\frac{4C}{\mu_\eta^*(z)}\right)}\cdot\log(T).$$

In order to apply (III.81), we need to have $\ell \geq \delta_0^{-1}\gamma\log(T) \geq \gamma\log(T)$ hence we can lower bound $\delta(t)$ by $\delta(\gamma\log(T))$. We recover the definition of ℓ_0, see (III.82). We conclude accordingly. □

(**STEP 4**) *Introduce the initial co-exploration GLR function:*

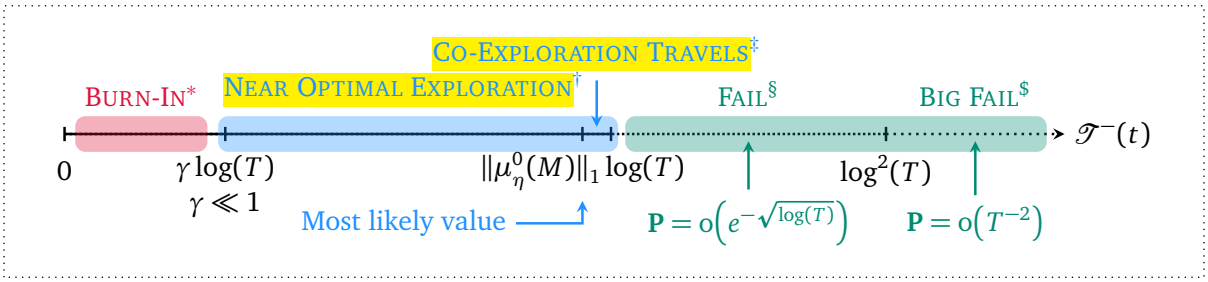
$$\mathrm{GLR}^{\pm}(t) := \inf\left\{\sum_{z\in\mathscr{Z}} N_t(z)\mathrm{KL}_z(\hat{M}_t||M^\dagger) : M^\dagger \in \mathrm{Alt}^{\epsilon(t)}(\hat{M}_t) \text{ and } \hat{M}_t = M^\dagger \text{ on } \mathscr{Z}_t\right\} \tag{III.87}$$

We say that t is an initial co-exploration travel time ($t \in \mathscr{T}^-_{\mathrm{co}}$) if it is an exploration time triggered by the co-exploration rule forcing exploration by trying to reach a sub-visited component of $\mathscr{Z}^{\epsilon(t)}_{**}(\hat{M}_t)$. More formally, $t \in \mathscr{T}^-_{\mathrm{co}}$ if (1) $t \in \mathscr{T}^-$ with (2) $Z_t \in \mathscr{Z}^{\epsilon(t)}_{**}(\hat{M}_t)$, (3) $\mathrm{GLR}^-(t) > (1+\delta(t))\log(t)$ and (4) $\mathrm{GLR}^{\pm}(t) \leq (1+\delta(t))\log(t)$. Let

$$\mathscr{I}_{\mathrm{co}}(T) := \left|\left\{t < T : t \in \mathscr{T}^-_{\mathrm{co}} \text{ and } |\mathscr{T}^-(t)| \geq \ell_0\right\}\right| \tag{III.88}$$

the number of initial co-exploration travel times prior to time T and happening after the exploration threshold $|\mathscr{T}^-(t)| \geq \ell_0$. There exist constants $\alpha, \beta > 0$ independent of δ_0, γ, T such that:

$$\mathbf{P}\Big(|\mathscr{T}^-(T)| \geq \ell_0 + \beta\Big(|\mathscr{I}_{\mathrm{co}}(T)| + \sqrt{\log(T)}\Big), \mathscr{E}\Big) = \mathrm{o}\Big(\exp\Big(-\alpha\sqrt{\log(T)}\Big)\Big) \tag{III.89}$$



Proof. Consider the event:

$$\mathscr{E}^* := \mathscr{E} \cap \left(\forall z \in \mathscr{Z}, |\mathscr{T}^-(t)| \geq \ell_0 \Rightarrow N^-_t(z) \geq \ell_0\mu^*_\eta(z)\left(1 - \tfrac{4C\delta_0}{\mu^*_\eta(z)}\right)\right). \tag{III.90}$$

Consider the stopping time enumeration of $\mathscr{T}^-$ starting from $\mathscr{T}^-(t) = \lceil \ell_0 \rceil$. More formally:

$$\tau_1 := \inf\{t \in \mathscr{T}^- : |\mathscr{T}^-(t)| \geq \ell_0\} \quad \text{and} \quad \tau_{j+1} := \inf\{t \in \mathscr{T}^- : t > \tau_j\}.$$

The sequence (τ_j) is partitionned into segments on which $\tau_{j+1} = \tau_j + 1$. Let $j(i)$ the starting index of the i-th segment of (τ_j). Since $\tau_j \geq \tau_1$ for all j, we have $\mathscr{T}^-(\tau_j) \geq \ell_0$. So, on $\mathscr{E}^*$, it follows from (III.84) that every segment $\{\tau_{j(i)}, \ldots, \tau_{j(i+1)-1}\}$ must be initiated by a initial co-exploration travel time for $i \geq 2$, i.e., $\tau_{j(i)} \in \mathscr{T}^-_{\mathrm{co}}$ for all $i \geq 2$. Hence:

(∗) On $\mathscr{E}^*$, the number of segments $\{\tau_{j(i)}, \ldots, \tau_{j(i+1)-1}\}$ up to T is at most $1 + |\mathscr{I}_{\mathrm{co}}(T)|$.

Meanwhile, every segment $\{\tau_{j(i)}, \ldots, \tau_{j(i+1)} - 1\}$ ends when $S_{\tau_{j(i+1)}}$ hits the states spawned by the current nearly optimal pairs $\mathscr{Z}^{\epsilon(t)}_{**}(\hat{M}_t)$, when the algorithm switches back to co-exploration or exploitation. Remarkably, by design of $\mathscr{E}$ and since $\mathscr{T}^-(t) \geq \ell_0 \geq \gamma\log(T)$, the nearly optimal pairs have converged to $\mathscr{Z}^{**}(M)$. From this, observe that

(∗∗) On $\mathscr{E}^*$, the time segment $\{\tau_{j(i)}, \ldots, \tau_{j(i+1)-1}\}$ can be, in the worst case, as long as the time required to reach the least visited component of $\mathscr{Z}^{**}(M)$ because as soon as it is reached, the algorithm *will* exploit or co-explore the component.

Let $\mathscr{Z}_{**}^1(M),\dots,\mathscr{Z}_{**}^m(M)$ the components of $\mathscr{Z}_{**}(M)$. Since, on $\mathscr{E}$, $\hat{M}_t \sim M$ with $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) = \mathscr{Z}^{**}(M)$, the lattest are also the components of $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ as estimated by the algorithm. Let $m(t)$ the component S_t is on at time t, if any (if it exists, it has to be unique). The least visited component at time t is denoted $m^-(t) := \min\{c : \min\{N_t(z) : z \in \mathscr{Z}_{**}^c(M)\} = \min\{N_t(z) : z \in \mathscr{Z}^{**}(M)\}\}$.

Given $i \geq 1$, we introduce the reward function $f^i(z) := \mathbf{1}(z \in \mathscr{Z}_{**}^{m_-(\tau_i)}(M))$ marking the least visited component. Let g^i, h^i the gain and bias functions under this reward function obtained by iterating the optimal exploration policy π_η^*. Its gap function is $\Delta^i(s,a) := g^i(s) + h^i(s) - f^i(s,a) - p(s,a)h^i$. Let $\beta^i := \mathrm{sp}(h^i) \vee \|\Delta^i\|_\infty$ and denote $\beta < \infty$ the maximum value that β^i can take over all (finitely many) values that f^i can take. We further have $g^i = \alpha^i e$ for some $\alpha^i > 0$ and pick $\alpha > 0$ the minimal value that α^i can take. Invoking $(*)$, we have:

$$
\begin{aligned}
0 &\geq \mathbf{1}(\mathscr{E}^*) \sum_{t=\tau_{j(i)}}^{\tau_{j(i+1)-1}} f^i(Z_t) \\
&\overset{(\dagger)}{=} \mathbf{1}(\mathscr{E}^*) \sum_{t=\tau_{j(i)}}^{\tau_{j(i+1)-1}} \big(\alpha^i + \big(e_{S_t} - p(Z_t)\big)h^i - \Delta^i(Z_t)\big) \\
&\overset{(\ddagger)}{\geq} \mathbf{1}(\mathscr{E}^*)\big(\alpha\big(\tau_{j(i+1)-1} - \tau_{j(i)} + 1\big) - \beta\big) + \mathbf{1}(\mathscr{E}^*) \sum_{t=\tau_{j(i)}}^{\tau_{j(i+1)-1}} \big(e_{S_{t+1}} - p(Z_t)\big)h^i \\
&\qquad\qquad - \mathbf{1}(\mathscr{E}^*) \sum_{t=\tau_{j(i)}}^{\tau_{j(i+1)-1}} \left\langle e_{Z_t} - \pi_t^-(S_t), \Delta^i \right\rangle \\
&\qquad\qquad - \mathbf{1}(\mathscr{E}^*) \sum_{t=\tau_{j(i)}}^{\tau_{j(i+1)-1}} \left\langle \pi_t^-(S_t) - \pi_\eta^*(S_t), \Delta^i \right\rangle \\
&\overset{(\S)}{\geq} \mathbf{1}(\mathscr{E}^*)\left(\alpha T_i - \beta(\delta_0 T_i + 1) + \sum_{t=\tau_{j(i)}}^{\tau_{j(i+1)-1}} \big(e_{S_{t+1}} - p(Z_t)\big)h^i + \sum_{t=\tau_{j(i)}}^{\tau_{j(i+1)-1}} \left\langle e_{Z_t} - \pi_t^-(S_t), \Delta^i \right\rangle \right)
\end{aligned}
$$

where $(\dagger)$ uses the Poisson equation, $(\ddagger)$ is just rewriting and $(\S)$ invokes the definition of $\mathscr{E}$ to upper bound $\|\pi_t^-(S_t) - \pi_\eta^*(S_t)\|_1$ while introducing the shorthand $T_i := \tau_{j(i+1)-1} - \tau_{j(i)} + 1$. We sum over i and we have $1 + |\mathscr{I}_{\mathrm{co}}(T)|$ by $(**)$ on $\mathscr{E}^*$. We obtain:

$$
0 \geq \mathbf{1}(\mathscr{E}^*)\left((\alpha - \beta\delta_0)\sum_i T_i - \beta(1 + |\mathscr{I}_{\mathrm{co}}(T)|) + \sum_i \sum_{t=\tau_{j(i)}}^{\tau_{j(i+1)-1}} \Big(\big(e_{S_{t+1}} - p(Z_t)\big)h^i + \left\langle e_{Z_t} - \pi_t^-(S_t), \Delta^i \right\rangle\Big) \right). \tag{III.91}
$$

By a time-uniform Azuma-Hoeffding inequality, with probability $\delta > 0$, we have:

$$
\sum_i \sum_{t=\tau_{j(i)}}^{\tau_{j(i+1)-1}} \Big(\big(e_{S_{t+1}} - p(Z_t)\big)h^i + \left\langle e_{Z_t} - \pi_t^-(S_t), \Delta^i \right\rangle\Big) \geq -2\beta\sqrt{\sum_i T_i \log\Big(\tfrac{1+\sum_i T_i}{\delta}\Big)}. \tag{III.92}
$$

Observe that α, β are independent of the choice of γ, δ_0 and T hence we can assume that $\alpha - \beta\delta_0 > 0$ up to reducing δ_0. So, on $\mathscr{E}^*$ and with probability $1-\delta$, (III.91) is rewritten as an equation of the form:

$$
u \leq \lambda_1 + \sqrt{\lambda_2 u \log(1+u)} + \sqrt{\lambda_3 u}
$$

where $u = \sum_i T_i$, $\lambda_1 = (\alpha - \beta\delta_0)^{-1}\beta(1 + |\mathscr{I}_{\mathrm{co}}(T)|)$, $\lambda_2 = (\alpha - \beta\delta_0)^{-1} \cdot 4\beta^2$ and $\lambda_3 = (\alpha - \beta\delta_0)^{-1}\log\big(\frac{1}{\delta}\big)$. The right-hand side can be decoupled by solving the weaker pair of equations:

$$
u \leq \lambda_1 + 2\sqrt{\lambda_2 u \log(1+u)} \quad \text{and} \quad u \leq \lambda_1 + 2\sqrt{\lambda_3 u}.
$$

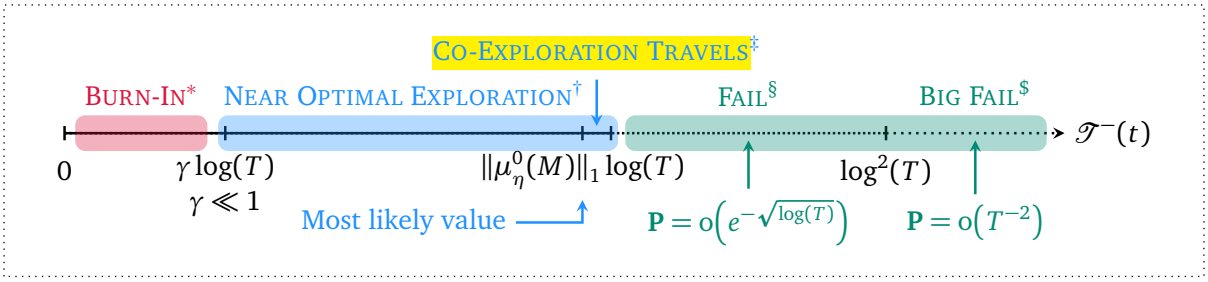
Using $\log(1+u) \le \sqrt{u}$, we find $u \le \max(2(\lambda_1+4\lambda_3), 2\lambda_1, 256\lambda_2^2)$. In other words, with probability $1-\delta$, we have:

$$\mathbf{1}(\mathscr{E}^*)|\mathscr{T}^-(T)| \le \ell_0 + \tfrac{2\beta}{\alpha-\beta\delta_0}(1+|\mathscr{I}_{\mathrm{co}}|(T)) + \tfrac{8}{\alpha-\beta\delta_0}\log\big(\tfrac{1}{\delta}\big) + 256\Big(\tfrac{4\beta^2}{\alpha-\beta\delta_0}\Big)^2. \tag{III.93}$$

Picking $\delta = \Theta\big(\exp\big(-\sqrt{\log(T)}\big)\big)$, we obtain the result. □

(**STEP 5**) *The number of initial co-exploration travel times prior to time T and happening after the exploration threshold $|\mathscr{T}^-(t)| \ge \ell_0$, see (III.88), satisfy:*

$$\mathbf{P}(|\mathscr{I}_{\mathrm{co}}(T)| > |\mathscr{S}|(1+2\log\log(T)), \mathscr{E}) \le \exp\big(-\delta_0^2\ell_0 + \log(1+\ell_0)\big). \tag{III.94}$$



Proof. We consider similar notations than in the proof of (**STEP 4**). We consider the event $\mathscr{E}^*$ introduce in (III.90) and the stopping time enumeration of $\mathscr{T}^-$ starting from $\mathscr{T}^-(t) = \lceil \ell_0 \rceil$. More formally:

$$\tau_1 := \inf\big\{t \in \mathscr{T}^- : |\mathscr{T}^-(t)| \ge \ell_0\big\} \quad \text{and} \quad \tau_{j+1} := \inf\big\{t \in \mathscr{T}^- : t > \tau_j\big\}.$$

Again, on $\mathscr{E}^*$, it follows from (III.84) that every segment $\{\tau_{j(i)}, \dots, \tau_{j(i+1)-1}\}$ must be initiated by a initial co-exploration travel time for $i \ge 2$, i.e., $\tau_{j(i)} \in \mathscr{T}_{\mathrm{co}}^-$ for all $i \ge 2$. Let $\mathscr{Z}_{**}^1(M), \dots, \mathscr{Z}_{**}^m(M)$ the components of $\mathscr{Z}_{**}(M)$. Since, on $\mathscr{E}$, $\hat{M}_t \sim M$ with $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) = \mathscr{Z}^{**}(M)$, the lattest are also the components of $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t)$ as estimated by the algorithm. Let $m(t)$ the component S_t is on at time t, if any (if it exists, it has to be unique). The least visited component at time t is denoted $m^-(t) := \min\{c : \min\{N_t(z) : z \in \mathscr{Z}_{**}^c(M)\} = \min\{N_t(z) : z \in \mathscr{Z}^{**}(M)\}\}$. Denote

$$L_{\tau_{j(i)},c} := \log\min\big\{N_{\tau_{j(i)}}(z) : z \in \mathscr{Z}_{**}^c\big\}, \quad c \in \{1,\dots,m\}. \tag{III.95}$$

Remark that, on $\mathscr{E}^*$, $L_{\tau_{j(i)},c}$ satisfies the following pair of equations:

$$\begin{cases} L_{\tau_{j(i)},m(\tau_{j(i)})} \le 2\min_c L_{\tau_{j(i)},c} \\ L_{\tau_{j(i+1)},m(\tau_{j(i)})} > 2\min_c L_{\tau_{j(i)},c} \end{cases} \qquad i \ge 2. \tag{III.96}$$

By induction, we deduce $L_{\tau_{j(i)},c} \ge 2^{\lfloor (i-1)/m \rfloor}$. Moreover, if $\tau_{j(i)} < T$, we have $L_{\tau_{j(i)},c} \le \log(T)$ for all $c = 1,\dots,m$. Last but not least, remember from (**STEP 4**) that on $\mathscr{E}^*$, the number of segments $\{\tau_{j(i)}, \dots, \tau_{j(i+1)-1}\}$ up to T is at most $1 + |\mathscr{I}_{\mathrm{co}}(T)|$. So, on $\mathscr{E}^*$.

$$|\mathscr{I}_{\mathrm{co}}(T)| \le m\Big(1 + \tfrac{\log\log(T)}{\log(2)}\Big) \le |\mathscr{S}|(1+2\log\log(T)).$$

This all happens on $\mathscr{E}^*$. Conclude with (**STEP 2**), see (III.81). □

We can finally conclude the proof. By (**STEP 4**), see (III.89), there exists an event $\mathscr{E}'$ with $\mathbf{P}(\mathscr{E}'^c \cap \mathscr{E}) = o(\exp(-\alpha\sqrt{\log(T)}))$ on which $|\mathscr{T}^-(T)| \le \ell_0 + \beta(|\mathscr{I}_{\mathrm{co}}(T)| + \sqrt{\log(T)})$. Furthermore, $\alpha, \beta > 0$ are independent of the choice of δ_0, γ and T. Up to intersecting events, on $\mathscr{E}' \cap \mathscr{E}$ we

further have the bound on $|\mathscr{I}_{\text{co}}(T)|$ provided by (**STEP 5**), see (III.94). Still up to intersecting events again, we can further assume that the upper-bound of (**STEP 2**), see (III.81) holds, while still guaranteeing $\mathbf{P}(\mathscr{E}'^c \cap \mathscr{E}) = o(\exp(-\alpha\sqrt{\log(T)}))$. We obtain the bound:

$$\mathbf{1}(\mathscr{E}' \cap \mathscr{E})N_T^-(z) \leq \ell_0 \mu_\eta^*(z)\Big(1 + \tfrac{4C\delta_0}{\mu_\eta^*(z)}\Big) + 2\beta|\mathscr{S}|(1+\log\log(T)) + \beta\sqrt{\log(T)}. \tag{III.97}$$

Outside of $\mathscr{E}' \cap \mathscr{E}$, we upper bound $N_T^-(z)$ almost surely by $|\mathscr{T}^-(T)|$, itself bounded using the $\log^2(T)$-exploration barrier (Lemma III.31), that bounds it by $C\log^2(T)$ with probability $1 - o(T^{-2})$. If the $\log^2(T)$-exploration barrier fails, we upper-bound $|\mathscr{T}^-(T)|$ by trivial bound T. Taking the expectation, we obtain:

$$\mathbf{E}[|\mathscr{T}^-(T)|] \leq \ell_0 \mu_\eta^*(z)\Big(1 + \tfrac{4C\delta_0}{\mu_\eta^*(z)}\Big) + O\Big(\sqrt{\log(T)}\Big) + (\mathbf{P}(\mathscr{E}'^c \cap \mathscr{E}) + \mathbf{P}(\mathscr{E}^c)) \cdot C\log^2(T) + o(T \cdot T^{-2}). \tag{III.98}$$

By the trigger effect (Lemma III.30), we know that $\mathbf{P}(\mathscr{E}^c) = o(\exp(-\sqrt{\log(T)}))$. Using this and unfolding the definition of the exploration threshold ℓ_0 (see (III.82)), we get:

$$\begin{aligned}
\mathbf{E}[N_T^-(z)] &\leq \ell_0 \mu_\eta^*(z)\Big(1 + \tfrac{4C\delta_0}{\mu_\eta^*(z)}\Big) + O\Big(\sqrt{\log(T)}\Big) + o\Big(\exp\Big(-\sqrt{\log(T)}\Big)\Big) \cdot C\log^2(T) + o(T \cdot T^{-2}) \\
&= \ell_0 \mu_\eta^*(z)\Big(1 + \tfrac{4C\delta_0}{\mu_\eta^*(z)}\Big) + O\Big(\sqrt{\log(T)}\Big) \\
&= \max\left\{ \frac{(1+\delta(\gamma\log(T)))\|\mu_\eta^0(M)\|_1}{1 - \delta_0\Big(1 + \frac{4C}{\mu_\eta^*(z)}\Big)} \cdot \log(T), \frac{\gamma\log(T)+2}{\delta_0} \right\} \mu_\eta^*(z)\Big(1 + \tfrac{4C\delta_0}{\mu_\eta^*(z)}\Big) \\
&\quad + O\Big(\sqrt{\log(T)}\Big).
\end{aligned} \tag{III.99}$$

Unfolding the definition of $\mu_\eta^*(z)$ and using that $\delta(\gamma\log(T)) \sim \frac{1}{\log\log(T)} \to 0$, we get:

$$\limsup_{T\to\infty} \frac{\mathbf{E}\Big[\sum_{t=0}^{T-1} \mathbf{1}(Z_t = z, t \in \mathscr{T}^-)\Big]}{\log(T)} = \limsup_{T\to\infty} \frac{\mathbf{E}[N_T^-(z)]}{\log(T)} \leq \frac{1 + \frac{4C\delta_0}{\mu_\eta^*(z)}}{1 - \delta_0\Big(1 + \frac{4C}{\mu_\eta^*(z)}\Big)} \cdot \mu_\eta^0(z, M) + \frac{\gamma}{\delta_0}. \tag{III.100}$$

This bound holds for arbitrary γ, δ_0 that are small enough. Make the discard coefficient $\gamma \to 0$, then the precision parameter $\delta_0 \to 0$. This concludes the proof. ■

10.C.4.2 Uniform exploration guarantees

Lemma III.29 (Uniform exploration). *There exists constants $\alpha, \beta > 0$ such that:*

$$\mathbf{P}\Big(\exists z \in \mathscr{Z}, \exists t \geq 0 :\ N_t(z) < \alpha\ell \text{ and } |\mathscr{T}^-(t)| \geq \ell\Big) = o(\exp(-\beta\ell)). \tag{III.101}$$

Proof. The proof shares similarities with Lemma III.21. We consider the model M_η with pair space $\mathscr{Z}$ which kernels given by:

$$p_\eta(s,a) := (1 - \eta|\mathscr{A}(s)|)p(s,a) + \eta \sum_{a' \in \mathscr{A}(s)} p(s,a').$$

The obtained model M_η is communicating, because the execution of the fully uniform policy on M_η is equivalent to the execution of the uniform policy on M. By construction, the execution of any η-uniform policy π of M can be seen as the execution of a randomized policy π_η of M_η.

Fix $z_0 \in \mathscr{Z}$ and introduce the reward function $f(z) := \mathbf{1}(z = z_0)$. Consider the model M_η^{-f} with kernel p_η and reward function $-f$. Let $g_\eta^{-f}, h_\eta^{-f}, \Delta_\eta^{-f}$ the associated gain, bias and gap functions. Because M_η^{-f} is communicating, $\mathrm{sp}(g_\eta^{-f}) = 0$ and by picking π_η a deterministic bias optimal policy of M_η^{-f}, we have by construction that π_η corresponds to an η-uniform policy of M under which all pairs are recurrent; so z_0 is recurrent under π_η in particular, and $\min(g_\eta^{-f}) = \max(g_\eta^{-f}) =: -\alpha_{x_0} < 0$. Consider the gaps $\Delta^{-f}(s,a) := g_\eta^{-f}(s) + h_\eta^{-f}(s) + f(s) - p(s,a)h_\eta^{-f}$. Following from the Bellman equations of M_η^{-f}, check that every η-uniform policy π satisfies:

$$\sum_{a \in \mathscr{A}(s)} \pi_\eta(a|s)\Delta_\eta^{-f}(s,a) = \sum_{a \in \mathscr{A}(s)} \pi(a|s)\Delta^{-f}(s,a) \geq 0 \tag{III.102}$$

where π_η is the randomized policy induced by π (i.e., bissimulating) in M_η. Let $\beta_{z_0} := \max(\mathrm{sp}(h_\eta^{-f}), |\Delta^{-f}|) < \infty$. Let (τ_i) the stopping time enumeration of $\mathscr{T}^-$. We have:

$$\begin{aligned}
N_t(z_0) &= \sum_{i=0}^{\infty} \mathbf{1}(\tau_i < t) f(S_{\tau_i}, A_{\tau_i}) \\
&= \sum_{i=0}^{\infty} \mathbf{1}(\tau_i < t)\Big(-g_\eta^{-f}(S_{\tau_i}) + \Big(p(S_{\tau_i}, A_{\tau_i}) - e_{S_{\tau_i}}\Big)h_\eta^{-f} + \Delta^{-f}(S_{\tau_i}, A_{\tau_i})\Big) \\
&= \alpha_{z_0}|\mathscr{T}^-(t)| + \underbrace{\sum_{i=0}^{\infty} \mathbf{1}(\tau_i < t)\Big(p(S_{\tau_i}, A_{\tau_i}) - e_{S_{\tau_i}}\Big)h_\eta^{-f}}_{\mathrm{A}_1} + \underbrace{\sum_{i=0}^{\infty} \mathbf{1}(\tau_i < t)\Delta^{-f}(S_{\tau_i}, A_{\tau_i})}_{\mathrm{A}_2}.
\end{aligned}$$

We control the error terms as follows. We start with A_1.

$$\begin{aligned}
\mathrm{A}_1 &\geq -\beta_{x_0} + \sum_{i=0}^{\infty} \mathbf{1}(\tau_i < t)\Big(h_\eta^{-f}(S_{\tau_i+1}) - h_\eta^{-f}(S_{\tau_{i+1}})\Big) + \sum_{i=0}^{\infty} \mathbf{1}(\tau_i < t)\Big(p(S_{\tau_i}, A_{\tau_i}) - e_{S_{\tau_i+1}}\Big)h_\eta^{-f} \\
&\overset{(\dagger)}{=} -(1 + |\mathscr{Z}|)\beta_{x_0} + \sum_{i=0}^{\infty} \mathbf{1}(\tau_i < t)\Big(p(S_{\tau_i}, A_{\tau_i}) - e_{S_{\tau_i+1}}\Big)h_\eta^{-f}
\end{aligned}$$

where (†) follows from the observation that, between exploration times, the algorithm is either co-explorating or exploiting. When it does, it does it until regeneration hence coming back to the state where exploration was suspended; At the exception of phases killed by panicking, but there are at most $|\mathscr{Z}|$ of them. We continue with A_2.

$$\begin{aligned}
\mathrm{A}_2 &:= \sum_{i=0}^{\infty} \mathbf{1}(\tau_i < t)\Delta^{-f}(S_{\tau_i}, A_{\tau_i}) \\
&= \sum_{i=0}^{\infty} \mathbf{1}(\tau_i < t) \sum_{a \in \mathscr{A}(S_{\tau_i})} \pi^-_{\tau_i}(a|S_{\tau_i})\Delta^{-f}(S_{\tau_i}, a) \\
&\quad + \sum_{i=0}^{\infty} \mathbf{1}(\tau_i < t) \sum_{a \in \mathscr{A}(S_{\tau_i})} \Big(\mathbf{1}(A_{\tau_i} = a) - \pi^-_{\tau_i}(a|S_{\tau_i})\Big)\Delta^{-f}(S_{\tau_i}, a) \\
&\overset{(\dagger)}{\geq} \sum_{i=0}^{\infty} \mathbf{1}(\tau_i < t) \sum_{a \in \mathscr{A}(S_{\tau_i})} \Big(\mathbf{1}(A_{\tau_i} = a) - \pi^-_{\tau_i}(a|S_{\tau_i})\Big)\Delta^{-f}(S_{\tau_i}, a)
\end{aligned}$$

where (†) follows from (III.102). We conclude that A_1 and A_2 both involve martingale obtained as the sum of $|\mathscr{T}^-(t)|$ martingales differences whose terms have span at most β_{z_0}. Invoking a

time-uniform Azuma-Hoeffding inequality (Lemma I.22) to bound each of them, we find that

$$N_t(z_0) \geq \alpha_{z_0}|\mathscr{T}^-(t)| - (1+|\mathscr{Z}|)\beta_{z_0} - 2\beta_{z_0}\sqrt{|\mathscr{T}^-(t)|\log\left(\frac{1+|\mathscr{T}^-(t)|}{\delta}\right)} \tag{III.103}$$

holds with probability $1-\delta$ at least. From (III.103), we get that for a fixed $\ell \in \mathbf{N}$,

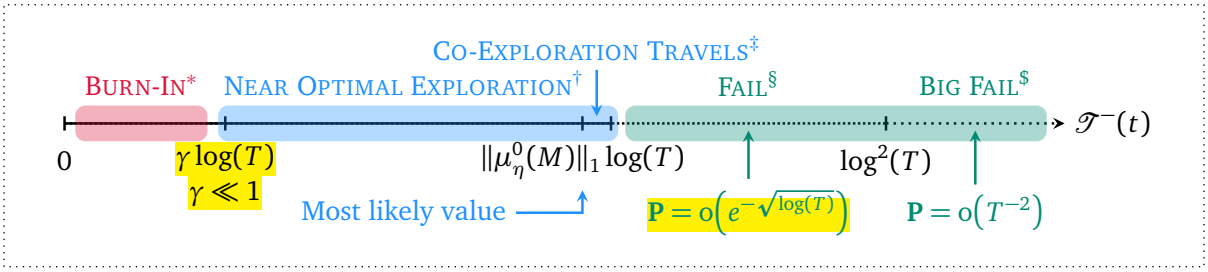
$$\mathbf{P}\left(\exists t \geq 0;\ N_t(z_0) < \alpha_{z_0}\ell - (1+|\mathscr{Z}|)\beta_{z_0} - 2\beta_{z_0}\sqrt{\ell\log\left(\tfrac{1+\ell}{\delta}\right)} \text{ and } |\mathscr{T}^-(t)| \geq \ell\right) \leq \delta. \tag{III.104}$$

For ℓ large enough, we have $(1+|\mathscr{Z}|)\beta_{z_0} < \frac{1}{3}\alpha_{z_0}\ell$. Also, straight forward algebra gives:

$$2\beta_{z_0}\sqrt{\ell\log\left(\tfrac{1+\ell}{\delta}\right)} < \tfrac{1}{3}\alpha_{z_0}\ell \iff \delta > \exp\left(-\left(\frac{\alpha_{z_0}}{6\beta_{z_0}}\right)^2\ell + \log(1+\ell)\right).$$

Accordingly, the tail of (III.104) is eventually sub-exponential, from which the result follows. □

10.C.4.3 The trigger effect



Lemma III.30 (Trigger effect). *The event $\mathscr{E} \equiv \mathscr{E}_{\epsilon_0,\gamma,T}$ given by:*

$$\mathscr{E} := \left(\forall t \leq T, |\mathscr{T}^-(t)| \geq \gamma\log(T) \Rightarrow \begin{pmatrix} \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) = \mathscr{Z}^{**}(M) \text{ and } \hat{M}_t \sim M \text{ and} \\ \|\mu_\eta^{\epsilon(t)}(\hat{M}_t) - \mu_\eta^0(M)\|_\infty < \eta\epsilon_0 \text{ and} \\ |K_\eta^{\epsilon(t)}(\hat{M}_t) - K_\eta^0(M)| < \epsilon_0 \end{pmatrix}\right)$$

For all $\epsilon_0, \gamma > 0$, we have $\mathbf{P}(\mathscr{E}^c) = o\left(\exp\left(-\sqrt{\log(T)}\right)\right)$ when $T \to \infty$.

Proof. The result is a combination of Lemma III.36 and Proposition III.12 and Theorem III.13. Fix $\epsilon > 0$. By Proposition III.12 and Theorem III.13, there exists $\epsilon_0' > 0$ such that, if $\epsilon(t) < \epsilon_0'$ and

$$\mathrm{KL}^*(\hat{M}_t||M) + \tfrac{1}{\epsilon(t)}\|\hat{M}_t - M\|^* < \epsilon_0' \tag{III.105}$$

then

$$(*) := \left(\begin{array}{c} \mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) = \mathscr{Z}^{**}(M) \text{ and } \hat{M}_t \sim M \text{ and} \\ \|\mu_\eta^{\epsilon(t)}(\hat{M}_t) - \mu_\eta^0(M)\|_\infty < \eta\epsilon_0 \text{ and } |K_\eta^{\epsilon(t)}(\hat{M}_t) - K_\eta^0(M)| < \epsilon_0 \end{array}\right)$$

holds. Because we are interested in times $t \leq T$, we can lower bound $\epsilon(t)$ by $\frac{1}{\log\log(T)}$. By Pinsker's inequality, we can bound $\|\hat{M}_t - M\|^*$ by $(\mathrm{KL}^*(\hat{M}_t||M))^{1/2}$ and up to assuming that $\epsilon_0' < 1$, we can change $\mathrm{KL}^*(\hat{M}_t||M)$ to $\mathrm{KL}^*(\hat{M}||M)$ in (III.105), to simplify (III.105) to:

$$\mathrm{KL}^*(\hat{M}_t||M) \leq \left(\tfrac{1}{2}\right)^2\epsilon_0'^2\epsilon(T)^2 = \left(\frac{\epsilon_0'}{2\log\log(T)}\right)^2. \tag{III.106}$$

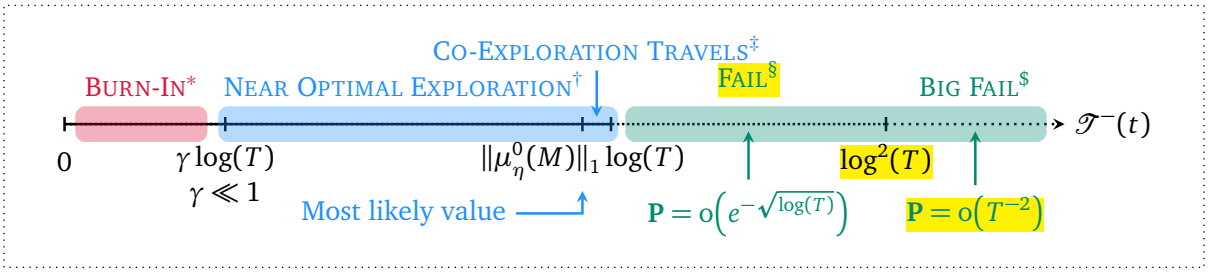
We now proceed as follows:

$$\mathbf{P}(\mathscr{E}^c) := \mathbf{P}(\exists t \leq T, |\mathscr{T}^-(t)| \geq \gamma\log(T) \text{ and } (*) \text{ is wrong})$$

$$
\begin{aligned}
&\leq \mathbf{P}\Big(\exists t \geq 0 : \begin{matrix}\min(N_t) < \alpha\gamma\log(T) \text{ and} \\ |\mathscr{T}^-(t)| \geq \gamma\log(T)\end{matrix}\Big) + \mathbf{P}\Big(\exists t \leq T : \begin{matrix}\min(N_t) \geq \alpha\gamma\log(T) \text{ and} \\ (*) \text{ is wrong}\end{matrix}\Big) \\
&\overset{(\dagger)}{\leq} o(\exp(-\beta\gamma\log(T))) + \mathbf{P}\Big(\exists t \leq T : \begin{matrix}\min(N_t) \geq \alpha\gamma\log(T) \text{ and} \\ \mathrm{KL}^*(\hat{M}_t||M) > \left(\frac{1}{2}\right)^2 \epsilon_0'^2 \epsilon(T)^2\end{matrix}\Big) \\
&\overset{(\ddagger)}{\leq} o(\exp(-\beta\gamma\log(T))) + \mathbf{P}\Big(\exists t \leq T : \begin{matrix}\min(N_t) \geq \alpha\gamma\log(T) \text{ and} \\ \mathrm{KL}(\hat{M}_t||M) > \left(\frac{1}{2}\right)^2 \epsilon_0''^2 \epsilon(T)^2\end{matrix}\Big) \\
&\overset{(\S)}{\leq} o(\exp(-\beta\gamma\log(T))) + 2|\mathscr{X}|\exp\Big(-\tfrac{\alpha\gamma\log(T)\epsilon_0''^2}{4|\mathscr{S}|\log\log(T)^2} + \log(1+\alpha\gamma\log(T)) + 1\Big) \\
&= o\Big(\exp\Big(-\sqrt{\log(T)}\Big)\Big).
\end{aligned}
$$

In the above, (†) bounds the left term using Lemma III.29 and the right term using (III.106), (‡) invokes Lemma III.37 by replacing ϵ_0' to an eventually smaller ϵ_0'', and (§) follows from concentrability results (Lemma III.36). □

10.C.4.4 The $\log^2(T)$ exploration barrier



Lemma III.31 ($\log^2(T)$ exploration barrier)**.** *There exists a constant $C > 0$ such that:*

$$\mathbf{P}\left(\exists t \leq T : |\mathscr{T}^-(t)| \geq C\log^2(T)\right) = o\left(T^{-2}\right). \tag{III.107}$$

Proof. Let (τ_j) the stopping time enumeration of $\mathscr{T}^-$. The sequence (τ_j) is made of segments where $\tau_{j+1} = \tau_j + 1$ and we denote $\tau_{j(i)}$ the starting index of the i-th segment. Let $\alpha, \beta > 0$ the constants provided by Lemma III.29. Introduce the event:

$$\mathscr{E}_t := \left(\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) = \mathscr{Z}^{**}(M) \text{ and } \mathscr{Z}_t = \mathscr{Z}\right),$$

stating that the skeleton $\mathscr{Z}_t$ is equal to $\mathscr{Z}$ and that the near optimal pairs are equal to $\mathscr{Z}^{**}(M)$, i.e., that the coexploration structure is correct. We have:

$$
\begin{aligned}
|\mathscr{T}^-(T)| &= \sum_{t=0}^{T-1} \mathbf{1}(t \in \mathscr{T}^-) = \sum_{i=1}^{\infty} \sum_{j'=j(i)}^{j(i+1)-1} \mathbf{1}(\tau_{j'} < T) \\
&\leq \frac{\log^2(T)}{\alpha} + \underbrace{\sum_{i=1}^{\infty} \sum_{j'=j(i)}^{j(i+1)-1} \mathbf{1}\left(\tfrac{\log^2(T)}{\alpha} \leq \tau_{j'} < T \text{ and } \mathscr{E}_{\tau_{j'}}\right)}_{\text{A}} + T \cdot \underbrace{\mathbf{1}\left(\exists j' \geq \tfrac{\log^2(T)}{\alpha}, \tau_{j'} < T \text{ and } \mathscr{E}_{\tau_{j'}}^c\right)}_{\text{B}}.
\end{aligned}
$$

In light of the proof of the trigger effect (Lemma III.30) and by Proposition III.12, there exists $\epsilon_0' > 0$ such that, if $\epsilon(T) \leq \epsilon(t) < \epsilon_0'$ and

$$\mathrm{KL}^*(\hat{M}_t||M) \leq \left(\tfrac{1}{2}\right)^2 \epsilon_0'^2 \epsilon(T)^2 = \left(\frac{\epsilon_0'}{2\log\log(T)}\right)^2 \tag{III.108}$$

then $\mathscr{Z}_{**}^{\epsilon(t)}(\hat{M}_t) = \mathscr{Z}^{**}(M)$. holds. By Proposition III.12 then invoking again the arguments of the proof of Lemma III.30, there exists $\epsilon_0' > 0$ such that if $\epsilon(T) \le \epsilon(t) < \epsilon_0'$ and $t \le T$, the condition

$$\mathrm{KL}^*(\hat{M}_t||M) \le \left(\tfrac{1}{2}\right)^2 \epsilon_0'^2 \epsilon(T)^2 = \left(\frac{\epsilon_0'}{2\log\log(T)}\right)^2 \tag{III.109}$$

implies $(*)$. We now proceed as follows:

$$\begin{aligned}
\mathbf{P}(\mathscr{E}^c) &:= \mathbf{P}(\exists t \le T, |\mathscr{T}^-(t)| \ge \gamma\log(T) \text{ and } (*) \text{ is wrong}) \\
&\le \mathbf{P}\left(\exists t \ge 0 : \begin{array}{c}\min(N_t) < \alpha\gamma\log(T) \text{ and} \\ |\mathscr{T}^-(t)| \ge \gamma\log(T)\end{array}\right) + \mathbf{P}\left(\exists t \le T : \begin{array}{c}\min(N_t) \ge \alpha\gamma\log(T) \text{ and} \\ (*) \text{ is wrong}\end{array}\right) \\
&\overset{(\dagger)}{\le} \mathrm{o}(\exp(-\beta\gamma\log(T))) + \mathbf{P}\left(\exists t \le T : \begin{array}{c}\min(N_t) \ge \alpha\gamma\log(T) \text{ and} \\ \mathrm{KL}^*(\hat{M}_t||M) > \left(\tfrac{1}{2}\right)^2 \epsilon_0'^2 \epsilon(T)^2\end{array}\right) \\
&\overset{(\ddagger)}{\le} \mathrm{o}(\exp(-\beta\gamma\log(T))) + \mathbf{P}\left(\exists t \le T : \begin{array}{c}\min(N_t) \ge \alpha\gamma\log(T) \text{ and} \\ \mathrm{KL}(\hat{M}_t||M) > \left(\tfrac{1}{2}\right)^2 \epsilon_0''^2 \epsilon(T)^2\end{array}\right) \\
&\overset{(\S)}{\le} \mathrm{o}(\exp(-\beta\gamma\log(T))) + 2|\mathscr{Z}|\exp\left(-\tfrac{\alpha\gamma\log(T)\epsilon_0''^2}{4|\mathscr{S}|\log\log(T)^2} + \log(1+\alpha\gamma\log(T)) + 1\right) \\
&= \mathrm{o}\left(\exp\left(-\sqrt{\log(T)}\right)\right).
\end{aligned}$$

In the above, (†) bounds the left term using Lemma III.29 and the right term using (III.106), (‡) invokes Lemma III.37 by replacing ϵ_0' to an eventually smaller ϵ_0'', and (§) follows from concentrability results (Lemma III.36). □

10.C.5 Adaptations of standard concentration results

Lemma III.32 ($\log^2$-concentration)**.** *We have:*

$$\mathbf{P}\left(\exists z \in \mathscr{Z} :\ N_t(z) > \log^2(t) \text{ and } \mathrm{KL}(\hat{M}_t(z)||M(z)) > \epsilon\right) = \mathrm{o}\left(t^{-\frac{\epsilon\log(t)}{2|\mathscr{S}|}}\right). \tag{III.110}$$

Proof. Recall that $\mathrm{KL}(\hat{M}_t(z)||M(z)) = \mathrm{KL}(\hat{r}_t(z)||r(z)) + \mathrm{KL}(\hat{p}_t(z)||p(z))$. Let $q \in \{r, p\}$ and let d the dimension of q, that is either 2 or $|\mathscr{S}|$. By Lemma III.36, we have:

$$\mathbf{P}\left(N_t(z) \ge \log^2(t) \text{ and } \mathrm{KL}(\hat{q}_t(z)||q(z)) > \epsilon\right) \le \exp\left(-\frac{\epsilon\log^2(t)}{d-1} + \log(1+\log^2(t)) + 1\right).$$

Observe that the right-hand side is $\mathrm{o}\left(t^{-\frac{\epsilon\log(t)}{2(d-1)}}\right)$ when $t \to \infty$. □

Lemma III.33 (All time Sanov)**.** *For $n \in (\mathbf{N}^*)^{\mathscr{Z}}$, denote $\hat{M}^n$ the empirical model obtained with visits counts $N_t(z) = n(z)$ for all $z \in \mathscr{Z}$. Let $\mathscr{M}^n$ the discrete space of all models that $\hat{M}^n$ can be equal to. Then:*

$$\forall M' \in \mathscr{M}^n, \quad \mathbf{P}\left(\hat{M}^n = M'\right) \le \exp\left(-\sum_{z\in\mathscr{Z}} n(z)\mathrm{KL}(M'(z)||M(z))\right). \tag{III.111}$$

Proof. For $m \ge 1$, write $r^m(z)$ and $p^m(z)$ the empirical reward and kernel at z that are observed under $N_t(z) = m$. Recall that $\mathrm{KL}(\hat{M}_t(z)||M(z)) = \mathrm{KL}(\hat{r}_t(z)||r(z)) + \mathrm{KL}(\hat{p}_t(z)||p(z))$. We have:

$$\mathbf{P}\left(\hat{M}^n = M'\right) = \prod_{z\in\mathscr{Z}} \mathbf{P}\left(\hat{r}^{n(z)}(z) = r'(z)\right)\mathbf{P}\left(\hat{p}^{n(z)}(z) = p'(z)\right). \tag{III.112}$$

By definition, $r'(z)$ is of the form $\mathrm{B}(k(z)/n(z))$ with $k \in \{0, ..., n(z)\}$. So:

$$\begin{aligned}\mathbf{P}\big(\hat{r}^{n(z)} = r'(z)\big) &= \binom{n(z)}{k(z)} r(z)^{k(z)}(1-r(z))^{n(z)-k(z)} \\ &= \exp(-n(z)\mathrm{KL}(r'(z)||r(z))) \cdot \binom{n(z)}{k(z)} \exp\big(-n(z)\mathrm{Ent}(r'(z))\big) \\ &\overset{(\dagger)}{\leq} \exp(-n(z)\mathrm{KL}(r'(z)||r(z)))\end{aligned}$$

where (†) follows from the classical inequality $\binom{a}{b} \leq \exp(a\mathrm{Ent}(\frac{b}{a}))$ where $\mathrm{Ent}(-)$ is the entropy (in base e). With the same computation for kernels, we show that:

$$\mathbf{P}\big(\hat{p}^{n(z)}(z) = p'(z)\big) \leq \exp(-n(z)\mathrm{KL}(p'(z)||p(z))).$$

Combining everything, we obtain the result. □

Lemma III.34 (Combinatorics)**.** *Let $k \geq 1$ and denote $\mathscr{P}_n[k]$ the set of probability distributions over $\{1, \dots, k\}$ of the form $(\frac{n_1}{n}, \dots, \frac{n_k}{n})$ with $n_i \in \mathbf{N}$. Then $|\mathscr{P}_n[k]| \leq (n+1)^k$.*

Lemma III.35 (Time-uniform empirical likelihood deviations, Jonsson et al. (2020))**.** *Let $d \geq 1$. Fix q a distribution on $\{1, \dots, d\}$ and denote q^n the empirical distribution on $\{1, \dots, d\}$ obtained after n i.i.d. samples of q.*

$$\forall \delta > 0, \quad \mathbf{P}\Big(\exists n \geq 1,\ n\mathrm{KL}(q^n||q) > \log\big(\tfrac{1}{\delta}\big) + (d-1)\log\big(e\big(1 + \tfrac{n}{d-1}\big)\big)\Big) \leq \delta. \tag{III.113}$$

Lemma III.36 (Threshold concentration)**.** *Let $m \geq 0$ and $d \geq 2$. Fix q a distribution on $\{1, \dots, d\}$ and denote q^n the empirical distribution on $\{1, \dots, d\}$ obtained after n i.i.d. samples of q. Then:*

$$\forall \epsilon > 0, \quad \mathbf{P}(\exists n \geq m : \mathrm{KL}(q^n||q) > \epsilon) \leq \exp\Big(-\frac{\epsilon m}{d-1} + \log(1+m) + 1\Big). \tag{III.114}$$

Proof. By Lemma III.35, we have:

$$\begin{aligned}\delta &\geq \mathbf{P}\big(\exists n \geq m, n\mathrm{KL}(q^n||q) \geq (d-1)\log\big(\tfrac{e}{\delta}\big(1 + \tfrac{n}{d-1}\big)\big)\big) \\ &\geq \mathbf{P}\big(\exists n \geq m, \mathrm{KL}(q^n||q) \geq \tfrac{d-1}{n}\log\big(\tfrac{e}{\delta}(1+n)\big)\big) \\ &\overset{(\dagger)}{\geq} \mathbf{P}(\exists n \geq m, \mathrm{KL}(q^n||q) \geq \epsilon)\end{aligned}$$

where (†) if for all $n \geq m$, we have:

$$\tfrac{d-1}{n}\log\big(\tfrac{e}{\delta}(1+n)\big) \leq \epsilon.$$

Solving in $\delta > 0$, we get the necessary condition that, for all $n \geq m$, we have:

$$\delta \geq e(1+n)\exp\big(-\tfrac{\epsilon n}{d-1}\big) =: \varphi(n)$$

By computing derivates, we find that $\varphi(n)$ is decreasing on $(\frac{d-1}{\epsilon} - 1, \infty)$, so for $m \geq \frac{d-1}{\epsilon} - 1$, we may set $\delta = e(1+m)\exp(-\frac{\epsilon m}{d-1})$. For $m < \frac{d-1}{\epsilon} - 1$, we see that $e(1+m)\exp(-\frac{\epsilon m}{d-1}) \geq 1$. □

Lemma III.37 (Converting KL^* to KL.)**.** *Let $d \geq 2$ and fix q a distribution on $\{1, \dots, d\}$. Denote $\epsilon := \min\{q(i) : q(i) > 0\} > 0$. If*

$$\mathrm{KL}(q'||q) < \log\left(\frac{1}{1-\epsilon}\right)$$

then $q \sim q'$, i.e., they have the same support and $\mathrm{KL}(q'||q) = \mathrm{KL}^*(q'||q)$.

Proof. It is known that if $\mathrm{KL}(q'||q) < \infty$, then $q' \ll q$. Assume that the support of q' is $\mathscr{S}' \subseteq \{1,\dots,d\}$ and isn't equal to the one of q. Consider the optimization problem: minimize $\mathrm{KL}(q^*||q)$ subject to $\mathrm{supp}(q^*) = \mathscr{S}'$. Using KKT conditions, we find that q^* and q must be proportional. So:

$$\mathrm{KL}(q'||q) = \sum_{i\in\mathscr{S}} q'(i)\log\left(\frac{q'(i)}{q(i)}\right) \geq \sum_{i\in\mathscr{S}} \frac{q(i)}{\|q\|_{1,\mathscr{S}'}}\log\left(\frac{1}{\|q\|_{1,\mathscr{S}'}}\right) = \log\left(\frac{1}{\|q\|_{1,\mathscr{S}'}}\right) \geq \log\left(\frac{1}{1-\epsilon}\right)$$

hence proving the result. □

10.D Deviation bounds of MDP specific quantities

The goal of this section is to provide a complete machinery to bound the variations of various policy related quantities, such as the gain, the bias, the diameter, the invariant measure and the reaching probabilities. One important consequence of the developed results is that if M, M' are two Markov decision processes, then

$$\mathscr{Z}^{\epsilon}_{**}(M') = \mathscr{Z}^{**}(M) \quad \text{when} \quad D_*(M)\|M'-M\|^* \ll \epsilon \tag{III.115}$$

where $D_*(M)$ is a notion of **worst diameter** (Equation (III.131)), and $\epsilon > 0$ is smaller than the gain gap of M. This shows that the near optimal pairs of M' can be explicitly related to the optimal pairs of M (Proposition III.48). The order of magnitude is tight in a minimax sense, although $D_*(M)$ could be changed to $D(M)$.[4] Similarly, we can show

$$\Delta^{\epsilon}_*(z,M') = \Delta^*(z,M) + \mathrm{O}\big(D_*(M)^2\|M'-M\|^*\big) \quad \text{when} \quad D_*(M)\|M'-M\|^* \ll \epsilon \tag{III.116}$$

which is also tight in order in magnitude (Proposition III.49). All bounds are developed using a martingale technique coupled with a Poisson equation, sometimes under a transform of the initial kernel and reward function. To provide a general flavor of the technique, consider two Markov reward processes (r_1,p_1) and (r_2,p_2), and denote g_j, h_j their respective gain and bias functions. We have the Poisson equation $g_j(s) + h_j(s) = r_j(s) + p_j(s)h_j$. We write S_t the random state at time t, and $\mathbf{P}^{p_j}_s(-), \mathbf{E}^{p_j}_s[-]$ the probability and expectation of the trajectory governed by p_j initialized at $S_1 = s$. Consider a stopping time τ such that $\mathbf{E}^{p_2}_s[\tau] < \infty$. The heart of the technique lies in the following computation:

$$\begin{aligned}
(-) &:= \mathbf{E}^{p_2}_s\left[\sum_{t=1}^{\tau-1} r_2(S_t)\right] \\
&= \mathbf{E}^{p_2}_s\left[\sum_{t=1}^{\tau-1}(r_2(S_t)-r_1(S_t))\right] + \mathbf{E}^{p_2}_s\left[\sum_{t=1}^{\tau-1} r_1(S_t)\right] \\
&\overset{(\dagger)}{=} \mathbf{E}^{p_2}_s\left[\sum_{t=1}^{\tau-1}(r_2(S_t)-r_1(S_t))\right] + \mathbf{E}^{p_2}_s\left[\sum_{t=1}^{\tau-1}\big(g_1(S_t) + \big(\mathbf{e}_{S_t} - p_1(S_t)\big)h_1\big)\right] \\
&= \mathbf{E}^{p_2}_s\left[\sum_{t=1}^{\tau-1} g_1(S_t)\right] + \mathbf{E}^{p_2}_s\left[\sum_{t=1}^{\tau-1}(r_2(S_t)-r_1(S_t))\right] + \mathbf{E}^{p_2}_s\left[\sum_{t=1}^{\tau-1}(p_2(S_t)-p_1(S_t))h_1\right] \\
&\qquad + \mathbf{E}^{p_2}_s[h_1(s)-h_1(S_\tau)] \\
&\overset{(\ddagger)}{\leq} \mathbf{E}^{p_2}_s\left[\sum_{t=1}^{\tau-1} g_1(S_t)\right] + \mathbf{E}^{p_2}_s[h_1(s)-h_1(S_\tau)] + \mathbf{E}^{p_2}_s[\tau]\big(\|r_2-r_1\|_\infty + \tfrac{1}{2}\mathrm{sp}(h_1)\|p_2-p_1\|_1\big)
\end{aligned}$$

[4] The proof would need to be improved. Because this is meant to be applied for the qualitative regret upper bound, we don't push the analysis to an optimal result.

where (†) follows from the Poisson equation of (r_1, p_1) and (‡) from standard norm bounds. Coupled with model transformation and careful choices of stopping times, this technique provides tight bounds for the variations of the gain, bias, diameter, reaching time and invariant measures. The obtained bounds on the variations of gain for instance (see Lemma III.41 and Lemma III.45) are much better than those obtained using algebraic approaches involving the Drazin inverse with very little structural assumptions, going way beyond the scope of ergodic Markov reward processes.

10.D.1 A multichain-friendly diameter notion for Markov chains

The diameter provided in the definition is below generalizes the well-known notion of diameter in the communicating/recurrent setting. It is tuned to provide a simple bound on the span of the bias function.

Definition III.8 (Diameter of a policy/Markov chain)**.** *Let p the kernel of a Markov chain with recurrent components $\mathcal{S}_1, \dots, \mathcal{S}_k$. The **(policy) diameter** of p is given by:*

$$D(p) := \sup_{(s_i) \in \prod_i \mathcal{S}_i} \sup_{s \in \mathcal{S}} \mathbf{E}_s^p[\tau_{\{s_1,\dots,s_k\}}] < \infty. \tag{III.117}$$

Lemma III.38 (Policy bias and diameter)**.** *Let (r, p) a Markov reward process and denote g, h its gain and bias functions. Then* $\operatorname{sp}(h) \le 2 \operatorname{sp}(r) D(p)$.

Proof. For $(s_i) \in \prod_{i=1}^k \mathcal{S}_i$ a covering of the recurrent components, we denote $\tau \equiv \tau_{\{s_1,\dots,s_k\}}$ for short. We have $\mathbf{E}_s^p[\tau] \le D(p) < \infty$ for all $s \in \mathcal{S}$ by construction. Because $\mu h = 0$ for every invariant measure of p, we see that for all $i = 1, \dots, k$, there must be $s_i \in \mathcal{S}_i$ such that $h(s_i) \ge 0$, and another $s_i \in \mathcal{S}_i$ such that $h(s_i) \le 0$. Assume that $h(s_i) \le 0$ for all $i = 1, \dots, k$. Then:

$$\begin{aligned} \max(r-g) D(p) &\ge \mathbf{E}_s^p\left[\sum_{t=1}^{\tau-1} (r(S_t) - g(S_t))\right] \\ &\overset{(\dagger)}{=} \mathbf{E}_s^p\left[\sum_{t=1}^{\tau-1} \big(\mathbf{e}_{S_t} - p(S_t)\big) h\right] \\ &= h(s) - \mathbf{E}_s^p[h(S_\tau)] = h(s) - \sum_{i=1}^k \mathbf{P}_s^p\big(\tau_{\mathcal{S}_i} < \infty\big) h(s_i) \overset{(\ddagger)}{\ge} h(s). \end{aligned} \tag{III.118}$$

where (†) follows from the Poisson equation $g(s) + h(s) = r(s) + p(s)h$ and (‡) by $h(s_i) \le 0$. With similar computations, and picking s_i such that $h(s_i) \ge 0$ for $i = 1, \dots, k$, we have:

$$\min(r-g) D(p) \le h(s) - \sum_{i=1}^k \mathbf{P}_s^p\big(\tau_{\mathcal{S}_i} < \infty\big) h(s_i) \le h(s). \tag{III.119}$$

Combining (III.118) and (III.119), we obtain:

$$\max(h) - \min(h) \le D(p)(\max(r-g) - \min(r-g)) \le 2 \operatorname{sp}(r) D(p) \tag{III.120}$$

where the second inequality uses that $\min(r) \le g(s) \le \max(r)$ for all s. □

Lemma III.39 (Variations of policy diameter)**.** *Let $p_1 \sim p_2$ two equivalent Markov chains with (common) recurrent components $\mathcal{S}_1, \dots, \mathcal{S}_k$. For all $(s_i) \in \prod_{i=1}^k \mathcal{S}_i$ covering of the recurrent components, the variations of the hitting time $\tau \equiv \tau_{\{s_1,\dots,s_k\}}$ are bounded as:*

$$\left| \sup_s \mathbf{E}_s^{p_2}[\tau] - \sup_s \mathbf{E}_s^{p_1}[\tau] \right| \le \frac{1}{2} \sup_s \mathbf{E}_s^{p_2}[\tau] \sup_s \mathbf{E}_s^{p_1}[\tau] \|p_2 - p_1\|_1. \tag{III.121}$$

In particular $D(p_2) \le D(p_1) + \frac{1}{2} D(p_1) D(p_2) \|p_2 - p_1\|_1$.

Proof. Let $(s_i) \in \prod_{i=1}^k \mathscr{S}_i$ a covering of the recurrent components and denote $\tau \equiv \tau_{\{s_1,\dots,s_k\}}$ for short. Consider the kernel p'_j obtained by making every s_i absorbing and consider the reward function $r'(s) := \mathbf{1}(s \notin \{s_1,\dots,s_k\})$. The gain and bias functions of the Markov reward process (r', p'_j) are denoted g'_j and h'_j for $j = 1, 2$. Remark that $g'_j = 0$ and that the bias satisfies

$$h'_j(s) = \operatorname*{Clim}_{T\to\infty} \mathbf{E}_s^{p'_j}\left[\sum_{t=1}^{\tau\wedge T-1} \left(r'(S_t) - g'_j(S_t)\right)\right] = \mathbf{E}_s^{p_j}[\tau] \tag{III.122}$$

hence h'_j are the reaching times of which we want to bound the variations. We have:

$$h'_2(s) = \mathbf{E}_s^{p_2}\left[\sum_{t=1}^{\tau-1} r'(S_t)\right] = h'_1(s) + \mathbf{E}_s^{p_2}\left[\sum_{t=1}^{\tau-1} (p_2(S_t) - p_1(S_t))h'_1\right].$$

The error term is bounded by $|\mathbf{E}_s^{p_2}[\sum_{t=1}^{\tau-1}(p_2(S_t) - p_1(S_t))h'_1]| \le \frac{1}{2}\mathrm{sp}(h'_1)\mathbf{E}_s^{p_2}[\tau]\|p_2 - p_1\|_1$. By (III.122), we clearly have $\mathrm{sp}(h'_1) = \max(h'_1) = \sup_s \mathbf{E}_s^{p_1}[\tau]$. Accordingly, we have obtained the self-bound:

$$\left|\sup_s \mathbf{E}_s^{p_2}[\tau] - \sup_s \mathbf{E}_s^{p_1}[\tau]\right| \le \tfrac{1}{2} \sup_s \mathbf{E}_s^{p_2}[\tau] \sup_s \mathbf{E}_s^{p_1}[\tau] \|p_2 - p_1\|_1. \tag{III.123}$$

This concludes the proof. □

The assumption $p_1 \sim p_2$ is not always necessary and can be dropped under a recurrent assumption.

Lemma III.40 (Variations of policy diameter, recurrent case)**.** *Let p_1, p_2 two recurrent Markov chains. We have:*

$$|D(p_1) - D(p_2)| \le \frac{1}{2} D(p_1) D(p_2) \|p_2 - p_1\|_1. \tag{III.124}$$

Proof. Same proof as Lemma III.39. □

The inequality $D(p_2) \le D(p_1) + \frac{1}{2}D(p_1)D(p_2)\|p_2 - p_1\|_1$ is a self-bound for $D(p_2)$. It can be decoupled when $\|p_2 - p_1\|_1$ is small. If $\|p_2 - p_1\|_1 < \frac{2}{D(p_1)}$, we obtain:

$$D(p_2) \le \frac{1}{1 - \frac{1}{2}D(p_1)\|p_2 - p_1\|_1} D(p_1). \tag{III.125}$$

For instance, if $\|p_2 - p_1\|_1 \le \frac{1}{D(p_1)}$, then $D(p_2) \le 2D(p_1)$.

10.D.2 Results for unichain Markov reward processes

The first lemma is a duplicate of Theorem II.1.

Lemma III.41 (Unichain gain variations)**.** *Let (r_1, p_1) and (r_2, p_2) two Markov reward processes. Denote g_j, h_j the gain and bias functions of (r_j, p_j) for $j = 1, 2$. Assume that $\mathrm{sp}(g_1) = 0$. Then:*

$$\|g_2 - g_1\|_\infty \le \|r_2 - r_1\|_\infty + \tfrac{1}{2}\mathrm{sp}(h_1)\|p_2 - p_1\|_1.$$

If in addition, $p_1 \sim p_2$, then $\mathrm{sp}(h_1)$ can be changed to $\mathrm{sp}(h_1|_{\mathscr{S}_1})$ where $\mathscr{S}_1$ is the collection of recurrent states under p_1.

Proof. This is direct adaptation of tutorial computation. Refer to the proof of Theorem II.1 for details. □

Lemma III.42 (Unichain invariant measure variations)**.** *Let $p_1 \sim p_2$ two Markov chains and assume that p_1 is unichain.*[5] *Denote μ_j the (unique) probability invariant measure under p_j. We have:*

$$\|\mu_2 - \mu_1\|_\infty \leq \min_{j\in\{1,2\}} D(p_j)\, \|p_2 - p_1\|_1.$$

Proof. Fix $s_0 \in \mathscr{S}$ a recurrent state. Consider the reward function $r(s) = \mathbf{1}(s = s_0)$, and denote g_j, h_j the gain and bias functions of the Markov reward process (r, h_j). Remark that $g_j(s) = \mu_j(s_0)$. By Lemma III.38, we have $\mathrm{sp}(h_j) \leq 2\mathrm{sp}(r)D(p_j) \leq 2D(p_j)$. Continuing with Lemma III.41, we have:

$$|\mu_2(s) - \mu_1(s)| \leq D(p_j)\|p_2 - p_1\|_1$$

for both $j = 1, 2$. This concludes the proof. □

Lemma III.43 (Unichain bias variations)**.** *Let (r_1, p_1) and (r_2, p_2) two Markov reward processes, assume that p_1 is unichain and that $p_1 \sim p_2$. Denote g_j and h_j the gain and bias function of (r_j, p_j) for $j = 1, 2$. We have:*

$$\|h_2 - h_1\|_\infty \leq 4D(p_2)\|r_2 - r_1\|_\infty + \big(2D(p_2)\mathrm{sp}(h_1) + \tfrac{1}{2}\mathrm{sp}(h_1^1)\big)\|p_2 - p_1\|_1$$

Proof. We start by bounding $\mathrm{sp}(h_2 - h_1)$. Fix s_0 a recurrent state and consider the transform p_j' that makes s_0 absorbing as well as the reward function $r_j'(s) := \mathbf{1}(s \neq s_0)(r_j(s) - g_j(s))$. Denote g_j' and h_j' the associated gain and bias functions. From direct computations, we see that $g_j' = 0$ and that $h_j'(s) = h_j(s) - h_j(s_0)$, so in particular $h_j' - h_j \in \mathbf{Re}$ so $\mathrm{sp}(h_j') = \mathrm{sp}(h_j)$ and $\mathrm{sp}(h_2' - h_1') = \mathrm{sp}(h_2 - h_1)$. Moreover, h_2' and h_1' are related as follows:

$$\begin{aligned}
h_2'(s) &= \mathbf{E}_s^{p_2'}\left[\sum_{t=1}^{\tau_{s_0}-1} r_2'(S_t)\right] \\
&\leq \mathbf{E}_s^{p_2}\big[\tau_{s_0}\big]\big\|r_2' - r_1'\big\|_\infty + \mathbf{E}_s^{p_2}\left[\sum_{t=1}^{\tau_{s_0}-1}\big(\mathbf{e}_{S_t} - p_1(S_t)\big)h_1'\right] \\
&\leq h_1'(s) + \mathbf{E}_s^{p_2}\big[\tau_{s_0}\big]\big(\|r_2 - r_1\|_\infty + \|g_2 - g_1\|_\infty + \tfrac{1}{2}\mathrm{sp}(h_1')\big\|p_2' - p_1'\big\|_1\big) \\
&\leq h_1'(s) + \mathbf{E}_s^{p_2}\big[\tau_{s_0}\big]\big(2\|r_2 - r_1\|_\infty + \mathrm{sp}(h_1)\big\|p_2' - p_1'\big\|_1\big)
\end{aligned}$$

where the last inequality is a consequence of Lemma III.41 and $\mathrm{sp}(h_1') \leq \mathrm{sp}(h_1)$. With the same technique, we lower bound $h_2'(s)$ by $h_1'(s)$ with the same error term. Remark that $\mathbf{E}_s^{p_j}[\tau_{s_0}] \leq D(p_j)$. So:

$$\begin{aligned}
\mathrm{sp}(h_2 - h_1) = \mathrm{sp}(h_2' - h_1') &= (h_2' - h_1')(s_{\max}) - (h_2' - h_1')(s_{\min}) \\
&\leq 2\, D(p_2)\big(2\|r_2 - r_1\|_\infty + \mathrm{sp}(h_1)\big\|p_2' - p_1'\big\|_1\big).
\end{aligned} \tag{III.126}$$

To transform the bound on the span to a bound on the norm, remark if $\mathbf{u}, \mathbf{v}$ are two vectors and $\mathbf{q}$ is a probability distribution, then $\|\mathbf{u} - \mathbf{v}\|_\infty \leq \mathrm{sp}(\mathbf{u} - \mathbf{v}) + |\mathbf{q}(\mathbf{u} - \mathbf{v})|$. Let μ_j the (unique) invariant probability measure under p_j. We have

$$\|h_2 - h_1\|_\infty \leq \mathrm{sp}(h_2 - h_1) + |\mu_2(h_2 - h_1)|. \tag{III.127}$$

The left term of (III.127) is taken care of with (III.126) and we are left to bound $|\mu_2(h_2 - h_1)|$. Remark that h_j^1, the 1-th higher order bias, is the bias of the Markov reward process $(-h_j, p_j)$.

[5] Hence p_2 is equally unichain with the same recurrent class.

We have:

$$\begin{aligned}|\mu_2(h_2-h_1)| = |\mu_2 h_1| &= \lim_{T\to\infty}\frac{1}{T}\left|\mathbf{E}_s^{p_2}\left[\sum_{t=1}^{T-1} h_1(S_t)\right]\right| \\ &= \lim_{T\to\infty}\frac{1}{T}\left|\mathbf{E}_s^{p_2}\left[\sum_{t=1}^{T-1}\big(\mathbf{e}_{S_t} - p_1(S_t)\big)h_1^1\right]\right| \\ &= \lim_{T\to\infty}\frac{1}{T}\left|\mathbf{E}_s^{p_2}\left[\sum_{t=1}^{T-1}(p_2(S_t)-p_1(S_t))h_1^1\right]\right| \leq \tfrac{1}{2}\mathrm{sp}(h_1^1)\|p_2-p_1\|_1.\end{aligned} \tag{III.128}$$

Combining (III.126), (III.127) and (III.128), we obtain the desired bound. □

10.D.3 Results for multichain Markov reward processes

We generalize the inequalities of the unichain setting to general multichain Markov processes. The new difficulty is that in the multichain setting, a perturbation of the kernel changes the probability of reaching a given recurrent component. The first result below is key. It is perhaps surprising, because it shows that the reaching probabilities vary additively and not multiplicatively with respect to kernel modifications.

Definition III.9. *We say that two Markov chains are* ***equivalent*** *and write $p_1 \sim p_2$ if $p_1 \ll p_2 \ll p_1$.*

Lemma III.44 (Reaching probabilities variations)**.** *Let $p_1 \sim p_2$ two equivalent Markov chains and let $\mathscr{S}_1^1, \dots, \mathscr{S}_1^k$ the recurrent classes of p_1. Let $\tau_\infty := \inf\{t \geq 1 : S_t \in \mathscr{S}_1^1 \cup \dots \cup \mathscr{S}_1^k\}$ the reaching time to one of the recurrent classes and $\tau_i := \inf\{t \geq 1 : S_t \in \mathscr{S}_1^i\}$ the reaching time to $\mathscr{S}_1^i$. We have:*

$$\left|\mathbf{P}_s^{p_1}(\tau_i < \infty) - \mathbf{P}_s^{p_2}(\tau_i < \infty)\right| \leq \min_{j\in\{1,2\}} \tfrac{1}{2}\mathbf{E}_s^{p_j}[\tau_\infty]\, \|p_1 - p_2\|_1.$$

In case, the bound can be simplified using $\mathbf{E}_s^{p_j}[\tau_\infty] \leq D(p_j)$.

Proof. Consider the reward function $r(S_t, S_{t+1}) = \mathbf{1}(S_t \notin \mathscr{S}_1^i \text{ and } S_t \in \mathscr{S}_1^i)$, providing reward when the walk crosses the frontier between $\mathscr{S}_1^i$ and its $\mathscr{S} \setminus \mathscr{S}_1^i$. For $j \in \{1,2\}$, we denote g_j, h_j the gain and bias functions under the Markov reward process (r, p_j). Since $p_1 \sim p_2$, they have the same recurrent classes, we have $\mathbf{E}^{p_j}[\tau_i] < \infty$ for $j = 1,2$ and $g_j(s) = 0$. Therefore, for $s \notin \mathscr{S}_1^i$, we have:

$$\begin{aligned}h_j(s) &:= \lim_{T\to\infty} \mathbf{E}_s^{p_j}[r(S_1,S_2) + \dots + r(S_{T-1}, S_T)] \\ &= \lim_{T\to\infty} \mathbf{E}_s^{p_j}\left[\sum_{t=1}^{T-1} \mathbf{1}(S_t \notin \mathscr{S}_1^i \text{ and } S_{t+1} \in \mathscr{S}_1^i)\right] \\ &= \lim_{T\to\infty} \mathbf{E}_s^{p_j}\left[\sum_{t=1}^{T-1} \mathbf{1}(\tau_i = t+1)\right] \\ &= \lim_{T\to\infty} \sum_{t=1}^{T-1} \mathbf{P}_s^{p_j}(\tau_i = t+1) = \mathbf{P}_s^{p_j}(2 \leq \tau_i < \infty) = \mathbf{P}_s^{p_j}(\tau_i < \infty)\end{aligned}$$

where the last line use that $s \notin \mathscr{S}_1^i$. For $s \in \mathscr{S}_1^i$, we have $p_j(s) = 0$. Our goal is therefore to bound $h_2(s)$ with respect to $h_1(s)$. For $s \notin \mathscr{S}_1^i$, we have:

$$h_2(s) = \lim_{T\to\infty} \mathbf{E}_s^{p_2}\left[\sum_{t=1}^{T-1} r(S_t, S_{t+1})\right]$$

$$\overset{(\dagger)}{=} \lim_{T\to\infty} \mathbf{E}_s^{p_2}\left[\sum_{t=1}^{T-1}\big(\mathbf{e}_{S_t} - p_1(S_t)\big)h_1\right] = h_1(s) + \lim_{T\to\infty} \mathbf{E}_s^{p_2}\left[\sum_{t=1}^{T-1}(p_2(S_t) - p_1(S_t))h_1\right]$$

where (†) follows from Poisson's equation. We are left to bound the RHS. Observe that once S_t belongs to one of the recurrent components of p_1, every state s' that is reachable under p_1 satisfies $h_1(s') = 0$. Because $p_1 \sim p_2$, this is also true for any state that is reachable under p_2. Therefore:

$$\left|\lim_{T\to\infty} \mathbf{E}_s^{p_2}\left[\sum_{t=1}^{T-1}(p_2(S_t) - p_1(S_t))h_1\right]\right| \le \tfrac{1}{2}\mathbf{E}_s^{p_2}[\tau_\infty]\mathrm{sp}(h_1)\|p_2 - p_1\|_1 \le \tfrac{1}{2}\mathbf{E}_s^{p_2}[\tau_\infty]\,\|p_2 - p_1\|_1.$$

In other words, $|p_1(s) - p_2(s)| \le \frac{1}{2}\mathbf{E}_s^{p_2}[\tau_\infty]$. Because $j = 1$ and $j = 2$ play a symmetric role, $\mathbf{E}_s^{p_2}[\tau_\infty]$ can be changed to $\mathbf{E}_s^{p_1}[\tau_\infty]$. □

Lemma III.45 (Multichain gain variations)**.** *Let $p_1 \sim p_2$ two equivalent Markov chains and let $\mathscr{S}_1, \dots, \mathscr{S}_k$ the (common) recurrent classes. Let $\tau_\infty := \inf\{t \ge 1 : S_t \in \mathscr{S}_1 \cup \dots \cup \mathscr{S}_k\}$ the reaching time to one of the recurrent classes. We have:*

$$\|g_2 - g_1\|_\infty \le \|r_2 - r_1\|_\infty + \min_{j\in\{1,2\}} \tfrac{1}{2}\Big(\max_i \mathrm{sp}(h_j|_{\mathscr{S}_i}) + \tfrac{1}{2}k\mathbf{E}_s^{p_j}[\tau_\infty]\Big)\|p_2 - p_1\|_1$$

In case, the bound can be simplified using $\frac{1}{2}(\max_i \mathrm{sp}(h_j|_{\mathscr{S}_i}) + \frac{1}{2}k\mathbf{E}_s^{p_j}[\tau_\infty]) \le (1 + \frac{1}{4}k)D(p_j)$.

Proof. Let $s \in \mathscr{S}$. We have:

$$\begin{aligned}
(*) &= |g_2(s) - g_1(s)| \\
&\le \left|\sum_{i=1}^{k} \mathbf{P}_s^{p_2}(\tau_i < \infty) g_2(\mathscr{S}_i) - \sum_{i=1}^{k} \mathbf{P}_s^{p_1}(\tau_i < \infty) g_1(\mathscr{S}_i)\right| \\
&\le \sum_{i=1}^{j} \mathbf{P}_s^{p_2}(\tau_i < \infty)|g_2(\mathscr{S}_i) - g_1(\mathscr{S}_i)| + \left|\sum_{i=1}^{j}\big(\mathbf{P}_s^{p_2}(\tau_i < \infty) - \mathbf{P}_s^{p_1}(\tau_i < \infty)\big) g_1(\mathscr{S}_i)\right| \\
&\overset{(\dagger)}{\le} \sum_{i=1}^{j} \mathbf{P}_s^{p_2}(\tau_i < \infty)\big(\|r_2 - r_1\|_\infty + \tfrac{1}{2}\mathrm{sp}(h_1|_{\mathscr{S}_i})\|p_2 - p_1\|_1\big) + \tfrac{1}{2}\left\|\big(\mathbf{P}_s^{p_2}(\tau_i < \infty) - \mathbf{P}_s^{p_1}(\tau_i < \infty)\big)_{i=1}^{k}\right\|_1 \\
&\overset{(\ddagger)}{\le} \|r_2 - r_1\|_\infty + \tfrac{1}{2}\Big(\max_i \mathrm{sp}(h_1|_{\mathscr{S}_i}) + \tfrac{1}{2}k\mathbf{E}_s^{p_1}[\tau_\infty]\Big)\|p_2 - p_1\|_1
\end{aligned}$$

where (†) follows by applying Lemma III.41 to the LHS and (‡) by applying Lemma III.44 to the RHS. □

Lemma III.46 (Multichain invariant measure variations)**.** *Let $p_1 \sim p_2$ two equivalent Markov chains and let $\mathscr{S}_1, \dots, \mathscr{S}_k$ the (common) recurrent classes. Let $\tau_\infty := \inf\{t \ge 1 : S_t \in \mathscr{S}_1 \cup \dots \cup \mathscr{S}_k\}$ the reaching time to one of the recurrent classes. Denote $\mu_j(-|s)$ the asymptotic empirical distribution of visits starting from s under p_j. We have:*

$$\forall s \in \mathscr{S}, \quad \|\mu_2(-|s) - \mu_1(-|s)\|_\infty \le \min_{j\in\{1,2\}} \tfrac{1}{2}\Big(\max_i D(p_j|_{\mathscr{S}_i}) + \tfrac{1}{2}k\mathbf{E}_s^{p_j}[\tau_\infty]\Big)\|p_2 - p_1\|_1.$$

In case, the bound can be simplified using $\frac{1}{2}(\max_i D(p_j|_{\mathscr{S}_i}) + \frac{1}{2}k\mathbf{E}_s^{p_j}[\tau_\infty]) \le (1 + \frac{1}{4}k)D(p_j)$.

Proof. This is a consequence of Lemma III.45 by considering the reward function $r(s) = \mathbf{1}(s = s_\infty)$ for $s_\infty \in \mathscr{S}$ a fixed state, similarly to Lemma III.42. Denote g_j, h_j denote the gain and bias

function of the Markov reward process (r, p_j). Remark that $g_j(s_0) = \mu_j(s_\infty|s_0)$. By Lemma III.45, we have:

$$|\mu_2(s_\infty|s_0) - \mu_1(s_\infty|s_0)| \le \min_{j\in\{1,2\}} \tfrac{1}{2}\Big(\max_i \operatorname{sp}(h_j|_{\mathscr{S}_i}) + \tfrac{1}{2}k\mathbf{E}_{s_0}^{p_j}[\tau_\infty]\Big)\|p_2 - p_1\|_1.$$

Because $\mathscr{S}_i$ is a recurrent class under p_j, it corresponds to the optimal bias of a communicating model, so by Lemma III.38, we have $\operatorname{sp}(h_j|_{\mathscr{S}_i}) \le \operatorname{sp}(r)D(p_j|_{\mathscr{S}_i}) \le D(p_j|_{\mathscr{S}_i})$. This provides the desired bound. □

Lemma III.47 (Multichain bias variations)**.** *Let (r_1, p_1) and (r_2, p_2) two Markov reward processes with $p_1 \sim p_2$ and let $\mathscr{S}_1, \dots, \mathscr{S}_k$ the (common) recurrent classes. Denote g_j, h_j the gain and bias functions of (r_j, p_j) for $j = 1, 2$. Let $\tau_\infty := \inf\{t \ge 1 : S_t \in \mathscr{S}_1 \cup \dots \cup \mathscr{S}_k\}$ the reaching time to one of the recurrent classes. We have*

$$\|h_2 - h_1\|_\infty \le 6D(p_2)\|r_2 - r_1\|_\infty + \big(\big(7 + \tfrac{1}{2}k\big)D(p_2)D(p_1) + 2D(p_1)^2\big)\|p_2 - p\|_1.$$

Following Lemma III.39, when $\|p_2 - p_1\|_1 \le \frac{1}{D(p_1)}$, the quantity $D(p_2)$ can be changed to $2D(p_1)$.

The result can be improved by more carefully tracking the Lipschitz constants throughout. To lighten the typography — and because our application do not require an optimal result, we overshoot the error term.

Proof. For $i = 1, \dots, k$, pick $s_0^i \in \mathscr{S}_i$ and denote $\mathscr{S}_0 = \{s_0^1, \dots, s_0^k\}$. For $j = 1, 2$, let p_j' the kernel obtained by copying p_j while making every s_0^i absorbing, and set $r_j'(s) := \mathbf{1}(s \notin \mathscr{S}_0)(r_j(s) - g_j(s))$. Denote g_j' and h_j' the gain and bias functions of the Markov reward process (r_j', p_j'). Observe that $g_j' = 0$ and $h_j'(s_0^i) = 0$ for all $i = 1, \dots, k$. Moreover, the biases h_j and h_j' are linked as follows:

$$\begin{aligned} h_j'(s) = \operatorname*{Clim}_{T\to\infty} \mathbf{E}_s^{p_j'}\left[\sum_{t=1}^{T-1} r_j'(S_t)\right] &= \operatorname*{Clim}_{T\to\infty} \mathbf{E}_s^{p_j'}\left[\sum_{t=1}^{\tau_{\mathscr{S}_0}\wedge T-1} \big(r_j(S_t) - g_j(S_t)\big)\right] \\ &= \mathbf{E}_s^{p_j}\left[\sum_{t=1}^{\tau_{\mathscr{S}_0}-1} \big(r_j(S_t) - g_j(S_t)\big)\right] \\ &= h_j(s) - \mathbf{E}_s^{p_j}\Big[h(S_{\tau_{\mathscr{S}_0}})\Big] = h_j(s) - \sum_{i=1}^{k} \mathbf{P}_s^{p_j}\big(\tau_{\mathscr{S}_i} < \infty\big)h_j(s_0^i). \end{aligned} \tag{III.129}$$

In particular, $\operatorname{sp}(h_j') \le 2\operatorname{sp}(h_j) \le 4D(p_j)$ by Lemma III.38. We have:

$$\begin{aligned} h_2'(s) = \mathbf{E}_s^{p_2}\left[\sum_{t=1}^{\tau_{\mathscr{S}_0}-1} r_2'(S_t)\right] &\le \mathbf{E}_s^{p_2}\big[\tau_{\mathscr{S}_0}\big]\big\|r_2' - r_1'\big\|_\infty + \mathbf{E}_s^{p_2}\left[\sum_{t=1}^{\tau_{\mathscr{S}_0}-1} \big(\mathbf{e}_{S_t} - p_1(S_t)\big)h_1'\right] \\ &\overset{(\dagger)}{\le} h_1'(s) + \mathbf{E}_s^{p_2}\big[\tau_{\mathscr{S}_0}\big]\Big(\big\|r_2' - r_1'\big\|_\infty + \tfrac{1}{2}\operatorname{sp}(h_1')\big\|p_2' - p_1'\big\|_1\Big) \\ &\overset{(\ddagger)}{\le} h_1'(s) + D(p_2)\big(\|r_2 - r_1\|_\infty + \|g_2 - g_1\|_\infty + 2D(p_1)\|p_2 - p_1\|_1\big) \\ &\overset{(\S)}{\le} h_1'(s) + D(p_2)\big(2\|r_2 - r_1\|_\infty + \big(3 + \tfrac{1}{4}k\big)D(p_1)\|p_2 - p_1\|_1\big). \end{aligned}$$

where (†) follows from Poisson's equation, (‡) bounds bias span and hitting times by diameter and expands the definition of r_j', and (§) invokes the (simplified version of) Lemma III.45 to

bound $\|g_2 - g_1\|_\infty$. With the same computations, we upper-bound $h_2'(s)$ relatively to $h_1'(s)$, leading to the equation:

$$\left|h_2'(s) - h_1'(s)\right| \leq D(p_2)\big(2\|r_2 - r_1\|_\infty + \big(3 + \tfrac{1}{4}k\big)D(p_1)\|p_2 - p_1\|_1\big). \tag{III.130}$$

Combining (III.129) and (III.130), we obtain:

$$\begin{aligned}
(*) &= |h_2(s) - h_1(s)| \\
&\leq \left|h_2'(s) - h_1'(s)\right| + \left|\sum_{i=1}^{k}\big(\mathbf{P}_s^{p_2}\big(\tau_{\mathscr{S}_i} < \infty\big)h_2(s_0^i) - \mathbf{P}_s^{p_1}\big(\tau_{\mathscr{S}_i} < \infty\big)h_1(s_0^i)\big)\right| \\
&\leq \left|h_2'(s) - h_1'(s)\right| + \left|\sum_{i=1}^{k}\mathbf{P}_s^{p_2}\big(\tau_{\mathscr{S}_i} < \infty\big)\big(h_2(s_0^i) - h_1(s_0^i)\big)\right| \\
&\qquad + \left|\sum_{i=1}^{k}\big(\mathbf{P}_s^{p_2}\big(\tau_{\mathscr{S}_i} < \infty\big) - \mathbf{P}_s^{p_1}\big(\tau_{\mathscr{S}_i} < \infty\big)\big)h_1(s_0^i)\right| \\
&\overset{(\dagger)}{\leq} 6D(p_2)\|r_2 - r_1\|_\infty + \big(\big(7 + \tfrac{1}{2}k\big)D(p_2)D(p_1) + \tfrac{1}{2}\mathrm{sp}(h_1^1)\big)\|p_2 - p\|_1 \\
&\overset{(\ddagger)}{\leq} 6D(p_2)\|r_2 - r_1\|_\infty + \big(\big(7 + \tfrac{1}{2}k\big)D(p_2)D(p_1) + 2D(p_1)^2\big)\|p_2 - p\|_1
\end{aligned}$$

where (†) is obtained by bounding the first term with (III.130), the second term with Lemma III.43 and the third term with Lemma III.44; and (‡) follows from $\mathrm{sp}(h_1^1) \leq 4D(p_1)^2$ by applying Lemma III.38 twice. □

10.D.4 Sensitivity of near optimal pairs and Bellman gaps

The **gain gap** Δ_g is given by $\Delta_g(M) := \min\{\|g^*(M) - g_\pi(M)\|_\infty : \pi \notin \Pi^*(M)\}$. The **worst diameter** of a communicating Markov decision process M is the worst diameter of its stationary deterministic policies (see Definition III.8), i.e.,

$$D_*(M) := \max_{\pi \in \Pi} D(p_\pi) < \infty \tag{III.131}$$

The worst diameter provides a simple description of the sensibility of $\mathscr{Z}_{**}^{\epsilon}$ and Δ_*^{ϵ} to model modifications.

Proposition III.48 (Sensitivity of $\mathscr{Z}_{**}^{\epsilon}$)**.** *Let $M \in \mathscr{M}$ and fix $\epsilon \in (0, \frac{1}{2}\Delta_g(M))$. For all M' such that:*

$$\|M' - M\|^* \leq \frac{\frac{1}{2}\epsilon}{1 + \big(1 + \frac{1}{4}|\mathscr{S}|\big)D_*(M)} =: \frac{\epsilon}{C_g(M)}, \tag{III.132}$$

we have $\mathscr{Z}_{**}^{\epsilon}(M') = \mathscr{Z}^{**}(M)$ and $\Pi_\epsilon^*(M') = \Pi^*(M)$.

Proof. Define $C(M) := 1 + (1 + \frac{1}{4}|\mathscr{S}|)D_*(M)$. For every policy $\pi \in \pi$, it follows from Lemma III.45 that:

$$\|g_\pi(M') - g_\pi(M)\|_\infty \leq \|r_\pi' - r_\pi\|_\infty + \big(1 + \tfrac{1}{4}|\mathscr{S}|\big)D(p_\pi)\|p_\pi' - p_\pi\|^* \leq C(M)\|M' - M\|^*.$$

For $\epsilon < \frac{1}{2}\Delta_g(M)$ and $C(M)\|M' - M\|^* \leq \frac{1}{2}\epsilon$, every policy satisfies:

$$\begin{aligned}
\|g_\pi(M') - g^*(M')\|_\infty &\leq \|g_\pi(M') - g_\pi(M)\|_\infty + \|g^*(M') - g^*(M)\|_\infty + \|g_\pi(M) - g^*(M)\|_\infty \\
&\leq \|g_\pi(M) - g^*(M)\|_\infty + \epsilon.
\end{aligned}$$

Therefore, every optimal policy of M is ϵ-gain optimal in M', meaning that $\mathscr{Z}_{**}^{\epsilon}(M') \supseteq \mathscr{Z}^{**}(M)$. Also, with similar computations, we find $\|g_\pi(M') - g^*(M')\|_\infty \geq \|g_\pi(M) - g^*(M)\|_\infty - \epsilon$. So every ϵ-gain optimal policy of M' is 2ϵ-gain optimal in M, and since $2\epsilon < \Delta_g(M)$, 2ϵ-gain optimal policies are gain optimal in M. So $\mathscr{Z}_{**}^{\epsilon}(M') \subseteq \mathscr{Z}^{**}(M)$. □

Proposition III.49 (Sensitivity of Δ^ϵ_*)**.** *Let $M \in \mathscr{M}$ and fix $\epsilon \in (0, \frac{1}{2}\Delta_g(M))$. If $C_g(M)\|M' - M\|^* < \epsilon$, then:*

$$\|h^\epsilon_*(M') - h^*(M)\|_\infty \le (18 + |\mathscr{S}|)D_*(M)^2\|M' - M\|^* =: C_h(M)\|M' - M\|^*; \tag{III.133}$$
$$\|\Delta^\epsilon_*(M') - \Delta^*(M)\|_\infty \le (1 + 2D_*(M) + 2C_g(M) + 2C_h(M))\|M' - M\|. \tag{III.134}$$

More simply, h^ϵ_* and Δ^ϵ_* are locally $\mathrm{O}(D_*(M)^2)$-Lipschitz for $\|{-}\|^*$.

Proof. Since $C_g(M)\|M' - M\|^* < 1$, unfolding the definition of $C_g(M)$ (Proposition III.48) we see that $D(p'_\pi) \le 2D(p_\pi)$ for every policy, see (III.125) and Lemma III.39. By Lemma III.47, for every policy π' we have:

$$\begin{aligned}\left\|h_\pi(M') - h_\pi(M)\right\|_\infty &\le 12D(p_\pi)\|r'_\pi - r_\pi\|_\infty + (18 + |\mathscr{S}|)D(p_\pi)^2\|p'_\pi - p_\pi\|_1^* \\ &\le (18 + |\mathscr{S}|)D_*(M)^2\|M' - M\|^*.\end{aligned} \tag{III.135}$$

Introduce $C_h(M) := (18 + |\mathscr{S}|)D_*(M)^2$. By Proposition III.48, we know that since $C_g(M)\|M' - M\|^* \le \epsilon$, $\Pi^*_\epsilon(M') = \Pi^*(M)$. Pick $\pi \in \Pi^*_\epsilon(M')$. Using (III.135), we find that, for all $s \in \mathscr{S}$,

$$h_\pi(s, M') \le h_\pi(s, M) + C_h(M)\|M' - M\|^* \le h^*(s, M) + C_h(M)\|M' - M\|^*$$

where the second inequality uses that π is gain-optimal in M. So $h^\epsilon_*(s, M') \le h^*(s, M) + C_h(M)\|M' - M\|^*$. Moreover, by picking π has a bias-optimal policy of M, we see that $\pi \in \Pi^*_\epsilon(M')$ and invoking (III.135), we obtain $h^\epsilon_*(s, M') \ge h^*(s, M) - C_h(M)\|M' - M\|^*$. Accordingly,

$$\left\|h^\epsilon_*(M') - h^*(M)\right\|_\infty \le C_h(M)\|M' - M\|^*. \tag{III.136}$$

Therefore,

$$\begin{aligned}(*) &= \|\Delta^\epsilon_*(s, a, M') - \Delta^*(s, a, M)\|_\infty \\ &\le \|g^*(M') - g^*(M)\|_\infty + \|h^\epsilon_*(M') - h^*(M)\|_\infty + \|r' - r\|_\infty + \|p'(s, a)h^\epsilon_*(M') - p(s, a)h^*(M)\|_\infty \\ &\le \|g^*(M') - g^*(M)\|_\infty + 2\|h^\epsilon_*(M') - h^*(M)\|_\infty + \|r' - r\|_\infty + \mathrm{sp}(h^*(M))\|p' - p\|_1 \\ &\le (1 + \mathrm{sp}(h^*(M)) + 2C_g(M) + 2C_h(M))\|M' - M\|^*.\end{aligned}$$

Conclude by bounding $\mathrm{sp}(h^*(M))$ by $2D_*(M)$. □

Part IV

Local Regret Considerations

In this last part, we focus on the local behavior of algorithms. It starts in Chapter 11 with the observation that optimistic methods, such as UCRL2 Auer et al. (2009), tend to play sub-optimal actions for arbitrarily long periods of time infinitely often. These bad periods of play are tracked with exploration times, leading to the design of a new learning metric that called the **regret of exploration**, and show that existing optimistic algorithms have linear regret of exploration, which is the worst possible. In Chapter 12, we address the poor regret of exploration guarantees of optimistic methods by revising how episodes are managed in the first place. We suggest the **performance test** (PT) to reduce the duration of episodes of sub-optimal play. This solution is simplified in Chapter 13 to the **vanishing multiplicative condition** (VM) and we discuss the ideas behind the proof, that are related to the local behavior of confidence region in the asymptotic regime of algorithms. Chapter 14 goes further, by investigating to which extend the management of episodes can be improved. To this end, Chapter 14 focuses on stochastic bandits where episodes are unnecessary and shows that optimistic methods, such as UCB Auer (2002), have infinitely many bursts of suboptimal play despite being episode-less. This property is generalized to a broader class of index algorithms and it is shown that this behavior can only be avoided with a form of randomization. This all is studied using the sliding regret, a new learning metric which is finer than the regret of exploration.

This part is a combination of three papers.

> Boone, V. and Gaujal, B. (2023b). The Regret of Exploration and the Control of Bad Episodes in Reinforcement Learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2824–2856. PMLR
>
> Boone, V. and Gaujal, B. (2024+). Local regret guarantees in average reward markov decision processes. To be submitted
>
> Boone, V. (2023). The Sliding Regret in Stochastic Bandits: Discriminating Index and Randomized Policies. arXiv:2311.18437 [cs, eess, math, stat]

The regret of exploration and the performance test (PT) are due to Boone and Gaujal (2023b), and the more mature analysis and episode rule (VM) to Boone and Gaujal (2024) which is yet to be published. The last Chapter 14 is adapted from my paper Boone (2023).

Chapter 11

Exploration Episodes and the Regret of Exploration

The story begins right where Part II stopped. In Part II, EVI based algorithms where introduced as a robust solution to average reward Markov decision processes. Such algorithms periodically compute an optimistic policy out of a confidence region that they deploy until it becomes obsolete. Because algorithms are supposed to be run, let us run the most classical one: UCRL2 from Auer et al. (2009).

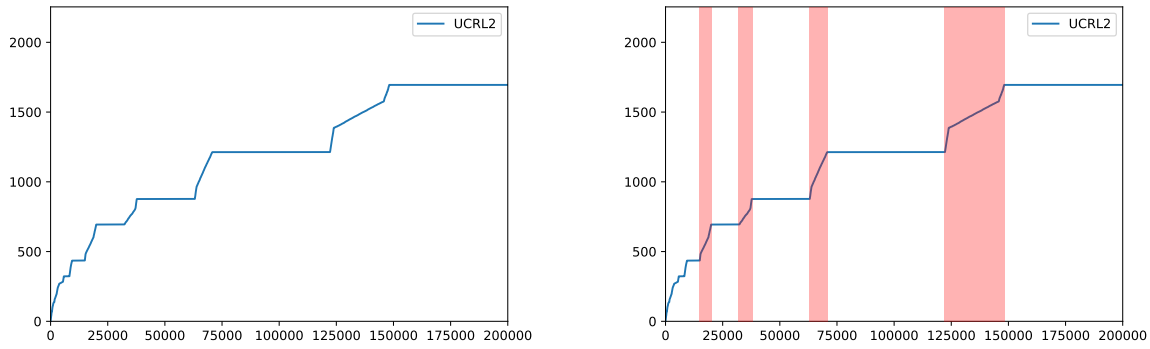


Figure 11.0.1: (On the left) The first order regret of UCRL2 Auer et al. (2009) on a random 10 state 2 action model, one run. (On the right) The same run, with highlighted periods of sub-optimal play.

The run is pictured on Figure 11.0.1 and is typical of UCRL2; We observe a first phase where the first order regret grows linearly and a second phase where the algorithm periodically alternates between periods of optimal play (where the regret is constant) and periods of suboptimal play (where the regret grows). Rather remarkably, even on the simplest model (a two-armed bandit), the periods of sub-optimal play of UCRL2 are increasingly large. In practice, UCRL2 eventually plays sub-optimal actions for arbitrary long period of time. As we will discuss shortly, this behavior is built in by the way episodes are managed.

In this part and the next Chapters 12 and 13, we patch this behavior.

Some will genuinely ask why this behavior should be patched in the first place; Especially if the regret guarantees are already satisfying. The answer is twofold. First, most theoretical guarantees are in strong probability or in expectation when, in many practical scenarios, algorithms are only run once. For consistency purposes, sub-optimal actions must be explored infinitely often yet with UCRL2, this exploration is not sporadically spread during play, but rather

concentrates in burst of suboptimal play that grow uncontrollably. Any service that is run using UCRL2 will display infinitely many periods of time when the quality of service drastically drops that, if rarer and rarer, take longer and longer.[1] Second, this question happens to be fruitful. It leads to the analysis of the **almost-sure regimes** of optimistic algorithms and to the local behavior time-wise behaviors of their confidence regions.

Overall, while Part II was focusing on the way the optimistic policy was computed and on the right choice of confidence region, the present chapter focuses on the **episode rule**, that decides when the policy should be changed. In most of the literature derived from UCRL2 Auer et al. (2009), episodes are updated using the **doubling trick** (DT):

$$t_{k+1} = \inf\big\{t > t_k : \exists a \in \mathscr{A}(S_t), N_t(S_t,a) \geq 1 \vee 2N_{t_k}(S_t,a)\big\}. \tag{DT}$$

In view of (DT), the observations of Figure 11.0.1 appear obvious. By design, the episodes of an algorithm using (DT) are increasingly large. So if the algorithm uses sub-optimal policies infinitely often, these "bad" periods of time grow in size. First, in Section 11.1, we address the increasingly burdensome lack of formalism concerning these "bad periods" and introduce the notion of **exploration times**. Then, in Section 11.3, we introduce a new learning metric that measures the worst regret at exploration times, the **regret of exploration**. Under this formalism, we show that UCRL2 and its variants (Part II) have a linear regret of exploration, then suggest ways to fix the issue by providing alternatives to the doubling trick (DT).

11.1 Exploration episodes and regret of exploration

Because we are now interested in how the regret scales locally, we overload the regret and first order regret notations by taking the initial time into account:

$$\mathrm{Reg}(\tau,\tau') := (\tau'-\tau)g^* - \sum_{t=\tau}^{\tau'-1} R_t \quad \text{and} \quad \mathrm{FOReg}(\tau,\tau') := \sum_{t=\tau}^{\tau'-1} \Delta^*(Z_t) \tag{IV.1}$$

where τ,τ' are stopping times of the natural filtration satisfying $\tau \leq \tau'$. In the whole Part IV, we focus on the study of **episodic algorithms**, generalizing the architecture of Algorithm II.1, see Algorithm IV.1. Such algorithms periodically update a policy π^k that they play for a period of time $\{t_k,\dots,t_{k+1}-1\}$ determined by some episode rule.

Algorithm IV.1 The architecture of episodic algorithms

```
k ← 0, initialize π^0;
for t = 0, 1, ... do
    if current policy π^k is obsolete then
        Update policy π^k;
        k ← k + 1;
        t_k ← t;
    end if
    π_t ← π^k;
    Iterate π_t, i.e., play A_t ∼ π_t(−|S_t), observe reward R_t and transition S_{t+1};
end for
```

Regarding Figure 11.0.1, one attempt to measure the regret endured when the algorithm explores is to measure the regret starting from episodes when the algorithm drops an optimal

[1] Some will ask which services are run with UCRL2. I am not aware of any.

policy for a sub-optimal one, i.e., at times

$$\left\{t_k : \pi^{k-1} \in \Pi^*(M) \text{ and } \pi^k \notin \Pi^*(M)\right\}. \tag{IV.2}$$

While (IV.2) captures the intention, it is not mathematically convenient and lacks many intuitive properties that one wishes it would have. The main concern is that it fails to capture the idea of exploration, and the main reason is that deployed policies can be partially optimal and multi-chain. For instance, the algorithm may drop a globally gain optimal policy for a sub-optimal policy π that is actually gain optimal from the current state, i.e., $g^*(S_t; M) = g^\pi(S_t; M)$ but $g^\pi(M) < g^*(M)$. The converse can also happen, as we want to track what happens when the algorithm drops a policy that was optimal from the current state (but no necessarily globally) for a policy that is sub-optimal from the current state. For these reasons, the final definition of **exploration times** (Definition IV.1) is slightly more complex than (IV.2).

Definition IV.1 (Exploration)**.** *An episode k is an **exploration episode** (and t_k is an **exploration time**) if the two conditions below are satisfied:*

(1) $g^*(M) = g^{\pi^{k-1}}(S_{t_k}; M)$*, i.e., the previously deployed policy was optimal from the current state;*
(2) *There is $z \in \mathrm{Reach}(\pi^k, S_{t_k}) \setminus \mathcal{Z}^*(M)$, i.e., there exists a sub-optimal pair that is reachable from the current state under the current policy.*

The set of exploration episodes is denoted $\mathcal{K}_{\mathrm{exp}}$.

Remark that when the underlying model is recurrent, the exploration times given by Definition IV.1 are equivalent to those defined using (IV.2).

We will commonly pick an enumeration $t_{k(i)}$ of $\mathcal{K}_{\mathrm{exp}}$ where $k(i)$ denotes the i-the exploration episode and $t_{k(i)}$ is the associated i-th initial exploration time. Formally, $t_{k(1)} := \inf \mathcal{K}_{\mathrm{exp}}$ and $t_{k(i+1)} := \inf\{t_k > t_{k(i)} : k \in \mathcal{K}_{\mathrm{exp}}\}$. We then define the **regret of exploration** as the asymptotically worst regret at exploration times.

Definition IV.2 (Regret of exploration)**.** *Let $(t_{k(i)})$ the enumeration of exploration episodes. The **regret of exploration** is given by:*

$$\mathrm{RegExp}(T) \equiv \mathrm{RegExp}(T; M) := \limsup_{i \to \infty} \mathbf{E}^M\left[\mathrm{Reg}(t_{k(i)}; t_{k(i)} + T)\right]. \tag{IV.3}$$

The regret of exploration is well-defined only if there are infinitely many exploration times $t_{k(i)}$, and this is not always guaranteed. There actually exist models where exploration is somehow unnecessary, making them conceptually easier to learn than bandits. As a matter of fact, these models are precisely those such that the regret lower bound (Part III) satisfies $K(M; \mathcal{M}) = 0$. Such models are discussed in Section 11.2. When $K(M; \mathcal{M}) = 0$, there is no guarantee that the number of exploration episodes is infinite in general. The regret of exploration may be ill-behaved when episodes are too short, because the performance observed during a short episode may be decorrelated from the actual performance of the deployed policy. In Chapter 12, we explain why optimistic algorithms cannot afford too many episodes. So, such cases will be cast out by assuming that there are sub-linearly many episodes, i.e., that $|\mathcal{K}(T)| = o(T)$.

11.2 Explorative Markov decision processes

When the enumeration of exploration episodes $(t_{k(i)})$ is almost surely finite, the regret of exploration is the limsup of a finite sequence of non-negative quantities, which is arguably zero. When this is the case, the regret of exploration is not an interesting metric. In this section, we look at when this can be the case.

11.2.1 A Markov decision process where UCRL2 has constant regret

There exist models that can be learned within a finite exploration phase. This is the case when the model dependent regret lower bound satisfies $K(M;\mathscr{M}) = 0$, for which Theorem III.14 essentially guarantees that regret rates of $o(\log(T))$ can be achieved. For such models, the sub-optimality of sub-optimal actions can be checked only by visiting optimal pairs $\mathscr{Z}^{**}(M)$. In fact, for such models, we can find consistent learners achieving regret $O(1)$. Such learners are specifically designed optimistic algorithms with finitely many exploration episodes. To the best of our knowledge, there is no mention of such cases in the literature.

We discuss the example displayed in Figure 11.2.1.

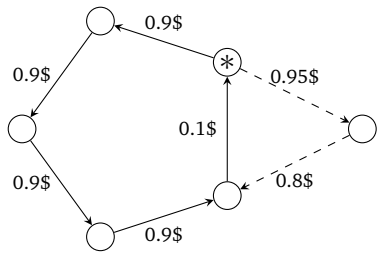


Figure 11.2.1: An example of model where $\mathrm{Cnf}(M) = \varnothing$ and where the regret of exploration may be null. There is a single choice of action in all states except at the marked state $(*)$ where there are two actions (dashed and solid lines). Under any state, a choice of action deterministically leads to the state indicated by the arrow. Rewards are Bernoulli, with means indicated by labels.

Consider $\mathscr{M}$ the class of models induced by M of Figure 11.2.1 by allowing for different reward parameters (still Bernoulli) and with the same transition kernel. We consider running UCRL2 Auer et al. (2009) on $M \equiv (r, p, \mathscr{Z})$ while giving it the transition kernel p as prior knowledge — making the algorithm very similar to the UCYCLE algorithm of Ortner (2010)). While it is known that UCRL2 has $O(\log(T))$ regret in general, it is $O(1)$ on the model of Figure 11.2.1.

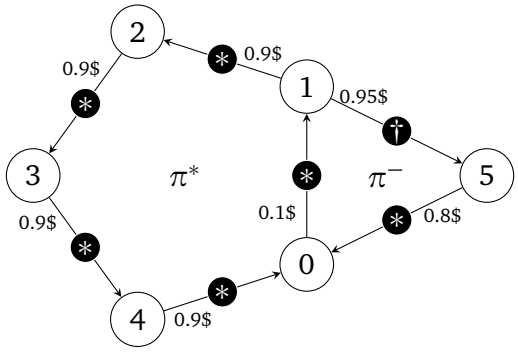
Proposition IV.1. *If* UCRL2 *is given the transition kernel of the model M of Figure 11.2.1, it achieves* $O(1)$ *expected regret.*

The idea of the proof is actually simple. On M, there are two possible policies, either looping on the 5-cycle or the 3-cycle. But by looping on the 5-cycle, the algorithm learns its rewards very well, hence can claim that the 3-cycle's average reward is upper bounded by $\frac{1+1+0.1+\varepsilon_t}{3}$ because unknown rewards are bounded by 1. This is smaller than a lower bound for the 5-cycle $\frac{0.9+0.9+0.9+0.9+0.1-\varepsilon_t}{5}$ (where ε_t is vanishing with t). Therefore, the algorithm has no need to visit the dashed arrows infinitely often. Interestingly, this example is robust to reward perturbations, meaning that it is **non-degenerate** in a sense that we make precise below (Definition IV.3).

We consider the version of UCRL2 adapted to models with deterministic (known) transitions and Bernoulli (unknown) rewards, so that the confidence region $\mathscr{M}_t$ on the model is

identified to the confidence region $\mathscr{R}_t$ on reward parameters (the parameters of the Bernoulli distributions). The algorithm given below relies on Hoeffding-type confidence bounds on rewards. The call $\texttt{EVI}(\mathscr{R},\epsilon)$ returns a policy achieving ϵ-optimal gain on $\mathscr{R}$, see Algorithm II.3. We write $g^\pi(s;\mathscr{R}) := \sup_{\tilde{M}\in\mathscr{R}} g^\pi(s;\tilde{M})$ and $g^*(s;\mathscr{R}) := \sup_\pi g^\pi(s;\mathscr{R})$ the optimistic gain (see Definition II.2).

Proof of Proposition IV.1. The fact that UCRL2 is consistent on M is a well-known result. Indeed, it is known that its regret is $O(\sqrt{SAT\log(T)})$, see Auer et al. (2009) and Ortner (2010). We focus on proving that it has bounded regret on the model M given in Figure 11.2.1. We will work with a labeled version of Figure 11.2.2, see Figure 11.2.2.



Algorithm IV.2 UCYCLE: UCRL2 for deterministic transition models

$$\mathscr{R}_t := \prod_{z\in\mathscr{Z}} \left\{ \tilde{r}(z)\in[0,1] : \tilde{r}(z) \le \hat{r}_t(z) + \sqrt{\tfrac{2\log(SAt)}{N_t(z)}} \right\}$$

1: $k \leftarrow 0$, initialize π^0;
2: **for** $t = 0,1,\dots$ **do**
3: **if** (DT) triggers **then**
4: $u^k \leftarrow \texttt{EVI}(\mathscr{R}_t,\epsilon_t,0^{\mathscr{S}})$;
5: $\pi^k \leftarrow$ any π s.t. $\mathscr{L}_t(u^k) = \mathscr{L}_t^\pi(u^k)$;
6: $k \leftarrow k+1$; $t_k \leftarrow t$.
7: **end if**
8: Set $\pi_t \leftarrow \pi^k$ and iterate π_t.
9: **end for**

Figure 11.2.2: A Markov decision process on which a version UCRL2 specialized to deterministic transition models (known as UCYCLE, see Ortner (2010)) has bounded regret.

The model M is accordingly identified with its reward vector r. Remark that the only pair with positive Bellman gap is $(1,\dagger)$ with Bellman gap $\Delta^*(1,\dagger) \le 1$. The regret is therefore upper-bounded by $|\{t\le T : \pi_t = \pi^-\}|$ and we are left to control the number of times the suboptimal policy π^- is being played. A simple property induced by the doubling trick (DT) is that $t_{k+1} \le 3t_k$. Hence, if $\pi_t = \pi^-$, then there exists $t' \in [\frac{1}{3}t, t]$ such that π^- is the result of EVI, i.e., $g^{\pi^-}(\mathscr{R}_{t'}) > g^{\pi^*}(\mathscr{R}_{t'}) + \frac{1}{t'}$.

Let $c := 3\cdot\frac{0.9+0.9+0.9+0.9+0.1}{5} - 2 = 0.22$, which is the threshold on the reward that one should have on $(0,*)$ in order to make π^- better than π^*. Since

$$g^{\pi^-}(\mathscr{R}_t) \le \frac{1}{3}\left(2 + \hat{r}_t(0,*) + \sqrt{\frac{2\log(SAt)}{N_t(0,*)}}\right) \quad\text{and}\quad g^{\pi^*}(\mathscr{R}_t) \ge \mathbf{1}(r\in\mathscr{R}_t) g^{\pi^*}(r),$$

we have:

$$\begin{aligned}
(*) &:= \mathbf{E}\left[\left|t\ge 1 : \pi_t \ne \pi^-\right|\right] \\
&\le 300 + \sum_{t\ge 300}\sum_{t'=t/3}^{t} \mathbf{P}\left(g^{\pi^-}(\mathscr{R}_{t'}) > g^{\pi^*}(\mathscr{R}_{t'}) + \tfrac{1}{100}\right) \\
&\le 300 + \sum_{t\ge 300}\sum_{t'=t/3}^{t}\left(\mathbf{P}\left(\hat{r}_{t'}(0,*) + \sqrt{\frac{2\log(SAt')}{N_{t'}(0,*)}} > 0.21\right) + \mathbf{P}\left(r\notin\mathscr{R}(t')\right)\right). \qquad \text{(IV.4)}
\end{aligned}$$

For the first term, remark that $N_{t'}(0,*) \geq \frac{1}{5}t'$ almost surely when $t' \geq 5$. For t' large enough so that $\sqrt{10\log(SAt')/t'} < 0.01$, we have

$$\begin{aligned}(**) &:= \mathbf{P}\left(\hat{r}_{t'}(0,*) + \sqrt{\frac{2\log(SAt')}{N_{t'}(0,*)}} > 0.21\right)\\ &\leq \mathbf{P}\left(\exists n \in [\tfrac{1}{5}t', t'] : N_{t'}(0,*) = n, \hat{r}_{t'}(0,*) + \sqrt{\tfrac{2\log(SAt')}{n}} > 0.21\right)\\ &\leq \sum_{n=\frac{1}{5}t'}^{\infty} \mathbf{P}(N_{t'}(0,*) = n, \hat{r}_{t'}(0,*) - r(0,*) > 0.2)\\ &\leq \sum_{n=\frac{1}{5}t'}^{\infty} \exp\left(-\frac{8}{10000}n\right) = \frac{\exp\left(-\frac{1}{6250}t'\right)}{1-\exp\left(-\frac{1}{1250}\right)} = \mathrm{O}\left(\exp(-\tfrac{1}{6250}t')\right).\end{aligned}$$

For the second term, we have

$$\begin{aligned}\mathbf{P}\left(r \notin \mathscr{R}(t')\right) &= \mathbf{P}\left(\exists z, |\hat{r}_{t'}(z) - r_z| > \sqrt{\frac{2\log(SAt')}{N_{t'}(z)}}\right)\\ &\leq \sum_z \sum_{n=1}^{\infty} \mathbf{P}\left(N_{t'}(z) = n, |\hat{r}_{t'}(z) - r(z)| > \sqrt{\frac{2\log(SAt')}{n}}\right)\\ &\leq 2SA \sum_{n=1}^{\infty} \exp\left(-4\log(SAt') \cdot n\right)\\ &\leq \frac{2SA}{(t'SA)^4} \cdot \frac{1}{1-(t'SA)^{-4}} \leq \frac{4}{(SA)^3 t'^4} = \mathrm{O}\left(t'^{-4}\right).\end{aligned}$$

Overall, injecting it all in (IV.4), we obtain $\mathbf{E}|t \geq 1 : \pi_t \neq \pi^-| < \infty$. Accordingly, $\mathbf{E}[\mathrm{Reg}(T)] = \mathrm{O}(1)$ on M. □

11.2.2 Explorative sub-spaces and exploration times

Proposition IV.1 suggests that given a class $\mathscr{M}$ of models, there may be consistent algorithms such that $\mathbf{E}^M[\mathrm{Reg}(T)] = \mathrm{O}(1)$ over large sub-spaces of $\mathscr{M}$. Such algorithms use sub-optimal policies finitely often, hence $\mathscr{K}_{\mathrm{exp}}$ is almost surely finite and there are finitely many exploration episodes. This motivates the following definitions.

Definition IV.3 (Non degeneracy). *A model $M \equiv (r,p)$ is said **non-degenerate** if there exists $\epsilon > 0$ such that, for all $r' \in \mathbf{R}^{\mathscr{Z}}$ with $\|r' - r\|_\infty < \epsilon$, the model $M' \equiv (r',p)$ satisfies $\mathscr{Z}^*(M') = \mathscr{Z}^*(M)$. In other words, if $\mathscr{Z}^*$ is locally robust to reward perturbations at M.*

Definition IV.4 (Explorative models). *Given a space of Markov decision processes $\mathscr{M}$, its **explorative sub-space** $\mathscr{M}^+$ is the set of non-degenerate models $M \in \mathscr{M}$ such that every algorithm (1) with sublinearly many episodes and (2) which is consistent on $\mathscr{M}$, has infinitely many exploration episodes almost surely.*

Non degeneracy goes back to Boone and Gaujal (2023a) and corresponds to models for which bias optimality can be correctly identified with arbitrary level of confidence. Theorem IV.2 below provides a complete characterization of exploration sub-spaces, linking them to the confusing

set. The main take-away from this result is that the regret of exploration is interesting to study whenever there is something to learn that cannot be learned with bounded regret.

Theorem IV.2. *Provided that M is non-degenerate, the following are assertions are equivalent.*

(1) $M \notin \mathscr{M}^+$;
(2) $\mathrm{Cnf}(M;\mathscr{M}^+) = \varnothing$;
(3) $K(M;\mathscr{M}) = 0$.

The equivalence between (2) and (3) follows by definition of the $K(M)$, see Theorem III.5. What we are interested in is (1) $\Rightarrow$ (2) or, equivalently, the converse $(2)^c \Rightarrow (1)^c$, namely that if the set of confusing model is non empty, then every consistent algorithm with sublinearly many episodes has infinitely many exploration episodes. The result is established under a technical assumption that is hard-coded in the definition of explorative sub-spaces, stating that there are sub-linearly many episodes. This condition is arguably mild, because most of state-of-the-art algorithms are made so that the number episodes grows sub-linearly (in fact polylogarithmically, if possible). The inverse implication (2) $\Rightarrow$ (1) is of second order importance and won't be shown in these pages, see Boone and Gaujal (2024) for a proof.

Proposition IV.3. *Let $M \in \mathscr{M}$ a non-degenerate model with $\mathrm{Cnf}(M) \neq \varnothing$. Every episodic algorithm which is (1) consistent and (2) with sub-linearly many episodes (in expectation) has infinitely many exploration episodes with probability one.*

This result is surprisingly difficult because of its generality. Recall that $k \geq 1$ is an **exploration episode** if (1) $g^*(M) = g(\pi^k, S_{t_k}, M)$ and (2) $\mathrm{Reach}(\pi^k, S_{t_k}, M) \cap \mathscr{Z}^-(M) \neq \varnothing$, see (Definition IV.1). In order to show that there are infinitely many exploration episodes, we have to show that the learning process alternates infinitely often between periods of times when the played policy is gain optimal, and others when there is a reachable sub-optimal pair. (**STEP 1**) is a preliminary technical fact. In (**STEP 2**), we show with (IV.8) that the process spends infinitely many times on the recurrent part of a gain-optimal policy. In (**STEP 2**), we show with (IV.9) that the process must play sub-optimal pairs infinitely often. Combining both in (**STEP 4**), we show that the number of exploration times is infinite, and each are finite with probability one.

Notation. If $\pi \in \Pi$, we write $\mathrm{Rec}(\pi)$ the set of states that are recurrent under π on M.

(**STEP 1**) *For all model $M \in \mathscr{M}$, there exists a constant $C(M) > 0$ such that whatever the driving mechanism, we have:*

$$\mathbf{E}^M\left[\sum_{t=1}^{T}(g^*(S_t,M) - g^{\pi_t}(S_t,M) + \mathbf{1}(S_t \notin \mathrm{Rec}(\pi_t)))\right] \leq \mathbf{E}^M[\mathrm{Reg}(T)] + C(M)\mathbf{E}^M[|\mathscr{K}(T)|]. \quad \text{(IV.5)}$$

Proof. In the proof below, we drop the dependency in M in the notations. If $\pi \in \Pi$, we denote $\mathrm{Rec}(\pi)$ the recurrent states of π in M. We have:

$$\begin{aligned}(*) &= \mathbf{E}[\mathrm{Reg}(T)] \\ &= \mathbf{E}\left[\sum_{t=1}^{T}\Delta^*(Z_t)\right] \\ &\overset{(\dagger)}{=} \mathbf{E}\left[\sum_{t=1}^{T}\big(g^*(S_t) - r(Z_t) + \big(e_{S_t} - p(Z_t)\big)h^*\big)\right]\end{aligned}$$

$$\geq \mathbf{E}\left[\sum_{t=1}^{T}(g^*-r(Z_t))\right]-\mathrm{sp}(h^*)$$
$$\overset{(\ddagger)}{\geq} \underbrace{\mathbf{E}\left[\sum_{k=1}^{|\mathscr{K}(T)|}\sum_{t=t_k}^{t_{k+1}-1}\mathbf{1}(S_t\in\mathrm{Rec}(\pi_t))(g^*(S_t)-r(Z_t))\right]}_{\mathrm{A}}-\underbrace{\mathbf{E}\left[\sum_{k=1}^{|\mathscr{K}(T)|}\sum_{t=t_k}^{t_{k+1}-1}\mathbf{1}(S_t\notin\mathrm{Rec}(\pi_t))\right]}_{\mathrm{B}}-\mathrm{sp}(h^*)$$

where (†) uses the Bellman equation $h^*(s)+g^*(s)=r(s,a)+p(s,a)h^*+\Delta^*(s,a)$, and (‡) uses that $g^*(S_t)-r(Z_t)\geq -1$. We bound A and B separately. Let $D_*:=\sup_\pi\sup_s\mathbf{E}_s^\pi[\inf\{t\geq 1:S_t\in\mathrm{Rec}(\pi)\}]<\infty$ the worst hitting time to a recurrent class in M. We have:

$$\mathrm{B}=\mathbf{E}\left[\sum_{k=1}^{|\mathscr{K}(T)|}\sum_{t=t_k}^{t_{k+1}-1}\inf\left\{t>t_k:S_t\in\mathrm{Rec}(\pi^k)\right\}\right]\leq D_*\mathbf{E}[|\mathscr{K}(T)|]. \tag{IV.6}$$

Meanwhile, introduce $t_k':=t_{k+1}\wedge\inf\{t>t_k:S_t\in\mathrm{Rec}(\pi^k)\}$ and $H:=\sup_\pi\mathrm{sp}(h^\pi)<\infty$ the worst bias span. We have:

$$\begin{aligned}
\mathrm{A}&=\mathbf{E}\left[\sum_{k=1}^{|\mathscr{K}(T)|}\sum_{t=t_k'}^{t_{k+1}-1}(g^*(S_t)-r(Z_t))\right]\\
&\overset{(\dagger)}{=}\mathbf{E}\left[\sum_{k=1}^{|\mathscr{K}(T)|}\sum_{t=t_k'}^{t_{k+1}-1}\left(g^*(S_t)-g^{\pi^k}(S_t)+\left(p(Z_t)-e_{S_t}\right)h^{\pi^k}\right)\right]\\
&\geq\mathbf{E}\left[\sum_{k=1}^{|\mathscr{K}(T)|}\sum_{t=t_k'}^{t_{k+1}-1}(g^*(S_t)-g^{\pi_t}(S_t))\right]-H\mathbf{E}[|\mathscr{K}(T)|]\\
&\overset{(\ddagger)}{\geq}\mathbf{E}\left[\sum_{t=1}^{T}(g^*(S_t)-g^{\pi_t}(S_t))\right]-H\mathbf{E}[|\mathscr{K}(T)|]
\end{aligned}\tag{IV.7}$$

where (†) uses the Poisson equation $h^{\pi^k}(s)+g^{\pi^k}(s)=r(s,\pi^k(s))+p(s,\pi^k(s))h^{\pi^k}$ and (‡) that $g^*(S_t)\geq g^{\pi_t}(S_t)$ for all $t\geq 1$. Combining (IV.6) and (IV.7), we get:

$$\mathbf{E}\left[\sum_{t=1}^{T}(g^*(S_t)-g^{\pi_t}(S_t))\right]+\mathbf{E}\left[\sum_{t=1}^{T}\mathbf{1}(S_t\notin\mathrm{Rec}(\pi_t))\right]\leq\mathrm{Reg}(T)+(2D_*+H)\mathbf{E}[|\mathscr{K}(T)|].$$

Conclude the proof by setting $C:=2D_*+H<\infty$. □

(**STEP 2**) *Assume that the algorithm is consistent and has sublinearly many episodes in expectation. Then:*

$$\mathbf{P}(\forall T,\exists t\geq T:g^*(S_t,M)=g^{\pi_t}(S_t,M)\text{ and }S_t\in\mathrm{Rec}(\pi_t))=1. \tag{IV.8}$$

Proof. Assume on the contrary that $\mathbf{P}(\forall T,\exists t\geq T:g^*(S_t,M)=g^{\pi_t}(S_t,M)\wedge S_t\in\mathrm{Rec}(\pi_t))=1-\delta$ with $\delta>0$. Accordingly, there exists $T_0\geq 1$ such that:

$$\mathbf{P}\Big(\forall t\geq T_0,g^*_{S_t}(M)>g^\pi_t(S_t,M)\text{ or }S_t\notin\mathrm{Rec}(\pi_t)\Big)\geq\tfrac{1}{2}\delta.$$

Let $\Delta_g:=\min\{g^*(s,M)-g^\pi(s,M):\pi\in\Pi,s\in\mathscr{S},g^*(s,M)>g^\pi(s,M)\}$ the gain-gap of M. We have $\Delta_g\in(0,1]$ and thus:

$$(*):=\mathbf{E}\left[\sum_{t=1}^{T}(g^*(S_t,M)-g^{\pi_t}(S_t,M))\right]+\mathbf{E}\left[\sum_{t=1}^{T}\mathbf{1}(S_t\notin\mathrm{Rec}(\pi_t))\right]$$

$$\begin{aligned}
&\geq \Delta_g \mathbf{E}\left[\sum_{t=1}^{T} \mathbf{1}(g^*(S_t,M) > g^{\pi_t}(S_t,M) \text{ or } S_t \notin \mathrm{Rec}(\pi_t))\right] \\
&\geq \Delta_g (T-T_0)\mathbf{P}(\forall t \geq T_0, g^*(S_t,M) > g^{\pi_t}(S_t,M) \text{ or } S_t \notin \mathrm{Rec}(\pi_t)) \\
&\geq \tfrac{1}{2}\Delta_g \delta (T-T_0) = \Omega(T).
\end{aligned}$$

Meanwhile, we know that $\mathbf{E}[\mathrm{Reg}(T)] = o(T)$ and $\mathbf{E}[|\mathscr{K}(T)|] = o(T)$, so that by (**STEP 1**) (IV.5), we also have $(*) = o(T)$, a contradiction. □

(**STEP 3**) *If* $\mathrm{Cnf}(M) \neq \emptyset$*, then every consistent algorithm satisfies*

$$\mathbf{P}^M(\forall T, \exists t > T : \Delta^*(Z_t) > 0) = 1. \tag{IV.9}$$

Proof. On the contrary, assume that $\mathbf{P}^M(\forall T, \exists t > T : \Delta^*(Z_t) > 0) = 1-\delta$ with $\delta > 0$. Accordingly, there exists $m \geq 1$ such that:

$$\tfrac{1}{2}\delta \leq \mathbf{P}^M(\forall t > m : \Delta^*(Z_t) = 0) \leq \mathbf{P}^M\left(\forall t \geq 1 : \sum_{z \in \mathscr{Z}^-(M)} N_t(z) \leq m\right). \tag{IV.10}$$

We show that $z \in \mathscr{Z}^-(M)$ can be changed to $z \notin \mathscr{Z}^{**}(M)$ in (IV.10), see (IV.11). To see this, introduce the reward function $f(z) := \mathbf{1}(z \in \mathscr{Z}^{**}(M))$ and let g^f, h^f and Δ^f the respective gain, bias and gap functions of the optimal policy π^* of M (defined by $\pi^*(s) = a$ the unique $a \in \mathscr{A}(s)$ such that $(s,a) \in \mathscr{Z}^*(M)$) under reward function f and kernel $p(M)$. Remark that $g^f(s) = 1$ for all $s \in \mathscr{S}$ and that, by construction of π^*, $\Delta^f(z) = 0$ for all $z \in \mathscr{Z}^*(M)$. Denote $H^f := \mathrm{sp}(h^f) \vee \max_z |\Delta^f(z)|$. We have:

$$\begin{aligned}
\sum_{z \in \mathscr{Z}^{**}(M)} N_z(T) &= \sum_{t=1}^{T} f(Z_t) \\
&\overset{(\dagger)}{=} \sum_{t=1}^{T} \big(1 + \big(e_{S_t} - p(Z_t)\big)h^f - \Delta^f(Z_t)\big) \\
&\geq T - H^f - \sum_{t=1}^{T} \Delta^f(Z_t) + \sum_{t=1}^{T} \big(e_{S_{t+1}} - p(Z_t)\big)h^f \\
&\overset{(\ddagger)}{\geq} T - H^f - H^f \sum_{t=1}^{T} \mathbf{1}(Z_t \notin \mathscr{Z}^*(M)) + \sum_{t=1}^{T} \mathbf{1}(Z_t \notin \mathscr{Z}^{**}(M))\big(e_{S_{t+1}} - p(Z_t)\big)h^f
\end{aligned}$$

where (†) uses the Bellman equation $1 + h^f(s) = f(s,a) + p^f(s,a)h^f + \Delta^f(s,a)$, and (‡) that $h^f(s) = 0$ for all $(s, \pi^*(s)) \in \mathscr{Z}^{**}(M)$. For $\mathscr{Z}' \subseteq \mathscr{Z}$, denote $N_T(\mathscr{Z}') := \sum_{z \in \mathscr{Z}'} N_T(z)$. The first sum is equal to $\sum_{t=1}^{T} \mathbf{1}(Z_t \notin \mathscr{Z}^*(M)) = N_T(\mathscr{Z}^-(M))$. The RHS of the above equation is bounded using a time-uniform Azuma-Hoeffding inequality (Lemma I.22), showing that:

$$\mathbf{P}\left(\exists T \geq 1, \sum_{t=1}^{T} \mathbf{1}(Z_t \notin \mathscr{Z}^{**}(M))\big(e_{S_{t+1}} - p(Z_t)\big)h^f < -H^f \sqrt{N_T(\mathscr{Z}^{**}(M)^c) \log\Big(\tfrac{4N_{\mathscr{Z}^{**}(M)^c}(T)}{\delta}\Big)}\right) \leq \tfrac{1}{4}\delta$$

Using that $N_T(\mathscr{Z}^{**}(M)^c) = T - N_T(\mathscr{Z}^{**}(M))$, we obtain that, with probability at least $\frac{1}{4}\delta$, for all $T \geq 1$, we have:

$$\begin{aligned}
T - N_T(\mathscr{Z}^{**}(M)^c) &\geq T - H^f\big(1 + N_T(\mathscr{Z}^-(M))\big) - H^f \sqrt{N_T(\mathscr{Z}^{**}(M)^c) \log\Big(\tfrac{4N_T(\mathscr{Z}^{**}(M)^c)}{\delta}\Big)} \\
&\geq T - H^f(1+m) - H^f \sqrt{N_T(\mathscr{Z}^{**}(M)^c) \log\Big(\tfrac{4N_T(\mathscr{Z}^{**}(M)^c)}{\delta}\Big)}.
\end{aligned}$$

Rearranging terms, we get that with probability at least $\frac{1}{4}\delta$, for all $T \geq 1$, we have:

$$N_T(\mathscr{Z}^{**}(M)^c) \leq H^f\left(1 + m + \sqrt{N_T(\mathscr{Z}^{**}(M)^c)\log(N_T(\mathscr{Z}^{**}(M)^c))} + \sqrt{N_T(\mathscr{Z}^{**}(M)^c)\log\left(\tfrac{4}{\delta}\right)}\right).$$

Denoting $n := N_T(\mathscr{Z}^{**}(M))$, we have an equation of the form $n \leq \alpha + \beta\sqrt{n\log(n)} + \gamma\sqrt{n}$. For $n \geq 3$, $n\log(n) \geq n$ hence we can simplify the upper-bound to $n \leq \alpha + (\beta+\gamma)\sqrt{n\log(n)}$. Dividing by $\log(n) \geq 1$ and setting $m := n/\log(n)$, we get $m \leq \alpha + (\beta+\gamma)\sqrt{m}$, and simple algebra leads to:

$$\frac{n}{\log(n)} = m \leq 2\big(\alpha + (\beta+\gamma)^2\big).$$

Further using $\log(n) \leq \sqrt{n}$, we get $n \leq 4(\alpha+(\beta+\gamma)^2)^2$. We conclude that there exists a constant m' such that:

$$\mathbf{P}^M\left(\forall t \geq 1, \sum_{z \notin \mathscr{Z}^{**}(M)} N_t(z) \leq m'\right) \geq \tfrac{1}{4}\delta. \tag{IV.11}$$

Now that (IV.11) is established, we finally derive a contradiction by relying on a change of measure argument. Let $M^\dagger \in \mathrm{Cnf}(M)$, which is non-empty by assumption. For short, the transition kernels and reward distributions of M (respectively $M^\dagger$) are denoted p and r (respectively $p^\dagger$ and $r^\dagger$). We introduce the log-likelihood-ratio of observations $H_t := (S_t, A_t, R_1, \ldots, A_{t-1}, R_{t-1}, S_t)$ as:

$$L(t) \equiv L(H_t) := \sum_{s,a}\sum_{i<t-1} \mathbf{1}(S_i = s, A_t = a)\log\left(\frac{q_{s,a}(R_i)}{q^\dagger_{s,a}(R_i)}\frac{p_{s,a}(S_{i+1})}{p^\dagger_{s,a}(S_{i+1})}\right).$$

It is known since Marjani et al. (2021) that if $\mathscr{E}$ is a $\sigma(H_t)$-measurable event, then $\mathbf{P}^{M^\dagger}(\mathscr{E}) = \mathbf{E}^M[\mathbf{1}(\mathscr{E})\exp(-L(t))]$. Since $M \ll M^\dagger$, there exists a constant $c > 0$ such that, for all $z \in \mathscr{Z}$, we have $\log[(r_z(\alpha)/r^\dagger_z(\alpha)) \cdot (p_z(s')/p^\dagger_z(s'))] \leq \log(c)$ with the convention $0/0 = 0$. For $z \in \mathscr{Z}^{**}(M)$, the LHS logarithm is null. Therefore, we have:

$$\begin{aligned}
\mathbf{P}^{M^\dagger}\left(\sum_{z\notin\mathscr{Z}^{**}(M)} N_t(z) \leq m'\right) &= \mathbf{E}^M\left[\mathbf{1}\Big(\sum_{z\notin\mathscr{Z}^{**}(M)} N_t(z) \leq m'\Big)\exp(-L(t))\right] \\
&\geq \mathbf{E}^M\left[\mathbf{1}\Big(\sum_{z\notin\mathscr{Z}^{**}(M)} N_t(z) \leq m'\Big)\exp\left(-\sum_{z\notin\mathscr{Z}^{**}(M)} N_t(z)\log(c)\right)\right] \\
&\geq c^{-m'}\mathbf{P}^M\left(\sum_{z\notin\mathscr{Z}^{**}(M)} N_t(z) \leq m'\right) \geq c^{-m'}\delta := \delta' > 0.
\end{aligned}$$

Accordingly, the algorithm has probability at least δ' to spend at most m' visits outside $\mathscr{Z}^{**}(M)$ when running on $M^\dagger$. This will be in contradiction $M^\dagger \in \mathrm{Cnf}(M)$ and the consistency of the algorithm. Indeed, since $M^\dagger \gg M$ coincides with M on $\mathscr{Z}^{**}(M)$, we see that the optimal policy π^* of M has unique recurrent class $\mathscr{Z}^{**}(M)$ in $M^\dagger$. Yet, $\pi^* \notin \Pi^*(M^\dagger)$, hence $\mathscr{Z}^{**}(M) \cap \mathscr{Z}^{-}(M^\dagger) \neq \varnothing$, i.e., there exists $z \in \mathscr{Z}^{**}(M)$ such that $\Delta^*(z; M^\dagger) > 0$. We further link the number of visits of this z to the total number of visits of $\mathscr{Z}^{**}(M)$ with the same technique that the one used to convert (IV.10) to (IV.11).

Introduce the reward function $f(z') := \mathbf{1}(z' = z)$, and let g^f, h^f, Δ^f the gain, bias and gaps functions of the policy π^* in $M^\dagger$. There exists $\epsilon > 0$ such that $g^f(s) = \epsilon$ for all $s \in \mathscr{S}$. Letting $C := \mathrm{sp}\big(h^f\big) \vee \max_{z'}\big|\Delta^f(z')\big| < \infty$. For all $T \geq 1$, we have

$$N_T(z) = \sum_{t=1}^T f(Z_t) = \sum_{t=1}^T\big(\epsilon + \big(e_{S_t} - p(Z_t)\big)h^f - \Delta^f(Z_t)\big)$$

$$\geq T\epsilon - C - CN_{\mathscr{Z}^{**}(M)^c}(T) + \sum_{t=1}^{T}\big(e_{S_{t+1}} - p(Z_t)\big)h^f$$
$$\overset{(\dagger)}{\geq} T\epsilon - C(1+m') - C\sqrt{T\log\big(\tfrac{2T}{\delta'}\big)} \sim T\epsilon.$$

where (†) holds with probability $\frac{1}{2}\delta' > 0$ uniformly for $T \geq 1$, by invoking a time-uniform Azuma-Hoeffding (Lemma I.22) to lower-bound the right-hand martingale. We accordingly obtain, when $T \to \infty$,

$$\mathbf{E}^{M^\dagger}[\mathrm{Reg}(T)] \gtrsim \tfrac{1}{2}\epsilon\delta'\Delta^*(z;M^\dagger)T = \Omega(T). \tag{IV.12}$$

So (IV.12) is in contradiction with the consistency of the algorithm. □

(**STEP 4**) *If the algorithm is consistent, has sub-linearly many episodes, then for all $M \in \mathscr{M}$ such that* $\mathrm{Cnf}(M) \neq \varnothing$*, we have:*

$$\mathbf{P}^M\big(\forall T, \exists t \geq T : g^*(M) = g^{\pi_{t-1}}(S_{t-1}, M) \text{ and } \mathrm{Reach}(\pi_t, S_t, M) \cap \mathscr{Z}^{-}(M) \neq \varnothing\big) = 1. \tag{IV.13}$$

Moreover, the stopping times t enumerating the time-instants such that $g^*(S_{t-1}, M) = g^{\pi_{t-1}}(S_{t-1}, M)$ and $\mathrm{Reach}(\pi_t, S_t, M) \cap \mathscr{Z}^{-}(M) \neq \varnothing$ are exploration times; Hence there are infinitely many of them with probability one.

Proof. This is obtained by combining (IV.8) of (**STEP 2**) and (IV.9) of (**STEP 3**). We have:

$$\mathbf{P}^M(\forall T, \exists t \geq T : g^*(S_{t-1}, M) = g^{\pi_t}(S_t, M) \text{ and } S_t \in \mathrm{Rec}(\pi_t)) = 1, \text{ and} \tag{IV.14}$$
$$\mathbf{P}^M\big(\forall T, \exists t \geq T : \mathrm{Reach}(\pi_t, S_t, M) \cap \mathscr{Z}^{-}(M) \neq \varnothing\big) = 1. \tag{IV.15}$$

By non-degeneracy of M, if $g^*(S_t, M) = g^{\pi_t}(S_t, M)$ and $S_t \in \mathrm{Rec}(\pi_t)$, then $\mathrm{Reach}(\pi_t, S_t, M) = \mathscr{Z}^{**}(M)$ which is disjoint from $\mathscr{Z}^{-}(M)$. Define:

$$\tau_1 := \inf\{t \geq 1 : g^*(S_t, M) = g^{\pi_t}(S_t, M) \text{ and } S_t \in \mathrm{Rec}(\pi_t)\},$$
$$\tau_{2i} := \inf\big\{t > \tau_{2i-1} : \mathrm{Reach}(\pi_t, S_t, M) \cap \mathscr{Z}^{-}(M) \neq \varnothing\big\},$$
$$\tau_{2i+1} := \inf\{t > \tau_{2i} : g^*(S_t, M) = g^{\pi_t}(S_t, M) \text{ and } S_t \in \mathrm{Rec}(\pi_t)\}$$

Then (τ_i) is an increasing sequence of stopping times, and by (IV.14) (IV.15) applied in tandem, we show by induction that $\mathbf{P}^M(\tau_i < \infty) = 1$ for all $i \geq 1$. By non-degeneracy of M, at $t = \tau_{2i+1}$, the current policy is gain-optimal and the process is currently on the optimal class $\mathscr{Z}^{**}(M)$. Because $\mathscr{Z}^{**}(M)$ is the disjoint union of sink components of $\mathscr{Z}^{*}(M)$, hence the only way to exit $\mathscr{Z}^{**}(M)$ is by playing a $z \in \mathscr{Z}^{-}(M)$. Therefore, we see that for $t = \tau_{2i}$, we must have $g^*(S_{t-1}, M) = g^{\pi_{t-1}}(S_{t-1}, M)$ with $\pi_{t-1} \neq \pi_t$. Accordingly, every τ_{2i} are change of episodes that are exploration episodes. □

This proves Proposition IV.3. ■

The last part of Theorem IV.2, that (2) ⇒ (1), is deferred to the appendix. The main take-away of this paragraph is the result below.

Corollary IV.4. *If $M \in \mathscr{M}$ satisfies* $\mathrm{Cnf}(M) \neq \varnothing$*, then every consistent episodic algorithm with sub-linearly many episodes have well-defined regret of exploration.*

While Corollary IV.4 is enough to properly analyze the regret of exploration of optimistic algorithms, I am still concerned about how difficult it is to satisfyingly define what is exploration. For instance, the definition of exploration times requires the algorithm to be episodic and to

maintain an internal policy that is used to navigate the environment. Without this internal policy, it is harder to define what exploration is. Actually, even if the algorithm maintains an internal policy, its performance may be decorrelated from how good this policy is, because the internal policy may be too volatile. Asking for the number of episodes to grow sub-linearly is a simple way to circumvent the issue, yet, I am not very satisfied with this solution. While many existing algorithms fit the episodic framework described with Algorithm IV.1, episodes are more of a matter of design than a matter of analysis or behavior, and the assumption on the sub-linearity of the number of episodes is only to make sure that the two are related. In the end, algorithms play actions, not policies. For instance, the algorithm ECoE (Algorithm III.2) does not fit into the framework of episodic algorithms like Algorithm IV.1. Perhaps, the most natural approach to exploration may be to claim that an exploration time should be a time instant when the played pair is the first to be sub-optimal after a long time period of optimal play. We immediately see that there is no canonical choice for the duration of the time period of optimal play. One may choose it to be 0 and claim that any time instant when a sub-optimal pair is played is an exploration time. Doing so reveals a troubling problem: The regret of exploration of UCB Auer (2002) is then linear, although UCB is episode-less and is perhaps the algorithm that emboddies optimism the best. This is why, in the definition of exploration times, we ask for the previous policy to be optimal. This guarantees that the starting episode is "**well conditionned**". Removing this conditionning leads to a finer notion that I call the **sliding regret**, and this is the subject of the last Chapter 14. This raises a new question, which is how poorly conditionned we may choose exploration times, because the condition (1) of Definition IV.1 may be stronger than required. While this direction seems interesting, it appeared disproportionally technical to address a phenomenon that was never investigated in the first place. I have made the choice of simplicity with restrictions over the difficulty of a completely general approach. Nonetheless, this interesting direction can be outlined with the following questions:

> *How should the trajectory of play be decomposed? Can exploitation and exploration be trajectorially isolated and distinguished? How can we describe and classify the various shapes of first order regret curves? Can we quantify how sub-optimal play is spread during a run?*

We barely scratch the surface of these questions in Chapter 14. In the next few chapters, we focus on the regret of exploration introduced by Definition IV.2.

11.3 The regret of exploration and the doubling trick

With Figure 11.0.1, we have observed that the regret at exploration times of UCRL2 seems to increase overall. This follows from a much more general principle that is quite intuitive: If a change of episode requires an increase of visit relatively to the initial visit count, and if deployed policies do not play actions with vanishing probabilities (see (IV.16)), then the regret of exploration grows linearly on recurrent models at least.

Theorem IV.5. *Fix a pair space $\mathscr{Z}$ and let $\mathscr{M}$ be the space of all recurrent models with pairs $\mathscr{Z}$. Let $f : \mathbf{N} \to (0,\infty)$ such that $\lim f(n) = +\infty$. Any $\mathscr{M}$-consistent episodic learner (π_t) satisfying:*

$$\begin{gathered}\forall k \geq 1, \exists z \in \mathscr{Z}, \quad N_{t_{k+1}}(z) \geq N_{t_k}(z) + f\big(N_{t_k}(z)\big) \\ \exists c > 0, \forall t \geq 0, \forall (s,a) \in \mathscr{Z}, \quad \pi_t(a|z) \geq c \text{ or } \pi_t(a|z) = 0\end{gathered} \tag{IV.16}$$

has linear regret of exploration on the explorative sub-space of $\mathscr{M}$, i.e., for all $M \in \mathscr{M}^+$, we have $\mathrm{RegExp}(T) = \Omega(T)$ a.s. when $T \to \infty$.

Proof. Let $M \in \mathscr{M}^+$. By Theorem IV.2, $\left|\mathscr{K}_{\mathrm{exp}}\right| = \infty$ almost surely. Denote $(t_{k(i)})$ the enumeration of exploration times. Because M is recurrent, every policy is recurrent on M thus $\mathrm{Reach}(\pi, M, s) \cap \mathscr{Z}^-(M) \neq \varnothing$ if, and only if $g^\pi(M) < g^*(M)$, where s is an arbitrary state. From (IV.16), we see that:

$$\mathbf{P}\Big(\lim_{t\to\infty} \min\{N_t(s,a) : \pi_t(a|s) > 0\} = \infty\Big) = 1. \tag{IV.17}$$

It follows that $\liminf(t_{k(i)+1} - t_{k(i)}) = \infty$. In particular, for all $T \geq 0$, we have:

$$\begin{aligned}
\mathrm{RegExp}(T) &\overset{(*)}{\geq} \limsup_{i\to\infty} \left(\mathbf{E}^{(\pi_t),M}\left[\sum_{t=t_{k(i)}}^{t_{k(i)}+T-1} \Delta^*(Z_t; M) \right] - \mathrm{sp}(h^*(M)) \right) \\
&\overset{(\dagger)}{\geq} \limsup_{i\to\infty} \Big(\mathbf{E}^{(\pi_t),M}\Big[T \min(\mu^{\pi_{t_{k(i)}}}(M)) \Delta^*_{\min}(M) - \mathrm{sp}(D(\pi_{t_{k(i)}}; M)) \Big] - h^*(M) \Big) \\
&\overset{(\ddagger)}{\geq} T\alpha - \beta
\end{aligned}$$

where $(*)$ follows from Proposition I.11; $(\dagger)$ is obtained by writing the Poisson equation of $\pi_{t_{k(i)}}$ for the reward function $f_i(z) = \mathbf{1}(z = z_i)$ where z_i is any sub-optimal pair played by $\pi_{t_{k(i)}}$; and $(\ddagger)$ introduces $\alpha := \min_\pi \min(\mu^\pi(M)) \Delta^*_{\min}(M) > 0$ and $\beta = \mathrm{sp}(h^*(M)) + \max_\pi D(\pi; M) < \infty$. □

This applies to UCRL2 and more generally to all algorithms relying on the doubling trick (DT) to manage episodes, where one can pick $f(n) = n \vee 1$. This includes UCRL2 Auer et al. (2009), REGAL Bartlett and Tewari (2009), KLUCRL Filippi et al. (2010), UCRL2B Fruit et al. (2020), SCAL Fruit et al. (2018), UCRL3 Bourel et al. (2020), EBF Zhang and Ji (2019) and also PMEVI (see Chapter 7) (up to mild modifications of (IV.16) for a few of them). I conjecture that Theorem IV.5 can be generalized beyond the recurrent setting, but a generalization of the proof will encounter many technical challenges that are not especially interesting nor informative to address, especially to cover weird exotic algorithms that do not exist.

Theorem IV.5 points out a disease: The local regret of the current optimistic algorithms is ill-behaved, because the regret of exploration of these methods grows linearly. Like every desease, we will attempt at curing it without too many side-effects. So, is it possible to alter these algorithms and have sub-linear regret of exploration without hurting the minimax regret guarantees? The answer is positive. This is achieved by changing the episode rule and is the subject of the next chapters.

Chapter 12

Managing Episodes with the Performance Test (PT)

In Chapter 11 with Theorem IV.5, the doubling trick (DT) has been shown to make the regret of exploration grow linearly. This is much of a surprise since the doubling trick essentially makes episodes double in size over time, and the phenomenon appears strikingly on experiments (Figure 11.0.1). In this chapter, we present the Performance Test (PT) to manage episodes and enjoy sublinear regret of exploration.

12.1 Managing episodes solely with optimism

The philosophy behind the performance test is purity: If the algorithm's design is centered around optimism, then optimism should solely drive the algorithm, and the doubling trick is only but a trick that is completely artificial. After all, with the doubling trick, an episode ends regardless of the data acquired during the episode and ends only because enough data have been gathered. Instead, an optimistic learner should play a policy because it is an optimistic policy, and only drop it because the policy is not optimistically optimal anymore. Caution is required however, because in opposition to multi-armed bandits where episodes are not necessary (UCB Auer (2002) is not episodic for instance), episode-less optimistic algorithms may endure linear regret for Markov decision processes. By episode-less algorithms, we mean that $t_{k+1} := t_k + 1$, hence that the policy is updated at every time-step. The necessity of episodes is discussed by Ortner (2010) for UCYCLE, a version of UCRL2 specialized to deterministic transition models.

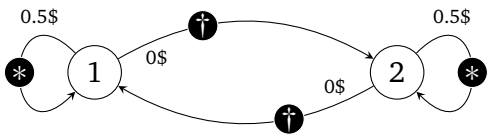


Figure 12.1.1: A degenerate model where optimistic algorithms must manage episodes carefully.

The problem arises when the model is degenerate (Definition IV.3), for instance with the model described by Figure 12.1.1. On this model, any gain optimal policy must either cycle with $(1,*)$ or with $(2,*)$ because the Bellman gaps of the transitioning pairs $(1,\dagger)$ and $(2,\dagger)$ are positive. In fact, $\mathscr{Z}^{**}(M) = \{(1,*),(2,*)\}$ has two communicating components. In practice, UCRL2 will hesitate between whether $(1,*)$ is better than $(2,*)$ or not. Ortner (2010) shows

that the episode-less version of UCRL2 will travel from 1 to 2 and 2 to 1 linearly often, hence its regret grows linearly.

Rather than changing episode as soon as the current policy is not optimistically optimal anymore, we suggest to change it when it not optimistically optimal **enough**:

$$t_{k+1} := \inf\{t > t_k : g^{\pi_{t-1}}(\mathcal{M}_t) + f(T)e \leq g^*(\mathcal{M}_t)\} \tag{PT}$$

where $f : \mathbf{N} \to \mathbf{R}_+$ is a non-increasing vanishing function that quantifies how lazy the updates should be and $g(\pi, \mathcal{M}_t)$ and $g^*(\mathcal{M}_t)$ are the policy-wise and global optimistic gain vectors (see Definition II.2). If f is sufficiently large, the back-and-forth behavior observed by Ortner (2010) disappears, as established by Theorem IV.6. We call this new rule the **performance test**. Given a standard optimistic algorithm (such as UCRL2), we use the suffix "-PT" to indicate that the episode rule is enriched with (PT).

There are two ideas behind the performance test. First, a sub-optimal policy cannot be played for too long because its optimistic value should drop quickly; Hence (PT) should offer better regret of exploration guarantees than (DT). Second, the sub-optimality tolerance $f(t)$ should be chosen correctly so that the number of episodes under (PT) remains under control and no back-and-forth behavior (Ortner (2010)) is possible.

12.2 Guarantees of the performance test

The performance test accordingly promises that (1) the minimax regret guarantees remain the same and (2) the regret of exploration guarantees are sub-linear. The sine qua non requirements are regret guarantees and are treated first in Section 12.2.2. Regret of exploration guarantees are investigated in Section 12.2.3.

For reference, the pseudo-code of UCRL2-PT is given with Algorithm IV.3. The confidence region is chosen as a product of ℓ_1-balls of radius $\sqrt{\xi_t(z)}$ that quantifies the optimistic bonus and is of form $\sqrt{S\log(Ct)/N_t(z)}$ by Weissman's inequality (Lemma I.23). The precise value of $C > 0$ doesn't matter much in the analysis, provided that the confidence region holds with high probability.

Algorithm IV.3 UCRL2-PT

$$\mathcal{M}_t := \left\{\tilde{r} : \forall z \in \mathscr{Z}, |\tilde{r}(z) - \hat{r}_t(z)|^2 \leq \sqrt{\tfrac{\log(Ct)}{N_t(z)}}\right\} \times \left\{\tilde{p} : \forall z \in \mathscr{Z}, \|\tilde{p}(z) - \hat{p}_t(z)\|_1^2 \leq \xi_t(z)\right\}$$

$k \leftarrow 0$, initialize π^k;
for $t = 0, 1, \ldots$ **do**
  if (DT) or (PT) **then**
    $u^k \leftarrow \mathrm{EVI}(\mathcal{M}_t, 0, 0^{\mathscr{S}})$;
    $\pi^k \leftarrow$ any π such that $\mathscr{L}_t(u^k) = \mathscr{L}_t^{\pi}(u^k)$
    $k \leftarrow k+1$; $t_k \leftarrow t$;
  end if
  Set $\pi_t \leftarrow \pi^k$ and iterate π_t;
end for

Remark that the architecture of Algorithm IV.3 is closer to EVI-based algorithm (Algorithm II.2) than it is to UCRL2 alone. The pseudo-code of Algorithm IV.3 is straightfully adapted to any EVI-based algorithm, providing KLUCRL-PT, UCRL2B-PT etc. However, the analysis is specific to UCRL2 although it probably could be generalized.

12.2.1 Minimax regret of UCRL2-PT

The regret analysis of UCRL2-PT is exactly the same than the one of UCRL2, and fit the optimistic framework exposed in Section 6.4. Specifically, we obtain

$$\mathrm{Reg}(T) \leq D(M)|\mathscr{K}(T)| + \mathrm{O}\Big(\sqrt{DSAT\log\big(\tfrac{T}{\delta}\big)}\Big) \tag{IV.1}$$

provided the confidence region holds universally in time, i.e., that $M \in \bigcap_{t=0}^{T-1} \mathscr{M}_t$, usually called "the good event". [1] Picking a time adaptive confidence level $\delta(t) := \frac{1}{t}$, one can also make sure that (IV.1) holds in expectation for all $T \geq 0$. The number of episodes $|\mathscr{K}(T)|$ of UCRL2-PT are left to be upper-bounded. This is done in the next section.

12.2.2 Number of episodes under (PT)

The number of episodes of UCRL2-PT are directly related to the slackness function $f(T)$. If $f(T)$ is large, (PT) does not trigger often and (DT) manages most of episodes. If $f(T)$ is small, the number of episodes is subjected to be much higher. Theorem IV.6 provides a bound on the number of episodes of UCRL2-PT. A notable difference from (DT) is that the bound holds with high probability rather than with probability one.

Theorem IV.6. *Introduce the* ***good event*** $\mathscr{E}(T)$ *as follows. Let* $\mathscr{E}(t,t';z) := (\|\hat{p}_{t:t'}(z) - p\|_1^2 \leq 4(N_{t'}(z) - N_t(z))\log\big(4SAt'^3/\delta\big))$ *and set* $\mathscr{E}(T) := \bigcap_{t=0}^{T-1}\bigcap_{t'=t}^{T-1}\bigcap_{z\in\mathscr{Z}}\mathscr{E}(t,t';z)$, *where* $\hat{p}_{t:t'}(z)$ *is the empirically observed kernel at* z *from time* t *to* $t'-1$. *On the good event* $\mathscr{E}(T)$, *the number of episodes of* UCRL2-PT *up to time* T *is upper-bounded by:*

$$|\mathscr{K}(T)| \leq \frac{2^4 D\sqrt{S\log(T)}}{f(T)} + \mathrm{O}\left(DS^{\frac{3}{2}}A\sqrt{\frac{\log\big(\frac{SAT^3}{\delta}\big)\log^2(T)}{f(T)}}\right). \tag{IV.2}$$

Sketch of proof. The doubling trick accounts only for logarithmically many episodes which is negligible in front of the number of other episodes. We thus ignore episodes interrupted by (DT). The fact that episode k ends at time t_{k+1} implies that

$$g^{\pi^k}(\mathscr{M}_{t_{k+1}}) + f(t_{k+1}) \leq g^{\pi^{k+1}}(\mathscr{M}_{t_{k+1}}) \tag{IV.3}$$

Because π^k is optimistically optimal at time t_k, it means that over $\{t_k,\dots,t_{k+1}\}$, either $g(\pi^k,\mathscr{M}_t)$ or $g(\pi^{k+1},\mathscr{M}_t)$ has varied by about $f(t_{k+1})$. However, we know that the gain is D-Lipschitz by Theorem II.1. Following this, we show that if $\pi \in \{\pi^k,\pi^{k+1}\}$, then

$$\big\|g^{\pi}(\mathscr{M}_{t_{k+1}}) - g^{\pi}(\mathscr{M}_{t_k})\big\|_\infty \leq D\big(\big\|\hat{P}_{t_{k+1}} - \hat{P}_{t_k}\big\|_1 + \big\|\xi_{t_{k+1}} - \xi_{t_k}\big\|_\infty\big)$$

Therefore, from (IV.3), we deduce that, over $\{t_k,\dots,t_{k+1}\}$, there must be a variation of (1) empirical kernels $\hat{P}_t$ or (2) optimistic bonuses ξ_t of order at least $D^{-1}\sqrt{\alpha\log(t_{k+1})/t_{k+1}}$. On the good event, these variations can be related to variations of time (i.e., $t_{k+1}-t_k$) and visit counts

[1] This event is too strong actually., To be precise and properly show that the algorithm is consistent, we should pick $\bigcap_{t=\sqrt{T}}^{T-1}\mathscr{M}_t$ instead, so that the probability that the good event holds is an increasing function of T. This is important to prove that UCRL2-PT is consistent, but we ignore this subtlety for simplicity.

(i.e., $N_{t_{k+1}}(z) - N_{t_k}(z)$). We then derive a collection of inequalities that guarantees that, when there is a change of episode, visit counts or time increase enough relatively to $f(t_{k+1})$, hence relatively to $f(T)$. The inequality that later accounts for the dominant part in the number of episodes is the following:

$$\frac{f(t_{k+1})}{2^4 D} \leq \sqrt{\frac{S}{N_{t_k}(z)}}\Big(\sqrt{\log(Ct_{k+1})} - \sqrt{\log(Ct_k)}\Big). \tag{IV.4}$$

This inequality quickly leads to the regret bound on $|\mathscr{K}(T)|$. The most technical part is to bound the second order term.

Formal proof of Theorem IV.6. For simplicity, **we assume that the rewards are known**, i.e., that $\mathscr{R}_t(z) = \{r(z)\}$ for all $z \in \mathscr{Z}$ and $t \geq 0$. The goal is to show that if there is a change of episode, then necessarily, transition kernels or confidence bound must have moved by some tractable quantity. Introduce the set of **standard** episodes $\mathscr{K}_0$

$$\mathscr{K}_0 := \{k \in \mathscr{K} : \forall z \in \mathscr{Z}, N_{t_{k+1}}(z) < 2N_{t_k}(z)\}. \tag{IV.5}$$

Therefore, an episode k is non-standard if either (1) there is $z \in \mathscr{Z}$ visited on $\{t_k, \dots, t_{k+1} - 1\}$ such that $N_{t_k}(z) = 0$ or (2) there is $z \in \mathscr{Z}$ such that $N_{t_k}(z) \geq 1$ and $N_{t_{k+1}}(z) \geq 2N_{t_k}(z)$. Accordingly, non-standard episodes are exactly those interrupted by the doubling trick (DT). On non-standard episodes, there must be a pair that doubles its visit count, so $\mathscr{K} \setminus \mathscr{K}_0$ is of logarithmic cardinal. This will be negligible in front of bounds on the cardinal of $\mathscr{K}_0$.

(**STEP 1**) *On the good event $\mathscr{E}(T)$, for all $k \in \mathscr{K}_0$, we have*

$$\frac{1}{2} f(T) \leq D\Big(\big\|\hat{p}_{t_k} - \hat{p}_{t_{k+1}}\big\|_1 + \big\|\xi_{t_k} - \xi_{t_{k+1}}\big\|_\infty\Big) \tag{IV.6}$$

Proof. By definition of $k \in \mathscr{K}_0$, we have:

$$g^{\pi^k}(\mathscr{M}_{t_{k+1}}) + f(t_{k+1}) \leq g^*(\mathscr{M}_{t_{k+1}}) = g^{\pi^{k+1}}(\mathscr{M}_{t_{k+1}}).$$

Further using that $g^{\pi^k}(\mathscr{M}_{t_k}) = g^*(\mathscr{M}_{t_k})$ and writing $\mathrm{A} := \Big[g^{\pi^k}(\mathscr{M}_{t_{k+1}}) - g^{\pi^k}(\mathscr{M}_{t_k})\Big]$ and $\mathrm{B} := \Big[g^{\pi^{k+1}}(\mathscr{M}_{t_{k+1}}) - g^{\pi^{k+1}}(\mathscr{M}_{t_k})\Big]$, this is equivalent to

$$\mathrm{A} - \mathrm{B} \leq g^{\pi^{k+1}}(\mathscr{M}_{t_k}) - g^{\pi^k}(\mathscr{M}_{t_k}) - f(t_k) \overset{(*)}{\leq} -f(T)$$

where $(*)$ follows from $g^{\pi^{k+1}}(\mathscr{M}_{t_k}) \leq g^{\pi^k}(\mathscr{M}_{t_k})$ and $f(t_k) \geq f(T)$. Accordingly, when the episode changes, either $\mathrm{A} \leq -\frac{1}{2} f(T)$ or $\mathrm{B} \geq \frac{1}{2} f(T)$. When the first inequality holds, we say that the episode is **type I** and when the second holds, the episode is said **type II**. Introduce the Hausdorff distance on subsets of $\mathscr{M}$

$$d_{\text{Hausdorff}}(\mathscr{M}_1, \mathscr{M}_2) := \max\left\{\sup_{p_1 \in \mathscr{M}_1} \inf_{p_2 \in \mathscr{M}_2} \|p_2 - p_1\|_1, \sup_{p_2 \in \mathscr{M}_2} \inf_{p_1 \in \mathscr{M}_1} \|p_1 - p_2\|_1\right\}. \tag{IV.7}$$

Recall that $\mathscr{M}_t^\pi := \{r\} \times \prod_{s \in \mathscr{S}} \sum_{a \in \mathscr{A}(s)} \mathscr{P}_t(s,a)\pi(a|s)$ is the confidence region associated to a policy. The policies deployed by UCRL2-PT are deterministic, so we focus on $\pi \in \Pi^{\mathrm{SD}}$. One checks that:

$$\begin{aligned} d_{\text{Hausdorff}}\Big(\tilde{\mathscr{M}}_{t_k}^\pi, \tilde{\mathscr{M}}_{t_{k+1}}^\pi\Big) &\leq \Big\|\hat{p}_{t_k}^\pi - \hat{p}_{t_{k+1}}^\pi\Big\|_1 + \Big\|\xi_{t_k}^\pi - \xi_{t_{k+1}}^\pi\Big\|_\infty \\ &= \sup_{z \in \pi}\big\|\hat{p}_{t_k}(z) - \hat{p}_{t_{k+1}}(z)\big\|_1 + \sup_{z \in \pi}\big\|\xi_{t_k}(z) - \xi_{t_{k+1}}(z)\big\|_\infty. \end{aligned}$$

Write $\mathrm{Proj}_{\mathscr{U}}(\cdot)$ a projection on $\mathscr{U} \subseteq \mathscr{M}$ for $\|-\|_1$, i.e., $\mathrm{Proj}_{\mathscr{U}}(p_1)$ is any $p_2 \in \mathscr{U}$ minimizing the ℓ_1-distance to p_1. In particular, for all $p_1 \in \mathscr{V}$, we have $\|p_1 - \mathrm{Proj}_{\mathscr{U}}(p_1)\|_1 \le d_{\mathrm{Hausdorff}}(\mathscr{U}, \mathscr{V})$. So, if the episode k is type I, we have $-\mathrm{A} \ge \frac{1}{2} f(T)$ thus:

$$\begin{aligned}
\tfrac{1}{2} f(T) &\le g^{\pi^k}(\mathscr{M}_{t_k}) - g^{\pi^k}(\mathscr{M}_{t_{k+1}}) \overset{(*)}{=} g(r, \tilde{p}^{\pi^k}_{t_k}) - g(r, \tilde{p}^{\pi^k}_{t_{k+1}}) \\
&\overset{(\dagger)}{\le} g(r, \tilde{p}^{\pi^k}_{t_k}) - g(r, \mathrm{Proj}_{\mathscr{M}^{\pi^k}_{t_{k+1}}}(\tilde{p}^{\pi^k}_{t_k})) \\
&\overset{(\ddagger)}{\le} \mathrm{sp}(h(r, \tilde{p}^{\pi^k}_{t_k})) \left(\sup_{z \in \pi^k} \left\| \hat{p}_{t_k}(z) - \hat{p}_{t_{k+1}}(z) \right\|_1 + \sup_{z \in \pi^k} \left\| \xi_{t_k}(z) - \xi_{t_{k+1}}(z) \right\|_\infty \right) \\
&\overset{(\S)}{\le} D\, \mathrm{sp}(r) \left(\left\| \hat{p}_{t_k} - \hat{p}_{t_{k+1}} \right\|_1 + \left\| \xi_{t_k} - \xi_{t_{k+1}} \right\|_\infty \right)
\end{aligned}$$

where $(*)$ introduces the optimistic models (Corollary II.12) of π^k at respective times t_k and t_{k+1}; $(\dagger)$ uses that $\tilde{p}^{\pi^k}_{t_{k+1}}$ is an optimistic model of π^k at time t_{k+1}; $(\ddagger)$ invokes the gain deviation inequality (Theorem II.1); and $(\S)$ uses that π^k is an output of `EVI` at time t_k, so that $\mathrm{sp}(h(r, \tilde{p}^{\pi^k}_{t_k})) = \mathrm{sp}(h^*(\mathscr{M}_{t_k})) \le D(\mathscr{M}_{t_k}) \le D(M)$ on the good event, see Proposition II.2. Type II episodes are handled with a similar computations, showing that

$$\tfrac{1}{2} f(T) \le D\, \mathrm{sp}(r) \left(\left\| \hat{p}_{t_k} - \hat{p}_{t_{k+1}} \right\|_1 + \left\| \xi_{t_k} - \xi_{t_{k+1}} \right\|_\infty \right) \tag{IV.8}$$

on the good event as well. So, on the good event, (IV.8) holds for all $k \in \mathscr{K}_0$. □

(**STEP 2**) *On the good event $\mathscr{E}(T)$, for all $z \in \mathscr{Z}$, we have*

$$\left\| \hat{p}_{t_{k+1}}(z) - \hat{p}_{t_k}(z) \right\| \le 4 \sqrt{S \log\left(\tfrac{4SAT^3}{\delta}\right)} \frac{\sqrt{N_{t_{k+1}}(z) - N_{t_k}(z)}}{N_{t_k}(z)} \tag{IV.9}$$

Proof. Pick $z \in \mathscr{Z}$. For short, denote $n = N_{t_k}(z)$ and $m = N_{t_{k+1}}(z) - N_{t_k}(z)$. Because $k \in \mathscr{K}_0$ is not interrupted by the doubling trick, we know that $m \le n$. Denote $W_t(z) := N_t(z) \hat{p}_t(z)$ the aggregate empirical distribution of the transition z. Then a straight forward computations shows that

$$\begin{aligned}
\hat{p}_{t_{k+1}}(z) - \hat{p}_{t_k}(z) &= \frac{1}{n+m} \left(W_{t_k}(z) + \sum\nolimits_{t=t_k}^{t_{k+1}-1} \mathbf{1}(Z_t = z) e_{S_{t+1}} \right) - \frac{1}{n} W_{t_k}(z) \\
&= \frac{1}{n+m} \left(-\frac{m}{n} W_{t_k}(z) + \sum\nolimits_{t=t_k}^{t_{k+1}-1} \mathbf{1}(Z_t = z) e_{S_{t+1}} \right)
\end{aligned}$$

Now, on the good event $\mathscr{E}(T)$,

$$\begin{aligned}
\left\| W_{t_k}(z) - N_{t_k}(z) P(z) \right\|_1 &\le \sqrt{4nS \log\left(\tfrac{4SAT^3}{\delta}\right)} \\
\left\| \sum\nolimits_{t=t_k}^{t_{k+1}-1} \mathbf{1}(Z_t = z) e_{S_{t+1}} - mP(z) \right\|_1 &\le \sqrt{4mS \log\left(\tfrac{4SAT^3}{\delta}\right)}.
\end{aligned} \tag{IV.10}$$

Combined, we obtain that on $\mathscr{E}(T)$,

$$\begin{aligned}
\left\| \hat{p}_{t_{k+1}}(z) - \hat{p}_{t_k}(z) \right\|_1 &\le \frac{1}{n+m} \left(\frac{m}{n} \sqrt{4nS \log\left(\tfrac{4SAT^3}{\delta}\right)} + \sqrt{2mS \log\left(\tfrac{4SAT^3}{\delta}\right)} \right) \\
&\le \sqrt{4S \log\left(\tfrac{4SAT^3}{\delta}\right)} \frac{\sqrt{m}\left(1 + \sqrt{m/n}\right)}{n+m} \\
&\le 4 \sqrt{S \log\left(\tfrac{4SAT^3}{\delta}\right)} \frac{\sqrt{m}}{n}
\end{aligned}$$

where the last inequality is obtained using $m \le n$. We conclude accordingly. □

(**STEP 3**) *For all $z \in \mathscr{Z}$,*

$$\left|\xi_{t_{k+1}}(z) - \xi_{t_k}(z)\right| \leq \frac{\sqrt{S\log(CT)}\left(N_{t_{k+1}}(z) - N_{t_k}(z)\right)}{\max\{1, N_{t_k}(z)^{3/2}\}} + \frac{(t_{k+1}-t_k)\sqrt{S}}{2t_k} \tag{IV.11}$$

Proof. Fix $z \in \mathscr{Z}$. Recall that $\xi_t(z) = \sqrt{S\log(Ct)/N_t(z)}$. We expand $\xi_{t_{k+1}}(z) - \xi_{t_k}(z)$ as:

$$\underbrace{\sqrt{S\log(Ct_{k+1})}\left(\frac{1}{\sqrt{N_{t_{k+1}}(z)}} - \frac{1}{\sqrt{N_{t_k}(z)}}\right)}_{\text{term A}} + \underbrace{\sqrt{\frac{S}{N_{t_k}(z)}}\left(\sqrt{\log(Ct_{k+1})} - \sqrt{\log(Ct_k)}\right)}_{\text{term B}}$$

Since $\left|\sqrt{1/(n+m)} - \sqrt{1/n}\right| \leq m/n^{3/2}$, term A is bounded as

$$\left|\sqrt{S\log(Ct_{k+1})}\left(\frac{1}{\sqrt{N_{t_{k+1}}(z)}} - \frac{1}{\sqrt{N_{t_k}(z)}}\right)\right| \leq \frac{\sqrt{S\log(CT)}\left(N_{t_{k+1}}(z) - N_{t_k}(z)\right)}{N_{t_k}(z)^{3/2}} \tag{IV.12}$$

Term B is left untouched. □

(**STEP 4**) *On the good event $\mathscr{E}(T)$, for all $k \in \mathscr{K}_0$, one of the following holds:*

$$\exists z \in \mathscr{Z}, \quad N_{t_{k+1}}(z) \geq \max\left\{N_{t_k}(z)\left(1 + \frac{f(T)N_{t_k}(z)}{2^{10}D^2 S\log\left(\frac{4SAT^3}{\delta}\right)}\right), 1\right\} \tag{type A}$$

$$\exists z \in \mathscr{Z}, \quad N_{t_{k+1}}(z) \geq \max\left\{N_{t_k}(z)\left(1 + \sqrt{\frac{f(T)N_{t_k}(z)}{2^8 D^2 S\log(CT)}}\right), 1\right\} \tag{type B}$$

$$\frac{f(T)}{2^4 D} \leq \sqrt{\frac{S}{N_{t_k}(z)}}\left(\sqrt{\log(Ct_{k+1})} - \sqrt{\log(Ct_k)}\right) \tag{type C}$$

Proof. By using the explicit variations of empirical kernels (IV.9) and of bonuses (IV.11) in (IV.6), we see that on the good event $\mathscr{E}(T)$, for all $k \in \mathscr{K}_0$, there must be a $z \in \mathscr{Z}$ such that one of the following holds:

$$\frac{f(T)}{2^4 D} \leq 4\sqrt{S\log\left(\frac{4SAT^3}{\delta}\right)}\frac{\sqrt{N_{t_{k+1}}(z) - N_{t_k}(z)}}{\max\{1, N_{t_k}(z)\}} \tag{type A}$$

$$\frac{f(T)}{2^4 D} \leq \sqrt{S\log(CT)}\frac{N_{t_{k+1}}(z) - N_{t_k}(z)}{\max\{1, N_{t_k}(z)^{3/2}\}} \tag{type B}$$

$$\frac{f(T)}{2^4 D} \leq \sqrt{\frac{S}{N_{t_k}(z)}}\left(\sqrt{\log(Ct_{k+1})} - \sqrt{\log(Ct_k)}\right). \tag{type C}$$

There are at most SA episodes such that the z achieving one of the conditions above has never been visited yet, i.e., such that $N_{t_k}(z) = 0$. Such episodes belong to $\mathscr{K} \setminus \mathscr{K}_0$, so can be ignored by assumption. Therefore, we can change $\max\{1, N_{t_k}(z)^\lambda\}$ to the simpler $N_{t_k}(z)^\lambda$. The main statement is obtaining by solving the equations in $N_{t_{k+1}}(z)$, $N_{t_{k+1}}(z)$ and t_{k+1} respectively. □

(**STEP 5**) *On the good event $\mathscr{E}(T)$, the number of episodes is bounded by:*

$$|\mathscr{K}(T)| \leq \frac{2^4 D\sqrt{S\log(T)}}{f(T)} + \mathrm{O}\left(DS^{\frac{3}{2}}A\sqrt{\frac{\log\left(\frac{SAT^3}{\delta}\right)\log^2(T)}{f(T)}}\right). \tag{IV.13}$$

Proof. Episodes are partitioned into standard episodes $\mathscr{K}_0$ and non-standard episodes $\mathscr{K}_0 \setminus \mathscr{K}$. The elements of $\mathscr{K}_0$ of (type A), (type B) and (type C) are respectively denoted $\mathscr{K}_A, \mathscr{K}_B$ and $\mathscr{K}_C$. Their respective cardinals are K_A, K_B and K_C for short.

We start by upper-bounding $\mathscr{K} \setminus \mathscr{K}_0$. Such episodes are due to the doubling trick (DT) triggering, and we know since Auer et al. (2009) that doubling trick induces at $O(SA\log(T))$ episode.

We now upper-bound the number of (type A) episodes. For such episodes, there is some $z \in \mathscr{Z}$ that accounts for $n \geq \frac{1}{SA}K_A$ elements of $\mathscr{K}_A$. Let $k_1, k_2, \dots, k_n$ the respective episodes. We have by Equation (type A):

$$\forall i < n, \quad N_{t_{k_{i+1}}}(z) \geq N_{t_{k_i+1}}(z) \geq N_{t_{k_i}}(z)\left(1 + \frac{f(T)N_{t_{k_i}}(z)}{2^{10}D^2S\log\left(\frac{4SAT^3}{\delta}\right)}\right)$$

with $N_{t_{k_2}}(z) \geq 1$. Setting $u_i := N_{t_{k_{i+1}}}(z)$, we set $\lambda := Tf(T)2^{-10}D^{-2}S^{-1}\log^{-1}\left(\frac{4SAT^3}{\delta}\right)$ and $\omega = 1$, then apply Lemma IV.9. Since $u_{n-1} = N_{t_{k_n}}(z) \leq T$, that

$$n - 1 \leq 3 \cdot 2^5 D\sqrt{Sf(T)^{-1}\log\left(\frac{4SAT^3}{\delta}\right)}\log(T)$$

Using that $n \geq \frac{1}{SA}K_A$ and solving in K_A, we obtain

$$K_A \leq SA + 2^7 \cdot DS^{3/2}A\sqrt{\frac{1}{f(T)} \cdot \log\left(\frac{4SAT^3}{\delta}\right)}\log(T). \tag{IV.14}$$

We continue by upper-bounding the number of (type B) episodes. The proof is similar. There is some $z \in \mathscr{Z}$ that accounts for $n \geq \frac{1}{SA}K_B$ elements of $\mathscr{K}_{0,B}$. Let $k_1, k_2, \dots, k_n$ the respective episodes. We have by Equation (type B):

$$\forall i < n, \quad N_{t_{k_{i+1}}}(z) \geq N_{t_{k_i+1}}(z) \geq N_{t_{k_i}}(z)\left(1 + \sqrt{\frac{f(T)N_{t_{k_i}}(z)}{2^8D^2S\log(CT)}}\right),$$

with $N_{t_{k_2}}(z) \geq 1$. Setting $u_i := N_{t_{k_{i+1}}}(z)$, we set $\lambda := Tf(T)2^{-8}D^{-2}S^{-1}\log^{-1}(CT)$ and $\omega = \frac{1}{2}$ and apply Lemma IV.9. Since $u_{n-1} = N_{t_{k_n}}(z) \leq T$, we obtain

$$n - 1 \leq 3 \cdot 2^{\frac{4}{3}}D^{\frac{2}{3}}S^{\frac{1}{3}}f(T)^{-\frac{1}{3}}\log^{\frac{1}{3}}(CT) \cdot \log(T)$$

Using that $n \geq \frac{1}{SA}K_B$ and solving in K_B, we obtain

$$K_B \leq SA + 3 \cdot 2^{\frac{4}{3}}D^{\frac{2}{3}}S^{\frac{4}{3}}Af(T)^{-\frac{1}{3}}\log^{\frac{1}{3}}(CT) \cdot \log(T) = O\left(\frac{\log^{\frac{4}{3}}(T)}{f(T)^{\frac{1}{3}}}\right) \tag{IV.15}$$

We finish with the upper-bound of the number of (type C) episodes. Denote $n = K_C$ and introduce $k_1, k_2, \dots, k_n$ the elements of $\mathscr{K}_{0,C}$. By Equation (type C), we have

$$2^{-4}D^{-1}\sum_{i=1}^{n} f(t_{k_i}) \leq \sum_{i=1}^{n}\sqrt{S}\left(\sqrt{\log(Ct_{k_i+1})} - \sqrt{\log(Ct_{k_i})}\right) \leq \sqrt{S\log(CT)}.$$

We have $\sum_{i=1}^{n} f(t_k) \geq nf(T) = K_C f(T)$. We obtain accordingly:

$$K_C \leq \frac{2^4 D}{f(T)} \cdot \sqrt{S\log(T)} \tag{IV.16}$$

This concludes the proof. □

This concludes the proof of Theorem IV.6. ■

Corollary IV.7. *If $f(T) = \omega(T^{-\frac{1}{2}})$, then the regret guarantees of* UCRL2-PT *are the same than those of* UCRL2 *Auer et al. (2009). Specifically,*

$$\mathbf{E}^M[\mathrm{Reg}(T)] = \mathrm{O}\Big(DS\sqrt{AT\log(T)}\Big) \tag{IV.17}$$

with the same numerical constants.

12.2.3 Regret of exploration under (PT)

The sine qua non requirement of any episode rule has been established: The performance test (PT) does not harm the minimax regret guarantees. The whole point is to show that (PT) further improves the regret of exploration guarantees. There is however a significant subtlety because a few technical assumptions are required. The first is that instead of using the vanilla performance test, we use the **regenerative** performance test below:

$$t_{k+1} := \inf\Big\{t > t_k : g^{\pi_{t-1}}(\mathscr{M}_t) + f(T)e \leq g^*(\mathscr{M}_t) \text{ and } \exists t' \in [t_k, t), S_t = S_{t'}\Big\} \tag{RPT}$$

Namely, the regenerative property forces the episode to wait until the current state S_t has already been seen during the episode. This doesn't change the various properties provided by Theorem IV.6 and Corollary IV.7 nor their proofs. I don't know if the technical modifications leading (RPT) are necessary over (PT), but the regret of exploration guarantees are only established for (RPT). In the original paper Boone and Gaujal (2023b), regret of exploration guarantees where only provided for deterministic Markov decision processes. In this manuscript, we extend this result to much broader settings using a more mature technique. The setting in which (RPT) offers sublinear regret of exploration guarantees is not general however, because of actual limitations of optimism that are better discussed in the next chapter. The result holds under (1) a non-degeneracy assumption of the model (Definition IV.3) and (2) a relative interior assumption for the hidden model (see Assumption 5), that is close in spirit to knowing the support of transitions beforehand.

Assumption 5 (Interior kernels)**.** *For all $t \geq 1$ and $z \in \mathscr{Z}$, $p(z)$ is interior to $\mathscr{P}_t(z)$, i.e., $\tilde{p}(z) \ll p(z)$ for all $\tilde{p}(z) \in \mathscr{P}_t(z)$.*

Theorem IV.8. *Let $M \in \mathscr{M}^+$ a non-degenerate explorative model. Assume that the confidence region of* UCRL2-PT *satisfies Assumption 5. Provided that $f(t) = \mathrm{o}\Big(\frac{1}{\log(t)}\Big)$, there exists a constant $C(M) < \infty$ such that* UCRL2-PT *satisfies:*

$$\mathrm{RegExp}(T) \leq C(M)\log(T) + \mathrm{o}(\log(T)). \tag{IV.18}$$

Regarding how UCRL2-PT is currently written (Algorithm IV.3), Assumption 5 only covers ergodic models. However, in the design of UCRL2-PT, we can also impose that $\mathscr{P}_t(z) \subseteq \{p'(z) : p'(z) \ll p(z)\}$, hence incorporating prior knowledge on the support of kernels, to obtain regret of exploration guarantees beyond ergodic models. This covers deterministic models and known kernels settings as special cases. The proof of Theorem IV.8 is postponed to the next chapter, that is dedicated to the proof technique.

Combining Corollary IV.7 and Theorem IV.8, the performance test allows for any slackness function f which is simultaneously $\mathrm{o}\Big(\frac{1}{\log(t)}\Big)$ and $\omega\Big(\frac{1}{\sqrt{t}}\Big)$. In practice, if the diameter of the

model is small, taking $f(t)$ as small as possible seems better. If the diameter of the model is large, one may want to be careful and pick $f(t)$ closer to $\frac{1}{\log(t)}$, for instance $\frac{c}{\log^2(t)}$ with $c \ll 1$. If $f(t)$ is large, the performance test doesn't trigger and the behavior of UCRL2-PT is the same as UCRL2's.

12.2.4 Experimental insights

In the remaining of the chapter, we rather provide experimental insights regarding the behavior of (PT). The simplest setting is a two-arm bandit example, say with two arms a_1, a_2 of respective mean rewards $r(a_1) = 0.5$ and $r(a_2)$.

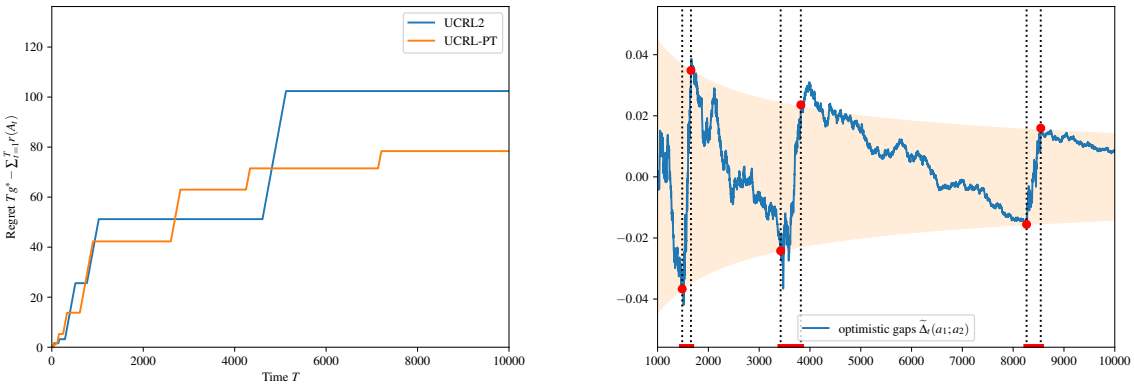


Figure 12.2.1: The behavior of UCRL2-PT on a two-arm bandit.

In Figure 12.2.1 (to the left), we display the regret of UCRL2 and UCRL2-PT over a single run. When none of the algorithms has better regret than the other, we nonetheless observe that the bad episodes of UCRL2-PT (where the regret slope is not null) are shorter than UCRL2's, suggesting that the regret of exploration is indeed better. To the right of Figure 12.2.1, we deep-dive into the behavior of UCRL2-PT by displaying the **optimistic gaps** of the suboptimal arm over time. The current policy π^k corresponds to which arm is drawn over $\{t_k, \dots, t_{k+1}-1\}$ and we know that the algorithm changes episodes when $g^{\pi^k}(\mathcal{M}_t) + f(t) \leq g^*(\mathcal{M}_t)$, hence the behavior of the algorithm is driven by the value of the optimistic gap $\tilde{\Delta}_t(a_1;a_2) := g^{a_1}(\mathcal{M}_t) - g^{a_2}(\mathcal{M}_t)$, plotted in Figure 12.2.1 (to the right). On the time-window $\{t, \dots, t+T-1\}$, the regret is proportional to the amount of time spent drawing arm a_2, highlighted in red on the time-axis. A bad episode starts when $\tilde{\Delta}_t(a_1;a_2)$ drops below $-f(t)$, then is interrupted as soon as it goes above $f(t)$ again. The filled region represents the points (x, y) for $|y| \leq f(x)$.

We observe that when a bad episode occurs, arm a_2 is triggered and its bonus decreases, making the associated optimistic gap drop quickly despite the noise. The update of empirical estimates is observed noise and is clearly non-negligible with respect to the upward drift induced by a decrease of the optimistic bonus of a_2. Still, that drift forces the bad episode to end rather quickly. We then switch to a good episode where the noise amplitude is much smaller. This is because the optimal arm a_1 is visited more often than a_2, hence its empirical reward estimate changes slower. We also observe a drift (there downward) which is due to the evolution of reward bonuses, but the drift is weaker because the visit counts of a_1 is much higher.

We accordingly observe that good episodes last, while bad episodes don't. Hence the regret of exploration is small.

12.3 Computational heaviness of the performance test

The performance test suffer from obviously heavy computational performance. At every time step, the optimistic value of the current policy π^k and of the whole confidence region $\mathscr{M}_t$ must be computed by running `EVI`. These additional computations can make (PT) unreliably performance hungry when the numbers of states and actions grow large. Thankfully, (PT) can be substantially accelerated with the combination of two methods.

(1) **Memorization** (M): From t to $t+1$, the confidence region is barely changed. Therefore, by initializing the `EVI` at time $t+1$ with the result of `EVI` at time t, one should expect `EVI` to converge much faster. This doesn't modify the algorithm's behavior.

(2) **Sparse** (PT) (S): Even if `EVI` converges almost instantly, running `EVI` at each time-step significantly slows down the algorithm. To address this, instead of always checking (PT), only test it when $t - t_k$ is a power of 2. Formally, (PT) is replaced by (PT*):

$$\log_2(t-t_k) \in \mathbf{N} \quad \text{and} \quad g^{\pi_{t-1}}(\mathscr{M}_t) + f(T)e \leq g^*(\mathscr{M}_t) \tag{PT*}$$

Although this modification slightly alter the behavior of the algorithm, its analysis is similar to (PT)'s.

As shown in Table 12.3.1, the combination of these two modifications makes the running times of (PT) variants acceptable in comparison to the originals.

	Original	Pure (PT)	(PT) + M	(PT) + S	(PT) + M + S
`UCRL2`	200k	1.0k	5.2k	74k	82k
`KLUCRL`	157k	0.2k	2.7k	36k	62k
`UCRL2B`	167k	0.5k	2.9k	59k	77k

Table 12.3.1: Iterations per second of UCRL2, KLUCRL and UCRL2B; the originals and the (PT) corrections with various acceleration options. The environment is a 5-state RiverSwim. These values have been obtained by running each algorithm for 100k iterations and take the average per-step time.

Appendix of Chapter 12

We provide below a numerical lemma used in the proof of Theorem IV.6.

Lemma IV.9. *Let $T \geq 3$ and fix $\lambda \leq T$. Let $\omega \in (0,1]$ and $(x_n \mid n \geq 1)$ an integer-valued sequence with $x_1 := 1$ such that*

$$x_{n+1} \geq \left(1 + \left(\frac{\lambda x_n}{T}\right)^{\omega}\right) x_n. \tag{IV.19}$$

If n is such that $x_n \leq T$, then $n \leq 3\left(\frac{T}{\lambda}\right)^{\frac{\omega}{\omega+1}} \log(T)$.

We further conjecture that the $\log(T)$ is not necessary, i.e., that if $x_n \leq T$ then $n = O((T/\lambda)^{\omega/(\omega+1)})$.

Proof. Define the integer valued sequence $x_{n+1} = \lceil (1 + (\lambda x_n/T)^{\omega}) x_n \rceil$ initialized to $x_1 = 1$ and analyze the increments of (x_n). Observe that $x_{n+1} > x_n$, so $x_{n+1} \geq x_n + 1$ and the sequence diverges to infinity. Setting $\beta := \frac{1}{\omega+1} \in (0,1)$, for $k \geq 1$, we get

$$x_{n+1} = x_n + k \iff x_n \in \left(\left(\tfrac{T}{\lambda}\right)^{1-\beta}(k-1)^{\beta}, \left(\tfrac{T}{\lambda}\right)^{1-\beta} k^{\beta}\right] =: I_k.$$

The length of I_k is decreasing with k and in particular $\mathrm{Leb}(I_k) \leq \mathrm{Leb}(I_1) = (\frac{T}{\lambda})^{1-\beta}$. Accordingly, the integer-valued sequence (y_n) with $y_1 = 1$ defined by its increments

$$y_{n+1} = y_n + k \iff y_n \in \left(\left(\tfrac{T}{\lambda}\right)^{1-\beta}(k-1), \left(\tfrac{T}{\lambda}\right)^{1-\beta} k\right] \tag{IV.20}$$

satisfies: $\forall n \geq 1, y_n \leq x_n$. Moreover,

$$\begin{aligned}
\forall n \geq 1, \quad y_{n+1} &= y_n + \left\lceil y_n \left(\tfrac{\lambda}{T}\right)^{1-\beta} \right\rceil && \text{(IV.21)}\\
&\geq y_n \left(1 + \left(\tfrac{\lambda}{T}\right)^{1-\beta}\right) && \text{(IV.22)}\\
&\geq \dots && \text{(IV.23)}\\
&\geq \left(1 + \left(\tfrac{\lambda}{T}\right)^{1-\beta}\right)^n. && \text{(IV.24)}
\end{aligned}$$

Let $n \geq 1$ such that $x_n \leq T$. Then $y_n \leq T$, hence $(1 + (\lambda/T)^{1-\beta})^{n-1} \leq T$. Thus

$$(n-1) \log\left(1 + \left(\tfrac{\lambda}{T}\right)^{1-\beta}\right) \leq \log(T). \tag{IV.25}$$

Since $\lambda \leq T$, we have $(\lambda/T)^{1-\beta} \in (0,1]$ so $\log(1 + (\lambda/T)^{1-\beta}) \geq \frac{1}{2}(\lambda/T)^{1-\beta}$. We obtain:

$$n \leq 1 + 2\left(\tfrac{T}{\lambda}\right)^{1-\beta} \log(T) \leq 3\left(\tfrac{T}{\lambda}\right)^{1-\beta} \log(T) \tag{IV.26}$$

and as $1 - \beta = \frac{\omega}{\omega+1}$, this proves the claim. □

Chapter 13

The Vanishing Multiplicative Condition (VM)

In Chapter 12, we provided (PT) as an enhancement of episode management and claimed that it guarantees sub-linear regret of exploration (Theorem IV.8). However, the claim is short of a proof; The minimax guarantees under (PT) rely on a difficult upper-bound on the number of episodes (Theorem IV.6) that works for ℓ_1-confidence regions (II.5) that is not easily generalized to KL-confidence regions (II.6) or Bernstein confidence regions (II.7), and even harder to generalize if the algorithm is PMEVI-based (Algorithm II.5) rather than EVI-based (Algorithm II.3); (PT) is computationally heavy and requires substantial optimizations to be run (Section 12.3); Lastly, the slackness function $f(t)$ of (PT) is difficult to choose. The last point is more important that it sounds. The slackness function guarantees that the number of episodes grows reasonably slow (Theorem IV.6), but the theoretical guarantees are very pessimistic and are way off in practice. Morally, this means that $f(t)$ may be chosen aggressively small, at the risk of entering the back-and-forth issue of Section 12.1, yet choosing $f(t)$ small is very important to observe sub-linear regret of exploration in practice because the guarantees are otherwise too asymptotic.

The hardness of tuning the slackness function of (PT) is produced by a common flaw of design: We are **explicitly** controlling the error range of the algorithm. It is very difficult to precisely quantify the speed at which optimistic gain of a played sub-optimal policy decreases, hence designing the slackness function is a hard problem. In the chapter, we provide another episode rule, the **Vanishing Multiplicative condition** (VM), that is **implicit** rather than explicit.

> **Important remark.** In the remaining of the chapter, we write $u_z(t)$ instead of $u_t(z)$.

13.1 Optimizing (PT): the vanishing multiplicative condition

This new episode rule, despite its simplicity, was designed by reverse engineering the proof of regret of exploration guarantees of Boone and Gaujal (2023b).

13.1.1 The return of visit counts

When investigating the experimental behavior of (PT) in Section 12.2.4, we have indirectly observed that the behavior of confidence regions is very different at high and low visit counts (to the right of Figure 12.2.1). At high visit counts, the optimistic estimates seem to behave like a random walk if a nearly negligible drift. At low visit counts, the optimistic estimates are very noisy but are pulled by a strong negative drift. At the end of the day, if sublinear regret of

exploration is achievable with optimistic methods in the first place, it is because the confidence region associated to a rarely visited pair evolves quickly when the number of visits increases. Hence, waiting for **relative increases** of visit counts just like (DT) is actually a good idea: If the pair z is rarely visited, its confidence region changes quickly and it is mandatory to frequently update the current policy when z is played, and if z is visited often, its confidence region doesn't change much upon visiting it and the current policy may not need that many updates. The issue of (DT) is that the visit counts of all pairs are subjected to be unbounded, hence waiting for a visit count to double is too much. We suggest to replace (DT) by a less demanding visit requirement. Specifically, the current episode k is ended as soon the visit count of a pair is about to increase **multiplicatively** with respect to a **vanishing time dependent multiplicative factor**, i.e.,

$$N_t(S_t, \pi^k(S_t)) > (1 + f(t_k))N_{t_k}(S_t, \pi^k(S_t)) \tag{VM}$$

where $f \in [0,1]^{\mathbb{N}}$ is a non-increasing vanishing function of t. The above condition will be referred to as the f-**Vanishing Multiplicative condition**, or f-(VM), or even more simply (VM). Remark that (DT) is of the form (VM) with $f \equiv 1$, excepted that this function is not vanishing. This control function can be almost arbitrary among those valued in $[0,1]$, but as the analysis will show, (1) $f(t) = o\left(\frac{1}{\log(t)}\right)$ is mandatory to get sublinear guarantees on the regret of exploration, and (2) standard minimax regret guarantees impose f to be of order $f(t) = \Omega(t^{-1/2})$. Remark that (1-2) are the same requirement than (PT). This episode rule further satisfies the following convenient measurability property. Fixing an episode (k) and denoting $\tau_s^t := \inf\{t' > t : S_{t'} = s\}$ the reaching time to $s \in \mathscr{S}$ since t, observe that for all $t \in \{t_k, \dots, t_{k+1}\}$, either $\mathbf{1}(t_{k+1} \leq \tau_s^t)$ or $\mathbf{1}(t_{k+1} > \tau_s^t)$ is $\sigma(H_t)$-measurable. It means that at time t, one is able to tell whether there exists a state that kills the episode upon its next future visit or not.

The point of (VM) is to achieve two goals at once: leave the regret guarantees unharmed and ensure regret of exploration guarantees. The benefits of changing (DT) to (VM) in classical EVI-based algorithms is summarized in the table below. One of the advantages of (VM) over (PT) is that its theoretical guarantees extend to a larger range of algorithms, including KLUCRL Filippi et al. (2010); Talebi and Maillard (2018) and UCRL2-B Fruit et al. (2020).

Algorithm	Minimax regret	Model dependent regret*	Regret of exploration
UCRL2-(DT)	$DS\sqrt{AT\log(T)}$	$O(\log(T))$	$\Omega(T)$
UCRL2-(VM)	$DS\sqrt{AT\log(T)}$	$O(\log(T)\log\log(T))$	$O(\log(T))$
KLUCRL-(DT)	$S\sqrt{DAT\log(T)}$	$O(\log(T))$	$\Omega(T)$
KLUCRL-(VM)	$S\sqrt{DAT\log(T)}$	$O(\log(T)\log\log(T))$	$O(\log(T))$
UCRL2-B-(DT)	$S\sqrt{DAT\log^2(T)}$	unknown	$\Omega(T)$
UCRL2-B-(VM)	$S\sqrt{DAT\log^2(T)}$	$O(\log(T)\log\log(T))$	$O(\log(T))$

* The model dependent guarantees of (VM) requires additional technical assumptions on the model.

Table 13.1.1: Theoretical guarantees of classical algorithms with (VM).

13.1.2 Minimax regret guarantees under (VM)

Again, the sine qua non requirement for (VM) is that the minimax regret guarantees are untouched. Thankfully, regret guarantees are much easier to establish than under (PT) because the number of episodes is much easier to bound. Similarly to Section 12.2.1,

$$\mathrm{Reg}(T) \leq D(M)|\mathscr{K}(T)| + \sum_{k,t}\Big(g^*(\mathscr{M}(t_k)) - \tilde{r}^k(Z_t)\Big) + \mathrm{O}\Big(\sqrt{SAT\log\big(\tfrac{T}{\delta}\big)}\Big) \tag{IV.1}$$

where $(\tilde{r}^k, \tilde{p}^k)$ is an optimistic model of π^k at time t_k (see Section 6.4 for more details) provided that the confidence region holds universally in time. The number of episodes is bounded by Proposition IV.10 below.

Proposition IV.10. *The number of episodes under (VM) is bounded by:*

$$|\mathscr{K}(T)| \leq SA\left(1 + \frac{2\log(T)}{f(T)}\right) \tag{IV.2}$$

Proof. By considering a pair interrupting the maximal number of episodes, we see that

$$(1+f(T))^{\frac{|\mathscr{K}(T)|}{SA}-1} \leq T.$$

Solve in $|\mathscr{K}(T)|$ and use that $\log(1+x) \geq \frac{1}{2}x$ for $x \in [0,1]$. □

Corollary IV.11 (Regret guarantees). *If $f(T) = \omega(T^{-\frac{1}{2}})$, then the regret guarantees of* UCRL2-(VM)*,* KLUCRL-(VM) *and* UCRL2B-(VM) *match their (DT) equivalents, see Table 13.1.1.*

13.1.3 Regret of exploration guarantees under (VM)

We now present the main result of the chapter: providing regret of exploration guarantees for optimistic algorithms managing episodes with (VM). Our result holds under (1) a non-degeneracy assumption of the model, and (2) a relative interior assumption for the hidden model (see Assumption 5) that is close in spirit to knowing the support of transitions beforehand.

Theorem IV.12 (Main result). *Let $M \in \mathscr{M}_+$ a non-degenerate explorative model. Consider running an* EVI*-based algorithm with confidence region $\mathscr{M}(t) \equiv \mathscr{M}_{\delta(t)}(t)$ as in Section 7.A.2 with $\delta(t) := \frac{1}{t}$, and assume that $\mathscr{M}(t)$ satisfies Assumption 5. If the algorithm manages episodes according to f-(VM) with $f(t) = \mathrm{o}\big(\frac{1}{\log(t)}\big)$, then there exists a constant $C(M) < \infty$ such that:*

$$\mathrm{RegExp}(T) \leq C(M)\log(T) + \mathrm{o}(\log(T)).$$

This result can be explicitly applied to UCRL2 Auer et al. (2009), KLUCRL Filippi et al. (2010) and UCRL2-B Fruit et al. (2020) to obtain logarithmic regret of exploration guarantees for the variants obtained by managing episodes using (VM) rather than (DT).

The proof of Theorem IV.12 is difficult and its structure is illustrated with Figure 13.1.1. The remaining of the chapter is dedicated to explaining the main ideas behind it. In Section 13.2.1, we present a family of **coherence** properties, properties that imply logarithmic regret of exploration

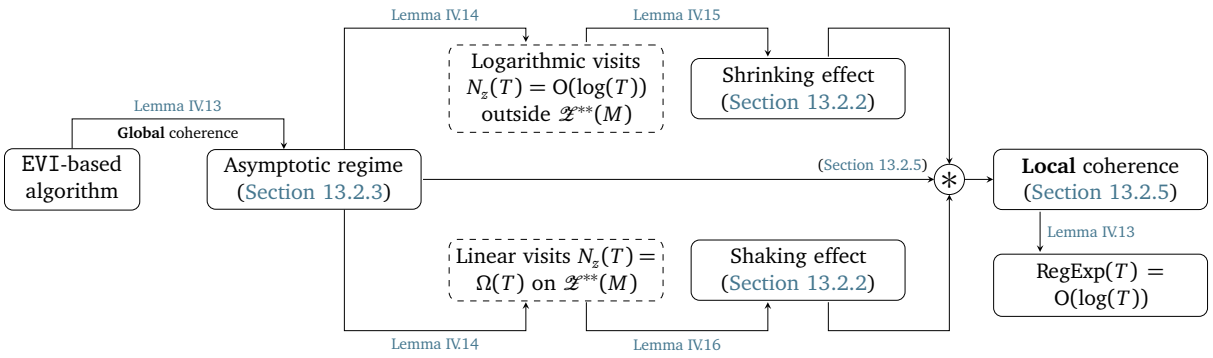


Figure 13.1.1: Architecture of Regret of Exploration Analysis.

and beyond for algorithms that satisfy them. In Section 13.2.2, we explain why optimistic algorithms enjoy coherence properties. This goes by characterizing the asymptotic visit rates of pairs in Section 13.2.3 and explain what non-degeneracy has to do with it; In Section 13.2.3.1, we explain why Assumption 5 is important. Among all the ideas invoked to establish coherence, the most important step is detailed in Section 13.2.4 where we investigate a **shrinking-shaking** behavior that confidence regions display, which plays a crucial role in making sub-optimal policies cast out by the whole EVI machinery. The throughout details are very voluminous and are deferred to appendix for the most part, excepted for the last Section 13.2.5 that connects all the arguments together (marked node $(*)$ of Figure 13.1.1).

13.2 Establishing regret of exploration guarantees

13.2.1 Coherent algorithms

Our proof strategy is valid for all EVI-based algorithms (Algorithm II.2) that use a confidence region $\mathscr{M}(t)$ as discussed in Section 6.3.2. The regret of exploration is upper-bounded by estimating the number of times it is expected to play a sub-optimal policy over time intervals $[t_k, t_k + T)$ starting at exploration times t_k, then by deducing that the regret is small on the same time period. This is achieved via Lemma IV.13 below, that controls the likely range of the regret under a stability condition specified in Definition IV.5, that we refer to as **coherence**.

Definition IV.5 (Coherence). *We say that an algorithm is* (F, τ, T, φ)*-coherent if* $F \equiv (F_t : t \geq 1)$ *is an adapted sequence of events,* τ *a stopping time,* $T \geq 1$ *is a scalar and* $\varphi : \mathbf{N} \to [0, \infty)$ *is a function such that, for all* $t \in \{\tau, \dots, \tau + T - 1\}$,

$$F_t \subseteq \left\{ g^{\pi_t}(S_t) < g^*(S_t) \Rightarrow \exists z \equiv (s,a) \in \mathrm{Reach}(\pi_t, S_t) : \begin{bmatrix} N_z(t) - N_z(\tau) \leq \varphi(\tau) \\ \text{and } g^{\pi_t}(s) < g^*(s) \end{bmatrix} \right\}$$

where $z \equiv (s,a) \in \mathrm{Reach}(\pi_t, S_t)$ *stands for* $\pi(a|s) > 0$ *and* $\mathbf{P}^{\pi}_{S_t}(\tau_s < \infty) > 0$.

Formally, coherence states that over $[\tau, \tau + T)$ and under a good event F_t, if the current policy π_t is sub-optimal from the current state, then there exists a sub-sampled pair (relatively to $\varphi(\tau)$) which is reachable with positive probability from the current state S_t with the current policy π_t and from which the current policy is sub-optimal. Morally, coherence states that the iteration of a sub-optimal policy is linked to a lack of information that has positive probability

to be recovered from by iterating that policy only. That lack of information is quantified by a budget $\varphi(\tau)$. The purpose of the coherence property is its link with local regret guarantees, as shown by Lemma IV.13 below. However, the coherence property may only be conveniently used if the episodes of the algorithm are **regenerative**, meaning that episodes may only end if the current state has already been visited during the episode. This property makes sure that the sub-sampled state-action pair, of which coherence ensures the existence, is reached and visited during the episode with positive probability.

Definition IV.6 (Regenerative episodes)**.** *We say that the episodes of an algorithm are* ***regenerative*** *if, for all $k \geq 1$, there exists $t \in [t_k, t_{k+1})$ such that $S_t = S_{t_{k+1}}$.*

Lemma IV.13 (Coherence and local regret)**.** *Assume that the underlying model M is non-degenerate (Definition IV.3). If the algorithm is (F, τ, T, φ)-coherent and has regenerative episodes, then there exist model dependent constants $C_1, C_2, C_3, C_4 > 0$ such that:*

$$\forall x \geq 0, \quad \mathbf{P}\left(\mathrm{Reg}(\tau, \tau+T) \geq x + C_4\varphi(\tau) \text{ and } \bigcap_{t=\tau}^{\tau+T-1} F_t\right) \leq C_1 T^{C_3} \exp\left(-\frac{x}{C_2}\right).$$

More specifically, C_1, C_2, C_3, C_4 only depend on M and are independent of F, τ, T and φ.

Using the shorthand $F_{\tau:\tau+T} := \bigcap_{t=\tau}^{\tau+T-1} F_t$, this means that on a good event $F_{\tau:\tau+T}$, the local regret $\mathrm{Reg}(\tau, \tau+T)$ has sub-exponential tails. The above result can also be written in the form $\mathbf{P}(\mathrm{Reg}(\tau, \tau+T) \geq C_1 + C_4\varphi(\tau) + (\eta C_2 + C_3)\log(T), F_{\tau:\tau+T}) \leq T^{-\eta}$ where $\eta > 0$ is arbitrary.

This result is invoked at several places in the analysis in intrinsically different scenarios. It plays a central role in the instance dependent regret analysis (done in expectation), but also in showing that non optimal pairs ($\mathcal{Z} \setminus \mathcal{Z}^{**}(M)$) are visited at most logarithmically often (almost surely). Perhaps more importantly, it plays the last key part in the proof of regret of exploration guarantees, as shown by Figure 13.1.1. The idea is simple: Applying Lemma IV.13 with $\eta = 1$, we get $\mathbf{E}[\mathrm{Reg}(\tau, \tau+T) | F_{\tau:\tau+T}] = \mathrm{O}(\log(T))$. We then typically choose τ as an exploration time of the process and F as a high probability concentration event whose construction is algorithm specific, and that depends on how coherence is established. This last part is the subject of the next section.

The proof of Lemma IV.13 is difficult and deferred to Section 13.A.

13.2.2 The shrinking/shaking behavior of confidence sets and coherence

In this part, we explain why EVI-based algorithms managing episodes with (VM) are coherent. Recall that the coherence property states that if a sub-optimal policy is being used, then $(*)$ a **reachable** pair has been sub-sampled and $(**)$ it is sub-sampled within $\mathrm{O}(\log(T))$ range. While $(*)$ leads to making assumptions on the space of models on which the algorithm runs, $(**)$ is linked to a shrinking-shaking behavior of confidence sets that leads to non-degeneracy.

13.2.3 Asymptotic regime of EVI-based algorithms, (VM) and non-degeneracy

The result below describes the almost-sure asymptotic regime of EVI-based algorithms managing episodes with (VM). Up to the non-degeneracy of the underlying model, the visit counts can be split into two regimes: $N_z(t)$ grows linearly with t for $z \in \mathcal{Z}^{**}(M)$ while $N_z(t)$ grows sub-logarithmically for $z \notin \mathcal{Z}^{**}(M)$ (including $\mathcal{Z}^{**}(M) \setminus \mathcal{Z}^{*}(M)$ in particular). Both results are mandatory steps in the proof of regret of exploration guarantees, see Figure 13.1.1.

Lemma IV.14 (Almost-sure asymptotic regime). *Let $M \in \mathscr{M}$ a non-degenerate model. Assume that the algorithm running is* EVI*-based with a confidence region $\mathscr{M}(t) \equiv \mathscr{M}_{\delta(t)}(t)$ as given by Section 7.A.2 with $\delta(t) := \frac{1}{t}$, managing episodes with f-(VM) with arbitrary $f > 0$. There exists $\lambda > 0$ s.t.:*

$$\forall z \notin \mathscr{Z}_{**}(M), \quad \mathbf{P}^M(\exists T, \forall t \geq T : N_z(t) < \lambda \log(t)) = 1, \text{ and}$$
$$\forall z \in \mathscr{Z}_{**}(M), \quad \mathbf{P}^M\Big(\exists T, \forall t \geq T : N_z(t) > \tfrac{1}{\lambda} t\Big) = 1.$$

Since the result holds for arbitrary f, it holds in particular for (DT). This is not much of a surprise, since UCRL2 is known to have model dependent logarithmic regret. This result is rather remarkable for (VM) since the number of episodes can be much larger than logarithmic. However, (DT) and (VM) differ in the amount of times a sub-sampled pair can be visited during an episode. Indeed, for $z \in \mathscr{Z}$ such that $\mathbf{P}^M(\exists T, \forall t \geq T : N_z(t) < \lambda \log(t)) = 1$, we see that:

$$N_z(t_{k+1}) \leq \lfloor (1 + f(t_k)) N_z(t_k) \rfloor + 1 = N_z(t_k) + 1 + \lfloor \lambda f(t_k) \log(t_k) \rfloor. \tag{IV.3}$$

For $f(t) = o\Big(\frac{1}{\log(t)}\Big)$, we have $\lfloor \lambda f(t_k) \log(t_k) \rfloor = 0$ provided that t_k is large enough, meaning that sub-sampled pairs can be visited at most once per episode in the long run. It means that under (VM), algorithms almost instantly refresh their policies when sub-optimal pairs are visited.

13.2.3.1 Recurrence or known supports

The first statement $(*)$, which is about the reachability of sub-sampled pairs, is not guaranteed to hold if we run an EVI-based algorithm on an arbitrary model. The issue lies in the fact that the high optimistic gain of a policy may be due states that are unreachable under the optimistically optimal policy. This is because in the confidence region $\mathscr{M}(t)$, there may be models with a richer transition structure than the true hidden model M. It is avoided using Assumption 5.

Assumption 1. *For all $t \geq 1$ and $z \in \mathscr{Z}$, $p(z)$ is interior to $\mathscr{P}_t(z)$, i.e., $\tilde{p}(z) \ll p(z)$ for all $\tilde{p}(z) \in \mathscr{P}_t(z)$.*

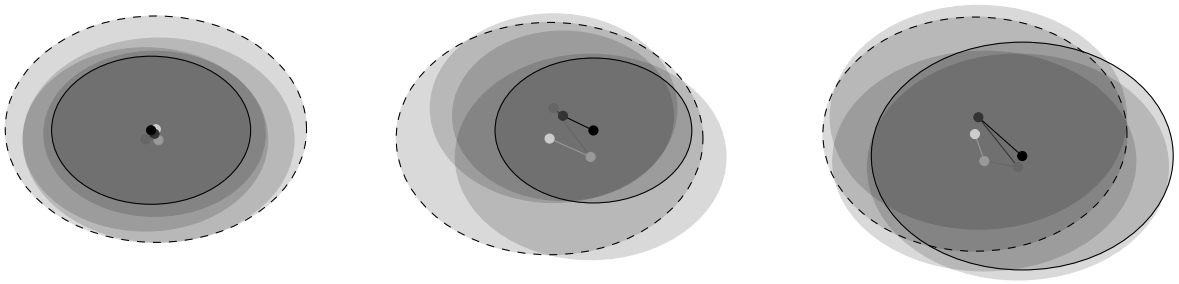
Assumption 5 is morally equivalent to stating that the support of the transitions of M are known in advance and we conjecture this assumption is mandatory without a significant rework of EVI. Reworking EVI however is not the subject of this chapter. Under Assumption 5, the optimistic gain of a policy π from a fixed state s only depends on $\mathscr{R}_z(t), \mathscr{P}_z(t)$ for pairs z that are reachable from s under π on M. This echoes the reachability requirement of sub-sampled pairs in $(*)$.

13.2.4 Shrinking and shaking confidence sets

The phenomenon presented in this section is an abstract view of what has been observed in the right part of Figure 12.2.1.

The second statement $(**)$ of coherence is that if $N_z(t) - N_z(\tau)$ is large enough, then EVI will reject sub-optimal policies, typically because their optimistic gain drops. This is achieved by showing that confidence regions collapse with high probability on sub-sampled pairs when their visit counts increase; Indeed, if the confidence region associated to a policy's kernel and reward vector are smaller, then so is the optimistic gain computed by EVI. The fact that confidence regions shrink quickly enough is not obvious and is a result of the $O(\log(T))$ visit rate of sub-optimal transitions, which is thankfully guaranteed by the characterization of the asymptotic regime given by Lemma IV.14. Then, the fact that the shrinking behavior of confidence regions

implies a decrease of the optimistic gain is not obvious either; But this last part is highly technical, hence completely deferred to the appendix.



$N_{s,a}(t) = o(\log(t))$ (Shrinking, Lemma IV.15) | $N_{s,a}(t) = \Theta(\log(t))$ (Shrinking, Lemma IV.15) | $N_{s,a}(t) = \Omega(\log(t))$ (Shaking, Lemma IV.16)

Figure 13.2.1: An artist view of the shrinking/shaking behavior of the confidence regions $\mathcal{Q}_{s,a}(t)$ as the number of new samples $N_{s,a}(t') - N_{s,a}(t) \ll N_{s,a}(t)$ increases (from dashed to solid line).

Formally, the **shrinking effect** is formalized with Lemma IV.15 below.

Lemma IV.15 (Shrinking effect). *Let $(t_{k(i)})$ the enumeration of exploration episodes, and let $T \geq 1$. Fix $\lambda, z \in \mathcal{Z} > 0$. For all $\delta > 0$, we can find $\epsilon, m, C > 0$ such that:*

$$\textit{kernel}: \quad \mathbf{P}\left(F_{t_{k(i)}}, \left[\exists t \in [t_{k(i)}, t_{k(i)} + T] : \begin{array}{c} \mathcal{P}_z(t) \not\subseteq \mathcal{P}_z(t_{k(i)-1}) \\ \text{and } N_z(t) > N_z(t_{k(i)}) + C\log\left(\frac{T}{\delta}\right) \end{array}\right]\right) \leq \delta,$$

$$\textit{reward}: \quad \mathbf{P}\left(F_{t_{k(i)}}, \left[\exists t \in [t_{k(i)}, t_{k(i)} + T] : \begin{array}{c} \sup \mathcal{R}_z(t) > \sup \mathcal{R}_z(t_{k(i)-1}) - \frac{N_z(t) - N_z(t_{k(i)})}{C\log(t_{k(i)})} \\ \text{and } N_z(t) > N_z(t_{k(i)}) + C\log\left(\frac{T}{\delta}\right) \end{array}\right]\right) \leq \delta$$

with $F_{t_{k(i)}} := (N_z(t_{k(i)}) < \frac{1}{\lambda}\log(t_{k(i)}), |\hat{r}_z(t_{k(i)-1}) - r_z| < \epsilon, \|\hat{p}_z(t_{k(i)-1}) - p_z\| < \epsilon, t_{k(i)} > m)$.

Remark that the shrinking is shown to be large on kernels, and strict on rewards. The good event $F_{t_{k(i)}}$ is asymptotically almost sure. Indeed, $(N_z(t_{k(i)}) < \frac{1}{\lambda}\log(t_{k(i)}))$ refers to the logarithmic visit rates provided by Lemma IV.14, the concentration events $(|\hat{r}_z(t_{k(i)-1}) - r_z| < \epsilon, \|\hat{p}_z(t_{k(i)-1}) - p_z\| < \epsilon)$ hold because of the law of large numbers, and $(t_{k(i)} > m)$ is just stating that $t_{k(i)}$ is large enough. The careful reader will remark that in the events above, $t_{k(i)}$ and $t_{k(i)-1}$ co-exist: The evolution of visit counts is measured relatively to their status at time $t_{k(i)}$, while the evolution of confidence regions is measured relatively to their status at time $t_{k(i)-1}$. This subtlety is technical, and comes from the fact that our goal is to bound the regret starting from $t_{k(i)}$, yet the confidence regions are better behaved at time $t_{k(i)-1}$ than at time $t_{k(i)}$.

A dual result holds for the confidence of highly visited pairs, showing that the associated confidence regions barely change on small time windows. This is the **shaking effect**, where the center of the confidence region moves faster than the region's diameter decreases.

Lemma IV.16 (Shaking effect). *Let $(t_{k(i)})$ the enumeration of exploration episodes, and let $T \geq 1$. Fix $\lambda, z \in \mathcal{Z}$ and for two sets $\mathcal{U}, \mathcal{V} \subseteq \mathbf{R}^n$, denote $d_{\mathrm{H}}(\mathcal{U}, \mathcal{V})$ the Hausdorff distance induced by the one-norm. We can find $c, m > 0$ such that:*

$$(\textit{kernels}) \quad F_{t_{k(i)}} \subseteq \left(\forall t \in [t_{k(i)}, t_{k(i)} + T] : d_{\mathrm{H}}(\mathcal{P}_z(t), \mathcal{P}_z(t_{k(i)-1})) \leq \sqrt{\frac{c\log(t)}{t}}\right),$$

$$\textit{(rewards)} \quad F_{t_{k(i)}} \subseteq \left(\forall t \in [t_{k(i)}, t_{k(i)} + T] : d_{\mathrm{H}}(\mathscr{R}_z(t), \mathscr{R}_z(t_{k(i)-1})) \leq \sqrt{\tfrac{c \log(t)}{t}} \right)$$

where $F_{t_{k(i)}} := (N_z(t_{k(i)-1}) > \lambda t_{k(i)-1}, t_{k(i)} > m) \cap (\forall t \in [t_{k(i)-1}, t_{k(i)}], M \in \mathscr{M}_{\delta(t)}(t))$.

13.2.5 Establishing coherence and proving Theorem IV.12

Based on the shrinking and shaking effect discussed upstream, we show that (VM) guarantees **local** coherence properties that, once combined with Lemma IV.14, become regret of exploration guarantees. The exact coherence property is detailed in Lemma IV.17 below. Once Lemma IV.17 is established, Theorem IV.12 follows instantly.

Lemma IV.17. *Let $M \in \mathscr{M}_+$ a non-degenerate explorative model. Consider running an* EVI*-based algorithm with confidence region $\mathscr{M}(t) \equiv \mathscr{M}_{\delta(t)}(t)$ as in Section 7.A.2 with $\delta(t) := \frac{1}{t}$, and assume that $\mathscr{M}(t)$ satisfies Assumption 5. Assume that the algorithm manages episodes according to f-(VM) with $f(t) = \mathrm{o}\left(\frac{1}{\log(t)}\right)$. Let $(t_{k(i)})$ the enumeration of exploration episodes. Then, there exists a constant $C(M) > 0$ such that, for all $T \geq 1$ and $\delta > 0$, there is an adapted sequence of events (E_t) and a function $\varphi : \mathbf{N} \to \mathbf{R}$ such that:*

(1) *For all $i \geq 1$, the algorithm is $(E_t, t_{k(i)}, T, \varphi)$-coherent;*
(2) $\mathbf{P}(\bigcup_{t=t_{k(i)}}^{t_{k(i)}+T-1} E_t^c) \leq \delta + \mathrm{o}(1)$ *when* $i \to \infty$;
(3) $\varphi(t_{k(i)}) \leq 1 + C \log(\frac{T}{\delta})$ *when* $i \to \infty$.

Proof. By correctness of the confidence region, $\mathbf{P}(\exists T, \forall t \geq T : \forall \pi, g^*(\pi, \mathscr{M}(t)) \geq g(\pi, M)) = 1$, hence a policy with optimistic gain less than $g^*(M)$ won't be optimistically optimal on this event, so won't be the result of EVI. Considering an exploration time $t_{k(i)}$, we know that the policy of the previous episode was optimal in M, hence $g^*(\mathscr{M}(t_{k(i)-1})) = g^{\pi^*}(\mathscr{M}(t_{k(i)-1}))$ where $\pi^* \in \Pi^*(M)$. By Assumption 5, we know that $g^*(\mathscr{M}(t_{k(i)-1}))$ only depends on $\mathscr{R}_z(t_{k(i)-1})$ and $\mathscr{P}_z(t_{k(i)-1})$ for $z \in \mathscr{Z}^{**}(M)$ where $N_z(t_{k(i)-1}) \geq \lambda t_{k(i)-1}$ by Lemma IV.14. Using Theorem II.1 to quantify the sensibility of the gain to kernel and reward perturbations, we get that

$$g^*(M) \leq g^*(\mathscr{M}(t_{k(i)-1})) \leq g^*(M) + \mathrm{O}\left(\sqrt{\frac{\log(t_{k(i)-1})}{t_{k(i)-1}}}\right) \tag{IV.4}$$

holds with probability one when $i \to \infty$.

Fix $t \in \{t_{k(i)}, \dots, t_{k(i)} + T - 1\}$. Recall that a policy that EVI outputs must have optimistic gain with span zero. Let π output by EVI at time $t' \in [t_{k(i)}, t]$, and assume that (1) π is sub-optimal in M from S_t, so that there exists $s \in \mathscr{S}$ such that $g^\pi(s; M) < g^*(s; M)$ and s is reachable from S_t under π; and (2) that $N_z(t) > N_z(t_{k(i)}) + C \log(T/\delta)$ for all $z \in \mathscr{Z}$, where C is given by the shrinking-shaking Lemmas IV.15 and IV.16. Without loss of generality, we can assume that s is recurrent under π on M and let $\mathscr{Z}' \subseteq \mathscr{Z}$ the associated recurrent component of pairs. By Assumption 5, we see that $g^\pi(s; \mathscr{M}(t))$ only depends on data on $\mathscr{Z}'$. Since π was output by EVI, $g^\pi(\mathscr{M}(t))$ only depends on data on $\mathscr{Z}'$. Let $\mathscr{Z}'_- := \mathscr{Z}' \setminus \mathscr{Z}^{**}(M)$ which is non-empty because $g^\pi(s; M) < g^*(s; M)$, and let $\mathscr{Z}'_+ := \mathscr{Z}' \cap \mathscr{Z}^{**}(M)$. We have:

$$\begin{aligned} g^\pi(\mathscr{M}(t)) &= \sup_{\tilde{r} \in \mathscr{R}_\pi(t)} \sup_{\tilde{p} \in \mathscr{P}_\pi(t)} g(r, p) = \sup_{\tilde{r} \in \mathscr{R}_{\mathscr{Z}'}(t)} \sup_{\tilde{p} \in \mathscr{P}_{\mathscr{Z}'}(t)} g(\tilde{r}, \tilde{p}) \\ &\overset{(\dagger)}{\leq} \sup_{\tilde{r} \in \mathscr{R}_{\mathscr{Z}'}(t_{k(i)-1})} \sup_{\tilde{p} \in \mathscr{P}_{\mathscr{Z}'}(t_{k(i)-1})} g\left(\tilde{r} - \frac{\log(T/\delta)}{\log(t_{k(i)})} \cdot e_{\mathscr{Z}'_-} + \sqrt{\frac{c \log(t_{k(i)})}{t_{k(i)}}} \cdot e_{\mathscr{Z}'_+}, \tilde{p} \right) \end{aligned}$$

$$\begin{aligned}
&\overset{(\ddagger)}{\leq} g^\pi(\mathscr{M}(t_{k(i)-1})) + \sqrt{\tfrac{c\log(t)}{t}} - \eta(M,\pi)\tfrac{\log(T/\delta)}{\log(t_{k(i)})} \\
&\sim g^\pi(\mathscr{M}(t_{k(i)-1})) - \tfrac{\eta(M,\pi)\log(T/\delta)}{\log(t_{k(i)})} \\
&\overset{(\text{IV.4})}{\leq} g^*(M) + \mathrm{O}\left(\sqrt{\tfrac{\log(t_{k(i)-1})}{t_{k(i)-1}}}\right) - \tfrac{\eta(M,\pi)\log(T/\delta)}{\log(t_{k(i)})} < g^*(M)
\end{aligned}$$

where the last inequality hold for $t_{k(i)}$ large enough. In the above, (†) holds on the events specified by the shrinking-shaking behavior of confidence regions, see Lemmas IV.15 and IV.16; and (‡) is a technical result on exit probabilities, stating that even though we take a supremum on $\tilde{p} \in \mathscr{P}_{\mathscr{Z}'}(t_{k(i)}-1)$, the choice of $\tilde{p}$ will put positive probability mass $\eta(M,\pi) > 0$ on $\mathscr{Z}'_-$ in its associated invariant probability measures.

This is justified as follows. On $\mathscr{Z}'_+ \equiv \mathscr{Z}' \cap \mathscr{Z}^{**}(M)$, the number of visits is $\omega(t_{k(i)-1})$ hence $\mathscr{P}_z(t_{k(i)-1})$ is nearly equal to $\{p_z\}$ for all $z \in \mathscr{Z}'_+$; In fact, for all fixed $\epsilon > 0$, we can assume that $\mathscr{P}_z(t_{k(i)-1}) \subseteq \{\tilde{p}_z : \|\tilde{p}_z - p_z\|_1 < \epsilon\}$ with overwhelming probability provided that $t_{k(i)-1}$ is large enough. Let $(\tilde{r}^\pi, \tilde{p}^\pi) \in \mathscr{M}^\pi(t_{k(i)-1})$ an optimistic model of π (see Corollary II.12) and let $\tilde{\mu}^\pi$ the empirical invariant measure of π starting from s under the optimistic model. Using that $\mathrm{sp}(g(\tilde{r}^\pi, \tilde{p}^\pi)) = 0$, we assume that $\tilde{p}^\pi$ has a single recurrent class $\mathscr{Z}''$ up to restricting to that class. By correctness of the confidence region, a policy output by EVI has optimistic gain higher than $g^*(M)$ and since the optimistic model is nearly equal to the true model on $\mathscr{Z}'_+$, we deduce that $\mathscr{Z}''$ must contain elements of $\mathscr{Z}'_-$ (otherwise π is optimal in M). We see that under $\tilde{p}^\pi$, for every element of $\mathscr{Z}'' \cap \mathscr{Z}'_+$ there must be a path to an element of $\mathscr{Z}'' \cap \mathscr{Z}'_-$ of length at most $|\mathscr{S}|-1$ and probability at least $c_\epsilon(M) := (\min_{z\in\mathscr{Z}'_+} \min\{p(s|z) > 0 : s \in \mathscr{S}\} - \epsilon)^{|\mathscr{S}|-1}$, which is well defined and positive for $\epsilon > 0$ small enough. So there must be $z \in \mathscr{Z}'' \cap \mathscr{Z}'_-$ such that $\tilde{\mu}(z) \geq |\mathscr{S}|^{-1} c_\epsilon(M)$. Set $\eta(M,\pi) := \frac{1}{2}c_0(M)$. For ϵ small enough and on mild concentration events, we have:

$$g\left(\tilde{r}^\pi - \tfrac{\log(T/\delta)}{\log(t_{k(i)})} \cdot e_{\mathscr{Z}'_-} + \sqrt{\tfrac{c\log(t_{k(i)})}{t_{k(i)}}} \cdot e_{\mathscr{Z}'_+}, , \tilde{p}^\pi\right) \leq g^\pi(\mathscr{M}(t_{k(i)-1})) + \sqrt{\tfrac{c\log(t)}{t}} - \eta(M,\pi)\tfrac{\log(T/\delta)}{\log(t_{k(i)})}.$$

This justifies (‡).

Overall, we have $g^\pi(\mathscr{M}(t)) < g^*(M) \leq g^*(\mathscr{M}(t))$ on the event $E_t := \bigcap_{z\in\mathscr{Z}} E_t^z$ with E_t^z given by:

$$\begin{cases}
\left(F^z_{t_{k(i)}}, \begin{bmatrix} \mathscr{P}_z(t) \subseteq \mathscr{P}_z(t_{k(i)-1}) \\ \text{or } N_z(t) \leq N_z(t_{k(i)}) + C\log\left(\frac{T}{\delta}\right) \end{bmatrix}, \begin{bmatrix} \sup \mathscr{R}_z(t) \leq \sup \mathscr{R}_z(t_{k(i)-1}) - \frac{N_z(t)-N_z(t_{k(i)})}{C\log(t_{k(i)})} \\ \text{or } N_z(t) \leq N_z(t_{k(i)}) + C\log\left(\frac{T}{\delta}\right) \end{bmatrix}\right) & \text{if } z \notin \mathscr{Z}^{**}(M) \\
\left(F^z_{t_{k(i)}}, d_{\mathrm{H}}(\mathscr{P}_z(t), \mathscr{P}_z(t_{k(i)-1})) \leq \sqrt{\tfrac{c\log(t)}{t}}, d_{\mathrm{H}}(\mathscr{R}_z(t), \mathscr{R}_z(t_{k(i)-1}))\right) & \text{if } z \in \mathscr{Z}^{**}(M)
\end{cases}$$

where, for $z \notin \mathscr{Z}^{**}(M)$, $F^z_{t_{k(i)}}$ is the event appearing in the shrinking effect lemma (Lemma IV.15), and for $z \in \mathscr{Z}^{**}(M)$, $F^z_{t_{k(i)}}$ is the event appearing in the shaking effect lemma (Lemma IV.16); In both cases, we have $\mathbf{P}(\exists i, \forall j \geq i : F^z_{t_{k(j)}}) = 1$ provided that the rate $\lambda > 0$ in the definition of $F^z_{t_{k(i)}}$ is chosen accordingly to the asymptotic regime of the algorithm (Lemma IV.14). We deduce that on E_t, π will be rejected as soon as (VM) triggers, because its optimistic gain is no more optimistically optimal. By (IV.3), as soon as a pair $z \notin \mathscr{Z}^{**}(M)$ is about to be visited for the second time in the episode, the episode will stop. We therefore have shown that while $g^\pi(S_t; M) < g^*(S_t; M)$ and on E_t, there exists $z \equiv (s,a)$ that is reachable from S_t under π such that $N_z(t) < N_z(t_k) + 1 + C\log(T/\delta)$ and $g^\pi(s; M) < g^*(s; M)$.

Accordingly, we have shown that the algorithm is $(E, t_{k(i)}, T, \varphi)$-coherent, with $\mathbf{P}(\exists t \in [t_{k(i)}, t_{k(i)} + T] : E_t^c) \leq \delta + \mathrm{o}(1)$ when $i \to \infty$ and $\varphi(t) = 1 + C\log(T/\delta)$. □

Proof of Theorem IV.12. Use the coherence property of Lemma IV.17 with $\delta = \frac{1}{T}$ and apply Lemma IV.13. We obtain:

$$(-) := \limsup_{i\to\infty} \mathbf{P}\big(\mathrm{Reg}(t_{k(i)}, t_{k(i)} + T) \geq x + C_4\varphi(t_{k(i)})\big)$$

$$\leq \limsup_{i\to\infty}\left\{\mathbf{P}\left(\mathrm{Reg}(t_{k(i)}, t_{k(i)}+T) \geq x + C_4\varphi(t_{k(i)}), \bigcap_{t=t_{k(i)}}^{t_{k(i)}+T-1} E_t\right) + \mathbf{P}\left(\bigcup_{t=t_{k(i)}}^{t_{k(i)}+T-1} E_t^c\right)\right\}$$
$$\leq \exp\left(-\frac{x}{C_2} + C_3\log(T) + \log(C_1)\right) + \frac{1}{T}$$

which is bounded by $\frac{2}{T}$ for $x \geq C_2(1+C_3)\log(T) + C_2\log(C_1)$, where C_1, C_2, C_3, C_4 are the constants provided by Lemma IV.13. Using that $\limsup_{i\to\infty}\varphi(t_{k(i)}) \leq 1 + 2C\log(T)$ and setting $\psi(T) := (C_2(1+C_3) + 2C_4C)\log(T) + C_2\log(C_1) + C_4$, we obtain:

$$\mathrm{RegExp}(T) \leq \limsup_{i\to\infty}\left\{\psi(T) + T\cdot\mathbf{P}\left(\mathrm{Reg}(t_{k(i)}, t_{k(i)}+T) \geq \psi(T)\right)\right\} \leq \psi(T) + 2. \tag{IV.5}$$

This concludes the proof of Theorem IV.12. □

13.3 Model dependent regret via coherence

In the proof of the regret of exploration guarantees, Lemma IV.13 is used twice and two different coherence properties are invoked. Coherence is first used in a *global* form to derive the almost sure asymptotic regime. Indeed, the first step of the proof (see Section 13.B) consists in showing that the algorithm is $((F_t), \lceil\log(T)\rceil, T, \varphi)$-coherent for $\varphi(\lceil\log(T)\rceil) = \mathrm{O}(\log(T))$ where the sequence of events (F_t) is asymptotically almost sure, i.e., $\mathbf{P}(\exists T, \forall t \geq T : F_t) = 1$. Then, coherence is used in a *local* form to derive the regret of exploration guarantees. Indeed, the whole point of Section 13.2.5 is to show that the algorithm is $(E, t_{k(i)}, T, \varphi)$-coherent where $(t_{k(i)})$ is the sequence of exploration episodes, $\mathbf{P}(\exists T, \forall t \geq T : E_t) = 1$ and $\varphi(t_{k(i)}) = \mathrm{O}(\log(T))$.

In this section, we show a third application of coherence properties: model dependent regret guarantees.

13.3.1 General model dependent regret bound via coherence

We provide first a general result.

Theorem IV.18. *Consider an episodic algorithm with* (1) *regenerative episodes and* (2) *such that there exists an adapted sequence of events* (F_t) *with* $\mathbf{P}(\bigcup_{t=T}^{\infty} F_t^c) = \mathrm{O}(\frac{1}{T})$ *such that the algorithm is* $((F_t), T, T, \varphi)$*-coherent for all* $T \geq 1$*. Then, for all non-degenerate model* M*,*

$$\mathbf{E}^M[\mathrm{Reg}(T)] = \mathrm{O}\left(\sum_{m=0}^{\lceil\log_2(T)\rceil-1} \varphi(2^m)\right) + \mathrm{O}(\log(T)) \tag{IV.6}$$

when $T \to \infty$*.*

Proof. Let $n := \lceil\log_2(T)\rceil$. For all $m \leq n$, the algorithm is $(F, 2^m, 2^m, \varphi)$-coherent, has regenerative episodes, and M is non-degenerate, so we invoke Lemma IV.13 and obtain, for $x \geq 0$,

$$(-) := \mathbf{P}\left(\mathrm{Reg}(2^m, 2^{m+1}) \geq x + C_4\varphi(2^n)\right)$$
$$\leq \mathbf{P}\left(\mathrm{Reg}(2^m, 2^{m+1}) \geq x + C_4\varphi(2^n), \bigcap_{t=2^m}^{2^{m+1}-1} F_t\right) + \mathbf{P}\left(\bigcup_{t=2^m}^{\infty} F_t^c\right)$$
$$\leq \exp\left(-\frac{x}{C_2} + C_3 m\log(2) + \log(C_1)\right) + \mathrm{O}\left(2^{-m}\right)$$

where C_1, C_2, C_3, C_4 are model dependent constants. For $x \geq C_2(C_1 + (1+C_3)\log(2)m)$, the RHS is $O(2^{-m})$. In other words, $\mathbf{E}[\mathrm{Reg}(2^m, 2^{m+1})] = O(\varphi(2^m))$. Summing for $m \geq 1$, we get:

$$\mathbf{E}[\mathrm{Reg}(T)] := \sum_{m=0}^{n-1} \mathbf{E}[\mathrm{Reg}(2^m, 2^{m+1})] = O\left(\sum_{m=0}^{n-1} \varphi(2^m) + 1\right) = O\left(\sum_{m=0}^{\lceil \log_2(T) \rceil - 1} \varphi(2^m)\right) + O(\log(T)).$$

This is the intended result. □

A few comments are in order. First, the requirement $\mathbf{P}(\bigcup_{t=T}^{\infty} F_t^c) = O(\frac{1}{T})$ is slightly overshoot and can be weakened depending on the asymptotic properties of φ and the desired bound. Second, the proof technique can be directly adapted to obtain bounds in probability rather than in expectation. Last, but perhaps the most important, is that this bound only holds for non-degenerate models (Definition IV.3). While every model can be made non-degenerate up to smooth reward perturbations, non-degenerate models are a bit special, because the weakly optimal pair is unique from every state (unique Bellman optimal policy), and $\mathscr{Z}^{**}(M)$ has a unique communicating component (unique gain optimal component), see Boone and Gaujal (2023a). The proof of Lemma IV.13, which is key here, inevitably relies on non-degeneracy. Yet, degenerate models are easy to find. Figure 12.1.1 is one simple example and is a good starting point to understand why coherence and regenerative episodes are insufficient to provide regret bounds on degenerate models.

13.3.2 A model dependent regret bound for (VM)

Theorem IV.18 is applied to EVI-based algorithms relying on (VM) by showing that such algorithms satisfy a $((F_t), T, T, \varphi)$-coherence property with a budget function $\varphi(T) = O(\log(T))$, leading to $O(\log(T)\log\log(T))$ regret bounds.

Theorem IV.19. *Let M a non-degenerate model. Consider running an EVI-based algorithm with confidence region $\mathscr{M}(t) \equiv \mathscr{M}_{\delta(t)}(t)$ as in Section 7.A.2 with $\delta(t) := \frac{1}{t}$, and assume that $\mathscr{M}(t)$ satisfies Assumption 5. If the algorithm manages episodes according to f-(VM) with $f > 0$, then*

$$\mathbf{E}^M[\mathrm{Reg}(T)] = O(\log(T)\log\log(T)). \tag{IV.7}$$

Proof. Consider the good events $E_t := (M \in \mathscr{M}(t))$ and $F_t := \bigcap_{t'=(t-|\mathscr{Z}|)/2}^{t} E_t$.

We know by design $\mathbf{P}(\bigcup_{t=T}^{\infty} E_t^c) = O(\frac{1}{T})$, see Section 6.3.2, so $\mathbf{P}(\bigcup_{t=T}^{\infty} F_t^c) = O(\frac{1}{T})$ as well. We show that the algorithm is (F_t, T, T, φ)-coherent for $\varphi(T) = O(\log(T))$. The result will then follow by Theorem IV.18 using that $\int \log(x)dx = x\log x - x$.

For all $\epsilon > 0$, it is easy to show that if $C \equiv C_\epsilon > 0$ is large enough with respect to ϵ, whether the confidence region is built out of ℓ_1-balls, KL-semi-balls or out of Bernstein's inequality, if $N_z(t) \geq C\log(t)$, we have:

$$\mathscr{P}_z(t) \subseteq \left\{\tilde{p}_z : \|\tilde{p}_z - \hat{p}_z(t)\|_1 < \tfrac{1}{2}\epsilon\right\} \quad \text{and} \quad \mathscr{R}_z(t) \subseteq \left\{\tilde{r}_z : \|\tilde{r}_z - \hat{r}_z(t)\|_\infty < \tfrac{1}{2}\epsilon\right\} \tag{IV.8}$$

Introduce the gain gap $\Delta_g := \min\{\|g^\pi(M) - g^*(M)\|_\infty : \pi \notin \Pi^*(M)\} > 0$. Whenever $M \in \mathscr{M}(t)$, we have $g^*(M) \leq g^*(\mathscr{M}(t))$. Let π a policy output by EVI at time t and assume that $N_z(t) \geq C\log(t)$ for all $z \in \mathscr{Z}$. It has optimistic bias with span at most $D(M)$, hence by Theorem II.1, we have:

$$\|g^\pi(\mathscr{M}_t) - g^\pi(M)\|_\infty \leq \epsilon\big(1 + \tfrac{1}{2}D(M)\big) \tag{IV.9}$$

yet $g^\pi(\mathscr{M}_t) \geq g^*(\mathscr{M}_t) \geq g^*(M)$. So, provided that $\epsilon(1 + \frac{1}{2}D(M)) < \Delta_g$, π necessarily achieves optimal gain. We assume from now on that $\epsilon(1 + \frac{1}{2}D(M)) < \Delta_g$ is true.

Now, assume that π_t is such that $g^{\pi_t}(S_t, M) < g^*(S_t, M)$. By construction of EVI-based algorithms, π_t is the output of EVI for t_k with $t \in [t_k, t_{k+1})$, hence is the optimistically optimal policy at time t_k. By assumption $g^{\pi_{t_k}}(S_t; M) < g^*(S_t; M)$, so assuming that

$$E_{t_k} \equiv (M \in \mathscr{M}(t_k)) \tag{IV.10}$$

holds, we deduce from the previous argument that there must be $z \in \mathscr{Z}$ such that $N_z(t_k) < C\log(t_k)$. Since $g^{\pi_{t_k}}(S_t, M) < g^*(S_t, M)$, $\mathrm{Reach}(\pi_{t_k}, S_t)$ must contain a recurrent component of π_{t_k} on which the achieved gain is sub-optimal. Pick one, denoted $\mathscr{Z}'$. Thanks to Assumption 5, the optimistic gain of $g^{\pi_{t_k}}(s, \mathscr{M}(t_k))$ for $s \in \mathscr{S}(\mathscr{Z}')$ only depends on pairs among $\mathrm{Reach}(\pi_{t_k}, s)$ and yet $g^{\pi_{t_k}}(s; \mathscr{M}(t_k)) \geq g^*(s, M)$. So there must be a sub-sampled pair in $\mathscr{Z}'$, i.e., there exists $(s,a) \in \mathscr{Z}'$ such that $N_{s,a}(t_k) < C\log(t_k)$; This pair is reachable from S_t under π_t and $g^{\pi_t}(s; M) < g^*(s; M)$ by construction of $\mathscr{Z}'$. Last, but not least, is that by construction of (VM), we have $t \leq 2t_k + |\mathscr{Z}|$ and $N_{s,a}(t) \leq 2N_{s,a}(t_k) + 1$. So, on the event $F_t := \bigcap_{t'=(t-|\mathscr{Z}|)/2}^{t} E_{t'}$,

$$\exists z \equiv (s,a) \in \mathrm{Reach}(\pi_t, S_t): \quad N_z(t) \leq 2C\log(t) + 1 \text{ and } g^{\pi_t}(s; M) < g^*(s; M). \tag{IV.11}$$

Setting $\varphi(t) := 2C\log(2t) + 1$, we have shown that the algorithm is $((F_t), T, T, \varphi)$-coherent. We have $\varphi(T) = \mathrm{O}(\log(T))$ and $\mathbf{P}(\bigcup_{t=T}^{\infty} F_t^c) = \mathrm{O}(\frac{1}{T})$. Conclude by applying Theorem IV.18. □

The result is remarkable in that f is basically arbitrary. It allows for $f(t)$ decreasing arbitrarily fast, hence for linearly many episodes, meaning that EVI-based algorithms can nearly be episode-less on non-degenerate models, at the expense of minimax guarantees (see Corollary IV.11). This remark is to be combined with the observation that EVI-based algorithm cannot be episode-less on degenerate models in general. This was discussed in Section 12.1 following an example from Ortner (2010). In tandem, this indicates that coherence alone cannot provide regret guarantees beyond non degenerate models. If the model dependent regret guarantees are obtained "for free" from coherence, extending such guarantees to degenerate models would require a different approach and most likely assumptions on the slackness function f.

From a high level viewpoint, we see that the proof of Theorem IV.19 can be generalized to any EVI-based algorithm with (1) not too long episodes, e.g., dominated by (DT) and (2) regenerative episodes. It is for instance directly applicable to (PT). Up to minor modifications of the proof, it can be adapted to PMEVI-based algorithms.

The extra $\log\log(T)$ is perhaps removable, but I have no proof direction in that matter.

13.4 Comments about (PT)

In Section 12.2.3, we have claimed that the present chapter would provide a technique to show that the performance test has sublinear regret of exploration. This technique is the coherence framework presented in Section 13.2, and consists in showing that (RPT) has consistency properties that are similar to (VM). The shrinking-shaking results (Lemmas IV.15 and IV.16) are general and hold similarly. However, one has to adapt the properties on the asymptotic regime (Lemma IV.14) to (RPT). This is nonetheless done similarly, because (VM) and (RPT) have very similar local and global almost-sure regret analysis.

For the asymptotic regime of (RPT), we follow the argument used in (VM) in Section 13.B. (1) By optimism, the algorithm only plays policies with optimistic gain larger than $g^*(M)$. But (2) there exists a constant $C > 0$ such that if every $z \notin \mathscr{Z}^{**}(M)$ has been visited at least $C\log(T)$ times, the optimistic gain of every sub-optimal policy is lower than $g^*(M)$. Combining (1) and (2) provides the coherence property that is used to derive the asymptotic regime of (RPT).

To adapt the local coherence property of (VM) established Section 13.2.5 to (RPT), remark that this property is obtained by showing that if sub-optimal pairs are visited enough in $[\tau, \tau + T)$,

then the optimistic gain of sub-optimal policies goes below $g^*(M)$ hence at the next episode update, such policies won't be picked again. This argument is however directly applicable to (RPT) as well. Almost by definition of (RPT) in fact. Therefore, the local coherence properties of (VM) and (RPT) are similar and both episode rules guarantee logarithmic regret of exploration for the same reasons.

The fact that the analysis of (VM) is almost directly applicable to (RPT) suggests that the two episode rules are perhaps instances of more general rule; Actually, the slackness function of (RPT) and the vanishing multiplicative factor of (VM) must satisfy the same conditions for minimax and regret of exploration guarantees! Nonetheless, the advantages of (VM) over (RPT) are that (VM) has minimax regret guarantees that are much easier to establish and computational advantages. With (RPT), optimistic gains must be monitored at all time while (VM) only updates optimistic gains when there is a chance that the optimistic policy changed.

13.5 Future directions

In the three previous chapters, we pointed the bad local behavior of EVI-based methods relying on the doubling trick and suggested a way to improve them. We have introduced two new episodes rules (PT) and (VM) that can be used to better manage episodes. To quantify the superiority of them over (DT), we introduced a new learning metric, the **regret of exploration**, shown that algorithms using (DT) have linear regret of exploration while (PT) and (VM) have logarithmic regret of exploration under technical assumptions on the underlying model.

A natural question is whether these assumptions (Assumption 5) may be dropped. This is not clear for reasons pointed out by (Lattimore and Szepesvári, 2020, §25.2) in their section entitled "*Clouds Looming for Optimism*." The policy-wise optimism of EVI-based makes policy played accordingly to their highest plausible gain. Sub-optimal policies are played under optimism because of a lack of information, but is playing the optimistic policy the best way to acquire information? In linear bandits for instance, this is not the case (Lattimore and Szepesvári, 2020, §25); And the behavior of EVI-based algorithms, that episodically play fixed deterministic policies, are in design miles away from the randomized nature of exploration described by the model dependent lower bound (Theorem III.5). In practice, when one takes a look at the optimistic models (Corollary II.12) produced by EVI, one sees that they systematically pick $\tilde{p}$ that does not satisfy $\tilde{p} \ll p$ when the confidence region allows it. It means that optimism tends to increase the set of recurrent states of policies, hence the optimistic policy π may be produced because of a lack of information on a pair that is not reachable under π. Remark that this is exactly the kind of behavior that coherence (Definition IV.5) requires to avoid. EVI-based algorithms have no protection against this nasty behavior, that they **will** display by design. The reason why they don't overcommit in iterating a policy π to overcome a lack of information on a region $\mathscr{X}' \subseteq \mathscr{X}$ that is not reachable using that policy, is that by iterating π, the reaching time to $\mathscr{X}'$ in $\mathscr{M}^\pi(t)$ explodes and eventually becomes larger than the diameter, hence π cannot be output by EVI anymore.

Another interesting direction, yet overwhelmingly technical perhaps, is to extend these regret of exploration of PMEVI-based algorithms. The analysis would have to quantify the evolution of the bias confidence region used by PMEVI (Chapter 7) over time. It probably also exhibits a shrinking-shaking behavior that would have to be quantified, and that would also further accelerate the rejection of sub-optimal policies, in theory at least. A genuine yet pertinent question is whether the regret of exploration of algorithms outside of the optimistic framework can be investigated. For instance, what can be said about Bayesian algorithms (PSRL Osband et al. (2013), TSDE Ouyang et al. (2017) or Optimistic PSRL Agrawal and Jia (2023))? Model dependent algorithms could be also considered, such as IMED-RL Pesquerel and Maillard (2022), IMED-KD Saber et al. (2024) or even ECoE (Chapter 10). Overall, the question of the

regret of exploration captures all algorithms in the existing literature.

Talking about model dependent algorithms, comes the question of the model dependent guarantees of EVI-based algorithms relying on (PT) or (VM). For instance, the original works of UCRL2 Auer et al. (2009) and KLUCRL Filippi et al. (2010) provides model dependent bound of these algorithms of order $O(\log(T))$. Via coherence again, Theorem IV.19 provides a partial answer, proving that (VM) provides guarantees of order $O(\log(T)\log\log(T))$ under a non-degeneracy assumption over the underlying model. Beyond that setting, the question of the model dependent regret guarantees of (PT) and (VM) are not easy, because contrary to the minimax guarantees, the proof techniques of the original (DT) cannot be borrowed directly. I leave a conjecture below.

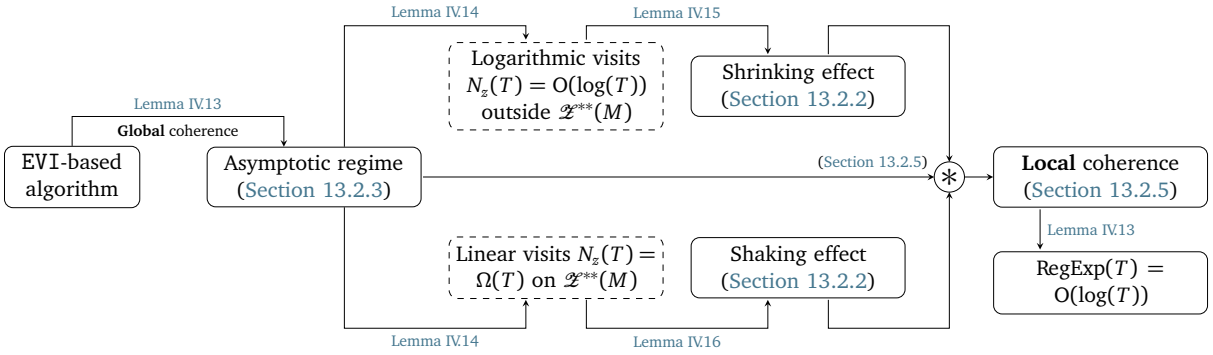
> **Conjecture.** UCRL2-(VM) has model dependent $O(\log(T)/f(T))$ in general; and $O(\log(T))$ if the underlying model is non-degenerate (Definition IV.3).

I believe however that the model dependent guarantees of EVI-based methods are not a good direction to investigate, because I believe that policy-wise optimism is different of nature from what the lower bound (Theorem III.5) says we should do.

Another interesting direction is to extend the local regret guarantees beyond exploration times. For instance, can $\sup_{t\to\infty} \mathbf{E}[\mathrm{Reg}(t, t+T)]$ grow sublinearly with T? Why did we even focus on exploration times in the first place? This direction is actually investigated in the next and last Chapter 14. There is a reason why the local regret guarantees of EVI-based methods stick to exploration times, and this is all the subject of Chapter 14.

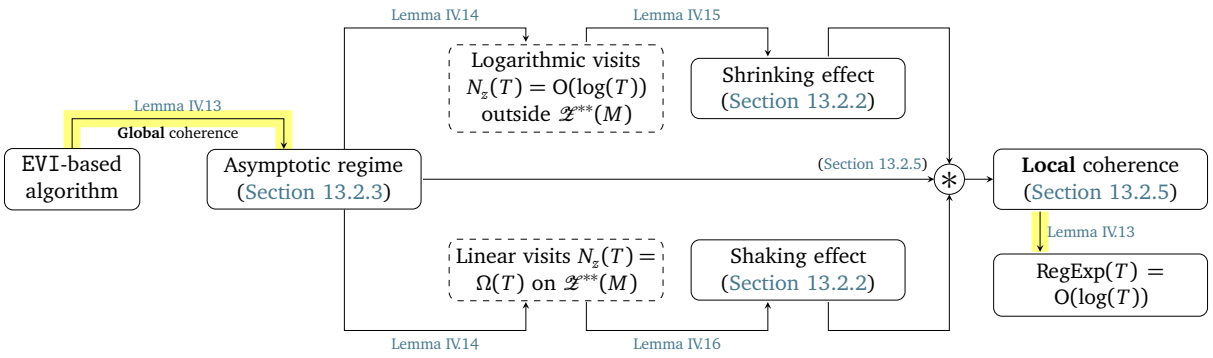
Appendix of Chapter 13

In this appendix, we provide the details to the coherence framework described in Section 13.1.2 and prove Theorem IV.12. Below is a map of the proof.



The map is reported throughout to track where we are in the proof.

13.A The coherence lemma: Proof of Lemma IV.13



13.A.1 Optimal/sub-optimal partioning of $[\tau, \tau + T)$

The time segment of interest $[\tau, \tau+T)$ is partioned into sub-segments $\biguplus_{i=1}^{I}[\tau_i, \tau_{i+1})$ as follows:

$$\tau_1 := \tau,$$
$$\tau_{i+1} := (\tau + T) \wedge \begin{cases} \inf\{t_k : t_k > \tau_i\} \\ \inf\{t > \tau_i : \mathbf{1}(g^{\pi_t}(S_t, M) = g^*(M)) \neq \mathbf{1}(g^{\pi_{\tau_i}}(S_{\tau_i}, M) = g^*(M))\} \end{cases}$$

and we write $i \in \mathscr{I}_{\text{opt}}$ if $g^{\pi_{\tau_i}}(S_{\tau_i}, M) = g^*(M)$ and $i \in \mathscr{I}_{\text{sub}}$ if $g^{\pi_{\tau_i}}(S_{\tau_i}) < g^*(M)$, that we refer to as **optimal** and **sub-optimal** segments. By design, every segment $[\tau_i, \tau_{i+1})$ is a subset of an episode and the sequence (τ_i) is a increasing sequence of stopping times. The regret is decomposed according to this partition:

$$\text{Reg}(\tau, \tau+T) = \sum_{i \in \mathscr{I}_{\text{sub}}} \sum_{t=\tau_i^0}^{\tau_{i+1}^0-1} \Delta_{Z_t} + \sum_{i \in \mathscr{I}_{\text{opt}}} \sum_{t=\tau_i^0}^{\tau_{i+1}^0-1} \Delta_{Z_t}. \tag{IV.12}$$

Both terms are bounded separately. The first corresponds to the regret on segments where the current policy is sub-optimal, while the second corresponds to the regret on segments where the current policy is asymptotically optimal.

13.A.2 Upper bounding the regret on sub-optimal segments

We have:

$$\sum_{i \in \mathscr{I}_{\text{sub}}} \sum_{t=\tau_i}^{\tau_{i+1}-1} \Delta_{Z_t} \le \Big(\max_{z \in \mathscr{Z}} \Delta_z\Big) \sum_{i \in \mathscr{I}_{\text{sub}}} (\tau_{i+1} - \tau_i). \tag{IV.13}$$

We bound $\sum_{i \in \mathscr{I}_{\text{sub}}} (\tau_{i+1} - \tau_i)$ directly.

(**STEP 1**) *There exists a constant $\epsilon > 0$ such that, on $\bigcap_{t=\tau}^{\tau+T-1} F_t$, we have:*

$$|\mathscr{Z}|(\varphi(\tau)+1) \ge \epsilon \sum_{i \in \mathscr{I}_{\text{sub}}} (\tau_{i+1} - \tau_i) + \sum_{i \in \mathscr{I}_{\text{sub}}} \sum_{t=\tau_i}^{\tau_{i+1}-1} \big(e_{S_{t+1}} - p(Z_t)\big) \sum_{z \in \mathscr{Z}} h^{\pi_{\tau_i}}(e_z, p) - \frac{|\mathscr{I}_{\text{sub}}|}{\epsilon} \tag{IV.14}$$

with $\text{sp}(\sum_{z \in \mathscr{Z}} h^{\pi_{\tau_i}}(e_z, p)) \le \frac{1}{\epsilon}$*, where $h^\pi(e_z, p)$ is the bias function of the policy π under the reward function e_z and kernel p. Moreover, ϵ can be chosen independently of F, τ, T and φ.*

Proof. Let $i \in \mathscr{I}_{\text{sub}}$ and fix $z \in \mathscr{Z}$. Because the segment $[\tau_i, \tau_{i+1})$ is a piece of episode, π_{τ_i} is used all throughout the segment. The gain and bias functions of π_{τ_i} on the model with reward function e_z (equal to one at z and null elsewhere) and kernel p are respectively denoted $g^{\pi_{\tau_i}}(-; e_z, p)$ and $h^{\pi_{\tau_i}}(-; e_z, p)$. Using the Poisson equation, we obtain:

$$\begin{aligned}
(-) &:= N_z(\tau_{i+1}) - N_z(\tau_i) \\
&= \sum_{t=\tau_i}^{\tau_{i+1}-1} g^{\pi_{\tau_i}}(S_t; e_z, p) + \mathbf{E}\big[h^{\pi_{\tau_i}}(S_{\tau_i}; e_z, p) - h^{\pi_{\tau_i}}(S_{\tau_{i+1}}; e_z, p)\big] + \sum_{t=\tau_i}^{\tau_{i+1}-1} \big(e_{S_{t+1}} - p(Z_t)\big) h^{\pi_{\tau_i}}(e_z, p) \\
&\ge \sum_{t=\tau_i}^{\tau_{i+1}-1} g^{\pi_{\tau_i}}(S_t; e_z, p) + \sum_{t=\tau_i}^{\tau_{i+1}-1} \big(e_{S_{t+1}} - p(Z_t)\big) h^{\pi_{\tau_i}}(e_z, p) - \frac{1}{\epsilon}
\end{aligned}$$

where ϵ is any positive quantity smaller than $(\max_\pi \max_z \text{sp}(h^\pi(e_z, p)))^{-1} > 0$.

Let $\mathscr{I}_{\text{sub}}^z := \{i \in \mathscr{I}_{\text{sub}} : z \in \text{Reach}(\pi_{\tau_i}, S_{\tau_{i+1}-1})\}$.

Because the segment $[\tau_i, \tau_{i+1})$ is a piece of episode, π_{τ_i} is used all throughout the segment hence a pair that is reachable at time $\tau_{i+1}-1$ is necessarily reachable during the entire segment. Therefore, if $i \in \mathscr{I}_{\text{sub}}^z$, then $g^{\pi_{\tau_i}}(S_t; e_z, p) > 0$ for all $t \in [\tau_i, \tau_{i+1}-1)$. Further assume that ϵ is smaller than $\min\{g^\pi(s; e_z, p) : z \in \text{Reach}(\pi, s), s \in \mathscr{S}, \pi \in \Pi\} > 0$. We obtain:

$$N_z(\tau_{i+1}) - N_z(\tau_i) \ge \epsilon(\tau_{i+1} - \tau_i) + \sum_{t=\tau_i}^{\tau_{i+1}-1} \big(e_{S_{t+1}} - p(Z_t)\big) h^{\pi_{\tau_i}}(e_z, p) - \frac{1}{\epsilon}.$$

Summing for i provides

$$\max_{i\in\mathscr{I}_{\rm sub}} N_z(\tau_{i+1})-N_z(\tau)\geq \epsilon\sum_{i\in\mathscr{I}^z_{\rm sub}}(\tau_{i+1}-\tau_i)+\sum_{i\in\mathscr{I}^z_{\rm sub}}\sum_{t=\tau_i}^{\tau_{i+1}-1}\big(e_{S_{t+1}}-p(Z_t)\big)h^{\pi_{\tau_i}}(e_z,p)-\frac{|\mathscr{I}^z_{\rm sub}|}{\epsilon}.$$

Recall that for $i\in\mathscr{I}_{\rm sub}$, the segment last until the next episode and $g^{\pi_{\tau_i}}(S_t,M)<g^*(M)$ holds for all $t\in[\tau_i,\tau_{i+1})$. Meanwhile, coherence guarantees that, on $\bigcap_{t=\tau}^{\tau+T-1}F_t$, we have $N_z(\tau_{i+1})\leq N_z(\tau)+\varphi(\tau)+1$ for all $i\in\mathscr{I}_{\rm sub}$ and $z\notin\mathscr{Z}^*(M)$. So, for all $z\notin\mathscr{Z}^*(M)$ and on $\bigcap_{t=\tau}^{\tau+T-1}F_t$, we have

$$\varphi(\tau)+1\geq \epsilon\sum_{i\in\mathscr{I}^z_{\rm sub}}(\tau_{i+1}-\tau_i)+\sum_{i\in\mathscr{I}^z_{\rm sub}}\sum_{t=\tau_i}^{\tau_{i+1}-1}\big(e_{S_{t+1}}-p(Z_t)\big)h^{\pi_{\tau_i}}(e_z,p)-\frac{|\mathscr{I}^z_{\rm sub}|}{\epsilon}.$$

By coherence and on $\bigcap_{t=\tau}^{\tau+T-1}F_t$ again, we see that $i\in\mathscr{I}_{\rm sub}$ belongs to one $\mathscr{I}^z_{\rm sub}$ for some $z\notin\mathscr{Z}^*(M)$ at least. Summing for $z\notin\mathscr{Z}^*(M)$, we obtain the claim. □

(**STEP 2**) *There exists a constant $\eta>0$ such that*

$$\forall x\geq 0,\quad \mathbf{P}\left(|\mathscr{I}_{\rm sub}|\geq x+\frac{1}{\eta}\varphi(\tau)\ \textit{and}\ \bigcap_{t=\tau}^{\tau+T-1}F_t\right)\leq\exp(-\eta x). \tag{IV.15}$$

Moreover, η can be chosen independently of F,τ,T and φ.

Proof. Denote $\mathscr{T}_{\rm sub}(\tau,\tau+T):=\bigcup_{i\in\mathscr{I}_{\rm sub}}[\tau_i,\tau_{i+1})$ the time instants when $g^{\pi_t}(S_t,M)<g^*(M)$.

Introduce the quantity $\phi(t):=\sum_z[\varphi(\tau)+N_z(\tau)-N_z(t)]_+$ for $t\in[\tau,\tau+T)$, which is non-increasing by construction. By coherence and on F_t, if $t\in\mathscr{T}_{\rm sub}(\tau,\tau+T)$ then there exists a reachable z such that $\varphi(\tau)+N_z(\tau)-N_z(t)>0$. The crucial remark is that for $i\in\mathscr{I}_{\rm sub}$ with $[\tau_i,\tau_{i+1})\subseteq[t_k,t_{k+1})$, two things may hold at time τ_{i+1}: (1) Either $i+1\in\mathscr{I}_{\rm opt}$, meaning that a state from which π_{τ_i} is optimal has been reached; (2) Or $i+1\notin\mathscr{I}_{\rm sub}$ and $\tau_{i+1}=t_{k+1}$, in which case $S_{\tau_{i+1}}$ has been already visited since τ_i. For (2), remark indeed that $S_{\tau_{i+1}}$ has been visited already since t_k by regenerativity of episodes (Definition IV.6), but if $t_k\neq\tau_i$ then $g^{\pi_{\tau_i}}(S_t,M)=g^*(M)$ for all $t\in[t_k,\tau_i)$ hence $S_{t_{k+1}}$ cannot appear within the collection of states visited in the time-range $[t_k,\tau_i)$. Combining (1) and (2), we conclude that conditionally on the history O_{τ_i}, every reachable pair $z\in{\rm Reach}(\pi_{\tau_i},S_{\tau_i})$ from which π_{τ_i} is sub-optimal have positive probability $\epsilon(S_{\tau_i},\pi_{\tau_i},z,M)$ to be visited until τ_{i+1}. Letting $\epsilon:=\min_{s,\pi,z}\epsilon(s,\pi,z,M)>0$, we get:

$$\begin{aligned}(-)&:=\mathbf{P}\big(\phi(\tau_{i+1})<\phi(\tau_i)\mid O_{\tau_i},i\in\mathscr{I}_{\rm sub},F_{\tau_i}\big)\\&\geq\min_{\substack{z\equiv(s,a)\in{\rm Reach}(S_{\tau_i},\pi_{\tau_i})\\ g^{\pi_{\tau_i}}(s,M)<g^*(M)}}\mathbf{P}\left(N_z(\tau_{i+1})>N_z(\tau_i)\middle|O_{\tau_i},i\in\mathscr{I}_{\rm sub},F_{\tau_i}\right)\\&\geq\epsilon.\end{aligned}$$

Let $\phi_0(\tau):=SA\varphi(\tau)$ and denote $F_{\tau:\tau+T}:=\bigcap_{t=\tau}^{\tau+T-1}F_t$. On $F_{\tau:\tau+T}$, ϕ can only decrease up to $\phi_0(\tau)$ times before reaching zero, and once it has reached zero, we cannot have $t\in\mathscr{T}_{\rm sub}(\tau,\tau+T)$ anymore. Accordingly, for all $m\geq1$, $|\mathscr{I}_{\rm sub}|\geq m+\phi_0(\tau)$ implies on $F_{\tau:\tau+T}$ that the first in the first $m+\phi_0(\tau)$ elements of $\mathscr{I}_{\rm sub}$, at least m of them are such that $\phi(\tau_{i+1})=\phi(\tau_i)$. Introduce the short-hand $U_{\tau_i}:=\mathbf{1}(\phi(\tau_{i+1})=\phi(\tau_i))$. For $\lambda>0$ and $m\geq1$, we have:

$$\psi(m):=\mathbf{P}(|\mathscr{I}_{\rm sub}|\geq m+\phi_0(\tau)\text{ and }F_{\tau:\tau+T})$$

$$
\begin{aligned}
&= \mathbf{P}\left(\sum_{j=1}^{m+\phi_0(\tau)} U_{\tau_j} \geq m \text{ and } F_{\tau:\tau+T} \right) \\
&= \mathbf{E}\left[\mathbf{1}(\exp\left(\lambda \sum_{j=1}^{m+\phi_0(\tau)} U_{\tau_j} \right) \geq \exp(\lambda m)) \mathbf{1}(F_{\tau:\tau+T}) \right] \\
&\leq \exp(-\lambda m)\mathbf{E}\left[\exp\left(\lambda \sum_{j=1}^{m+\phi_0(\tau)} U_{\tau_j} \right) \mathbf{1}(F_{\tau:\tau+T}) \right] \\
&\overset{(\dagger)}{\leq} \exp(-\lambda m)\mathbf{E}\left[\exp\left(\lambda \sum_{j=1}^{m+\phi_0(\tau)-1} U_{\tau_j} \right) \mathbf{1}(F_{\tau:\tau_{m+\phi_0(\tau)}}) \cdot \mathbf{1}(F_{\tau_{m+\phi_0(\tau)}}) \mathbf{E}\Big[\exp(\lambda U_{\tau_m+\phi_0(\tau)}) \Big| F_{\tau_{m+\phi_0(\tau)}} \Big] \right] \\
&\overset{(\ddagger)}{\leq} \exp(-\lambda m)\mathbf{E}\left[\exp\left(\lambda \sum_{j=1}^{m+\phi_0(\tau)-1} U_{\tau_j} \right) \mathbf{1}(F_{\tau:\tau_{m+\phi_0(\tau)}}) \cdot \exp\left(\lambda(1-\epsilon) + \frac{\lambda^2}{8} \right) \right] \\
&\vdots \\
&\leq \exp\left(-\lambda m + \lambda(1-\epsilon)(m+\phi_0(\tau)) + (m+\phi_0(\tau))\frac{\lambda^2}{8} \right).
\end{aligned}
$$

In the above, (†) use that $\mathbf{1}(F_{\tau:\tau_{m+\phi_0}}) \cdot \mathbf{1}(F_{\tau_{m+\phi_0}}) \leq \mathbf{1}(F_{\tau:\tau+T})$ and (‡) is an application of Hoeffding's Lemma together with the fact that $\mathbf{E}[U_{\tau_i}|F_{\tau_i}]\mathbf{1}(F_{\tau_i}) \leq 1-\epsilon$. Assume that m is large enough so that $\epsilon m > (1-\epsilon)\phi_0(\tau)$. Then we continue by factorizing the polynomial within the exponential and minimizing in λ, straight forward algebra shows that for $m \geq \frac{2\phi_0(\tau)}{\epsilon}$, we have:

$$
\mathbf{P}(|\mathscr{I}_{\text{sub}}| \geq m + \phi_0(\tau) \text{ and } F_{\tau:\tau+T}) \leq \exp\left(-\frac{3\epsilon^2 m}{4} \right). \tag{IV.16}
$$

We conclude accordingly by choosing $\eta = \Theta(1 + \frac{2}{\epsilon})$. □

(**STEP 3**) *There exists constants $C_0, C_1, C_2, C_3 > 0$ such that*

$$
\forall x \geq 0, \quad \mathbf{P}\left(\sum_{i \in \mathscr{I}_{\text{sub}}} (\tau_{i+1} - \tau_i) > x + C_3\varphi(\tau) \text{ and } \bigcap_{t=\tau}^{\tau+T-1} F_t \right) \leq C_1 T^{C_2} \exp(-C_0 x). \tag{IV.17}
$$

Moreover, C_0, C_1, C_2, C_3 can be chosen independently of F, τ, T and φ.

Proof. Using a time uniform Azuma-Hoeffding's inequality (Lemma I.22), we have:

$$
\forall \delta > 0, \quad \mathbf{P}\left(\sum_{i \in \mathscr{I}_{\text{sub}}} \sum_{t=\tau_i}^{\tau_{i+1}-1} \big(e_{S_{t+1}} - p(Z_t) \big) \sum_{z \in \mathscr{Z}} h^{\pi_{\tau_i}}(e_z, p) < -\frac{1}{\epsilon} \sqrt{ \sum_{i \in \mathscr{I}_{\text{sub}}} (\tau_{i+1} - \tau_i) \log\big(\tfrac{T}{\delta}\big) } \right) \leq \delta.
$$

Combined with (IV.14) from (**STEP 1**), we obtain an equation of the form $x \leq \alpha + \beta\sqrt{x}$ with $x = \sum_{i \in \mathscr{I}_{\text{sub}}} (\tau_{i+1} - \tau_i)$, $\alpha = \frac{1}{\epsilon}(|\mathscr{Z}|(\varphi(\tau)+1) + \frac{1}{\epsilon}|\mathscr{I}_{\text{sub}}|)$ and $\beta = \frac{1}{\epsilon}\sqrt{\log(T/\delta)}$. Simple algebra shows that $x \leq 2\alpha + 2\beta^2$. In other words, we have shown that:

$$
\forall \delta > 0, \quad \mathbf{P}\left(\sum_{i \in \mathscr{I}_{\text{sub}}} (\tau_{i+1} - \tau_i) > C_0 \log\big(\tfrac{T}{\delta}\big) + C_1\varphi(\tau) + C_2|\mathscr{I}_{\text{sub}}| \text{ and } \bigcap_{t=\tau}^{\tau+T-1} F_t \right) \leq \delta
$$

for some model dependent constants $C_0, C_1, C_2 > 0$. Use the sub-exponential tail property of $|\mathscr{I}_{\text{sub}}|$ (IV.15) from (**STEP 2**) to obtain a sub-exponential tail for $\sum_{i \in \mathscr{I}_{\text{sub}}} (\tau_{i+1} - \tau_i)$. □

(**STEP 4**) *There exist constants $C_0, C_1, C_2, C_3 > 0$ such that, for all $\eta > 0$,*

$$\mathbf{P}\left(\sum_{j \in \mathscr{J}^+_{\mathrm{sub}}} \sum_{t=\tau^+_j}^{\tau^+_{j+1}-1} \Delta_{Z_t} > x + C_3 \varphi(\tau) \text{ and } \bigcap_{t=\tau}^{\tau+T} F_t \right) \leq C_1 T^{C_2} \exp(-C_0 x). \tag{IV.18}$$

Moreover, C_0, C_1, C_2, C_3 can be chosen independently of F, τ, T and φ.

Proof. Combine (IV.13) with the result of (**STEP 3**). □

13.A.3 Upper bounding the regret on optimal segments

We start by merging consecutive optimal segments. This is done by setting:

$$\begin{aligned} \tau^+_1 &:= \inf\{\tau_i : i \in \mathscr{I}_{\mathrm{opt}}\} \\ \tau^+_{2j} &:= \inf\{\tau_i > \tau^+_{2j-1} : i \in \mathscr{I}_{\mathrm{sub}}\} \\ \tau^+_{2j+1} &:= \inf\{\tau_i > \tau^+_{2j} : i \in \mathscr{I}_{\mathrm{opt}}\} \end{aligned} \tag{IV.19}$$

that design a macroscopic decomposition of $[\tau, \tau+T-1)$ into time-segments, of which (τ_i) is a refinement. Remark that if j is even, then $[\tau^+_j, \tau^+_{j+1}) \subseteq \bigcup_{i \in \mathscr{I}_{\mathrm{sub}}} [\tau_i, \tau_{i+1})$ and conversely, if j is odd, then $[\tau^+_j, \tau^+_{j+1}) \setminus \bigcup_{i \in \mathscr{I}_{\mathrm{sub}}} [\tau_i, \tau_{i+1}) = \varnothing$. We write $j \in \mathscr{J}^+_{\mathrm{sub}}$ and $j \in \mathscr{J}^+_{\mathrm{opt}}$ respectively.

By non-degeneracy of the model M, all asymptotically optimal policies of M have the same (unique) invariant probability measure $\mu^* \in \mathscr{P}(\mathscr{Z})$. On segments $[\tau^+_j, \tau^+_{j+1})$ with $j \in \mathscr{J}^+_{\mathrm{opt}}$, Δ_{Z_t} can only be positive if the optimal recurrent states $\mathscr{S}(\mathrm{supp}(\mu^*))$ have not been reached yet. The proof consists in showing that when $j \in \mathscr{J}^+_{\mathrm{opt}}$, the optimal recurrent class is quickly reached on $[\tau^+_j, \tau^+_{j+1})$. Indeed, setting $\tau^*_{j+1} := \tau^+_{j+1} \wedge \inf\{t > \tau^+_j : \mu^*(S_t) > 0\}$ the reaching time to $\mathrm{supp}(\mu^*)$ after τ^+_j, we have:[1]

$$\sum_{j \in \mathscr{J}^+_{\mathrm{opt}}} \sum_{t=\tau^+_j}^{\tau^+_{j+1}-1} \Delta_{Z_t} \leq \left(\max_{z \in \mathscr{Z}} \Delta_z \right) \sum_{j \in \mathscr{J}^+_{\mathrm{opt}}} \left(N_{\mathscr{Z}^-(M)}(\tau^+_{j+1}) - N_{\mathscr{Z}^-(M)}(\tau^+_j) \right) = \left(\max_{z \in \mathscr{Z}} \Delta_z \right) \sum_{j \in \mathscr{J}^+_{\mathrm{opt}}} \left(\tau^*_{j+1} - \tau^+_j \right). \tag{IV.20}$$

We now upper bound the RHS.

(**STEP 1**) *There exists a constant $D_* > 0$ as well as an adapted sequence (h_t) with $\mathrm{sp}(h_t) \leq D^*$ s.t.:*

$$\sum_{j \in \mathscr{J}^+_{\mathrm{opt}}} \left(\tau^*_{j+1} - \tau^+_j \right) \leq 2D^* \left(\left| \mathscr{J}^+_{\mathrm{opt}} \right| + \sum_{j \in \mathscr{J}^+_{\mathrm{opt}}} \left| \left\{ t_\ell \in (\tau^+_j, \tau^+_{j+1}) : \mu^*(S_{t_\ell}) = 0 \right\} \right| \right) + \sum_{j \in \mathscr{J}^+_{\mathrm{opt}}} \sum_{t=\tau^+_j}^{\tau^*_j - 1} (e_{S_{t+1}} - p_{Z_t}) h_t.$$

Moreover, D_* is independent of F, τ, T and φ.

Proof. Notice that $[\tau^+_j, \tau^+_{j+1})$ is of the form $[t'_k, t_{k+1}) \uplus \biguplus_\ell [t_\ell, t_{\ell+1})$ where $[t'_k \in [t_k, t_{k+1}]$ is a time such that $g^{\pi_{t-1}}(S_{t-1}; M) < g^{\pi_t}(S_t; M) = g^*(S_t; M)$. Consider the reward function $f(z) := \mathbf{1}(z \notin \mathscr{Z}^*(M))$. Over an episode $[t_\ell, t_{\ell+1}) \subseteq [\tau^+_j, \tau^+_{j+1})$, the gain and the bias of π^ℓ associated to this reward function are respectively denoted $g^{(\ell)}$ and $h^{(\ell)}$. Because the recurrent pairs under π^ℓ from S_{t_ℓ} are $\mathrm{supp}(\mu^*)$, we have $g^{(\ell)}(s) = 0$ for all $(s,a) \in \mathrm{Reach}(S_{t_\ell}, \pi^\ell, M)$ and $h^{(\ell)}(s) = 0$

[1] μ is a measure on $\mathscr{Z}$. For $s \in \mathscr{S}$, we write $\mu(s) := \sum_{a \in \mathscr{A}(s)} \mu(s,a)$.

for all $(s,a) \in \mathrm{supp}(\mu^*)$. Let $D^* < \infty$ the maximum $\mathrm{sp}(h^{(\ell)})$ possible over all $\pi^\ell \in \Pi$. Using Poisson's equation $g^{(\ell)}(s) + h^{(\ell)}(s) = f(s,\pi^\ell(s)) + p(s,\pi^\ell(s))h^{(\ell)}$, we obtain:

$$\begin{aligned}
\tau^*_{j+1} - \tau^+_j &= N_{\mathscr{Z}^-(M)}(\tau^+_{j+1}) - N_{\mathscr{Z}^-}(\tau^+_j) \\
&= \sum_\ell \big(h^{(\ell)}(S_{t_\ell}) - h^{(\ell)}(S_{t_{\ell+1}})\big) + \sum_\ell \sum_{t=t_\ell}^{t_{\ell+1}-1} \big(e_{S_{t+1}} - p_{S_t,A_t}\big)h^{(\ell)} \\
&\le 2D^* + \sum_{\ell : t_\ell \in (\tau^+_j, \tau^+_{j+1})} \big(h^{(\ell)}(S_{t_\ell}) - h^{(\ell)}(S_{t_{\ell+1}})\big) + \sum_\ell \sum_{t=t_\ell}^{t_{\ell+1}-1} \big(e_{S_{t+1}} - p_{S_t,A_t}\big)h^{(\ell)} \\
&\overset{(\dagger)}{=} 2D^* + \sum_{\ell>k} \mathbf{1}(t_\ell < \tau^*_j)\big(h^{(\ell)}(S_{t_\ell}) - h^{(\ell)}(S_{t_{\ell+1}})\big) + \sum_\ell \sum_{t=t_\ell}^{t_{\ell+1}-1} \mathbf{1}(t < \tau^*_j)\big(e_{S_{t+1}} - p_{S_t,A_t}\big)h^{(\ell)} \\
&\overset{(\ddagger)}{\le} 2D^*\Big(1 + \Big|\Big\{t_\ell \in (\tau^+_j, \tau^+_{j+1}) : \mu^*(S_{t_\ell}) = 0\Big\}\Big|\Big) + \sum_{t=\tau_j}^{\tau^*_j - 1} \big(e_{S_{t+1}} - p_{S_t,A_t}\big)h_t
\end{aligned}$$

where (†) follows from $h^{(\ell)} = 0$ on the support of μ^*, and (†) introduces h_t as the unique $h^{(\ell)}$ such that $t \in [t_\ell, t_{\ell+1})$. Conclude by summing over $i \in \mathscr{J}^+_{\mathrm{opt}}$. □

(**STEP 2**) *There exist constants $C_1, C_2, C_3, C_4 > 0$ such that, for all $\eta > 0$,*

$$\forall x \ge 0, \quad \mathbf{P}\left(\sum_{j \in \mathscr{J}^+_{\mathrm{opt}}} \left(\tau^*_{j+1} - \tau^+_j\right) > x + C_4\varphi(\tau) \text{ and } \bigcap_{t=\tau}^{\tau+T} F_t\right) \le C_1 T^{C_2} \exp(-C_3 x). \tag{IV.21}$$

Moreover, C_1, C_2, C_3, C_4 can be chosen independently of F, τ, T and φ.

Proof. We bound every term appearing in (**STEP 1**).

The **first term** involves $|\mathscr{J}^+_{\mathrm{opt}}|$. Because elements of $\mathscr{J}^+_{\mathrm{opt}}$ and $\mathscr{J}^+_{\mathrm{sub}}$ are intertwined, we $|\mathscr{J}^+_{\mathrm{opt}}| \le 1 + |\mathscr{J}^+_{\mathrm{sub}}|$. Moreover, since macroscopic segments $[\tau^+_j, \tau^+_{j+1})$ are unions of segments $[\tau_i, \tau_{i+1})$, we have $|\mathscr{J}^+_{\mathrm{sub}}| \le |\mathscr{I}_{\mathrm{sub}}|$ that has been bounded in (IV.15) already. Accordingly, $|\mathscr{J}^+_{\mathrm{sub}}|$ has sub-exponential tails on the good event $\bigcap_{t=\tau}^{\tau+T-1} F_t$:

$$\forall x \ge 0, \quad \mathbf{P}\left(\left|\mathscr{J}^+_{\mathrm{sub}}\right| \ge x + \frac{1}{c}\varphi(\tau) \text{ and } \bigcap_{t=\tau}^{\tau+T-1} F_t\right) \le \exp(-cx) \tag{IV.22}$$

where $c > 0$ is a model dependent constant.

For the **second term**, remark that for each $t_\ell \in [\tau_j, \tau_{j+1})$ with $j \in \mathscr{J}^+_{\mathrm{opt}}$, the probability that the episode ends with $S_{t_{\ell+1}} \in \mathrm{supp}(\mu^*)$ is positive because of the regenerativity property (Definition IV.6) of (VM). This is also true for the first (possibly) truncated episode $[t'_k, t_{k+1})$ that starts the macroscopic segment $[\tau^+_j, \tau^+_{j+1})$ because as the gain $g^{\pi_t}(S_t; M)$ increases from $t'_k - 1$ to t'_k to the optimal $g^{\pi_t}(S_t, M) = g^*(S_t; M)$, all states that are reachable from S_t under π_t cannot have been visited yet during the episode. In the end, the probability of reaching $\mathrm{supp}(\mu^*)$ by the end of the episode is lower bounded by some $\epsilon'(\pi_{t_\ell}, S_{t_\ell}, M) > 0$ and we denote $\epsilon' > 0$ the minimum for all possible values of π^ℓ and S_{t_ℓ}. We conclude that $\mathbf{P}(\mu^*(S_{t_{\ell+1}}) > 0 \mid O_{t_\ell}) > \epsilon'$. Accordingly,

$$U_{\tau^+_j} := \Big|\Big\{t_\ell \in (\tau^+_j, \tau^+_{j+1}) : \mu^*(X_{t_\ell}) = 0\Big\}\Big|$$

is stochastically dominated by a geometric distribution $\mathrm{G}(\epsilon')$. Using bounds on tails of geometric random variables (Lemma I.29), we obtain:

$$\mathbf{P}\left(\sum_{j\in\mathscr{J}^+_{\mathrm{opt}}}\left|\left\{t_\ell\in(\tau^+_j,\tau^+_{j+1}):\mu^*(S_{t_\ell})=0\right\}\right|>\left|\mathscr{J}^+_{\mathrm{opt}}\right|\left(1+\tfrac{2}{\epsilon'}\right)+\tfrac{2\eta\log(T)}{\log\left(\frac{1}{1-\epsilon'}\right)}\right)\leq T^{-\eta}. \tag{IV.23}$$

The **third term** $\sum_{j\in\mathscr{J}^+_{\mathrm{opt}}}\sum_{t=\tau_j}^{\tau^*_{j+1}-1}(e_{S_{t+1}}-p_{Z_t})h_t$ is the sum of a martingale difference sequence, each term having span at most D^* by (**STEP 1**). By applying a time-uniform Azuma-Hoeffding's inequality (Lemma I.22), we obtain:

$$\mathbf{P}\left(\sum_{j\in\mathscr{J}^+_{\mathrm{opt}}}\sum_{t=\tau_j}^{\tau^*_{j+1}-1}(e_{S_{t+1}}-p_{Z_t})h_t>D^*\sqrt{\sum_{j\in\mathscr{J}^+_{\mathrm{opt}}}\left(\tau^*_{j+1}-\tau^+_j\right)\left(\tfrac{1}{2}+\eta\right)\log(1+T)}\right)\leq T^{-\eta}. \tag{IV.24}$$

Combining the bound of the first term, (IV.23) and (IV.24), we see that there exists C_1,C_2,C_3,C_4 such that for all $\eta>0$, with probability $1-3T^{-\eta}$,

$$\sum_{j\in\mathscr{J}^+_{\mathrm{opt}}}\left(\tau^*_{j+1}-\tau^+_j\right)\leq C_1+(C_2+\eta C_3)(\log(T)+\varphi(\tau))+C_4\sqrt{\sum_{j\in\mathscr{J}^+_{\mathrm{opt}}}\left(\tau^*_{j+1}-\tau^+_j\right)\left(\tfrac{1}{2}+\eta\right)\log(T)}.$$

This is an equation of the form $x\leq\alpha+\beta\sqrt{x}$ that implies in particular $x\leq 2(\alpha+\beta^2)$. We conclude by rearranging terms of the equation. □

(**STEP 3**) *There exist constants $C_1,C_2,C_3,C_4>0$ such that, for all $\eta>0$,*

$$\mathbf{P}\left(\sum_{j\in\mathscr{J}^+_{\mathrm{opt}}}\sum_{t=\tau^+_j}^{\tau^+_{j+1}-1}\Delta_{Z_t}>x+C_4\varphi(\tau)\text{ and }\bigcap_{t=\tau}^{\tau+T}F_t\right)\leq C_1T^{C_2}\exp(-C_3x). \tag{IV.25}$$

Moreover, C_1,C_2,C_3,C_4 can be chosen independently of F,τ,T and φ.

Proof. Invoke (IV.20) and apply the result of (**STEP 2**). □

13.A.4 Combining everything

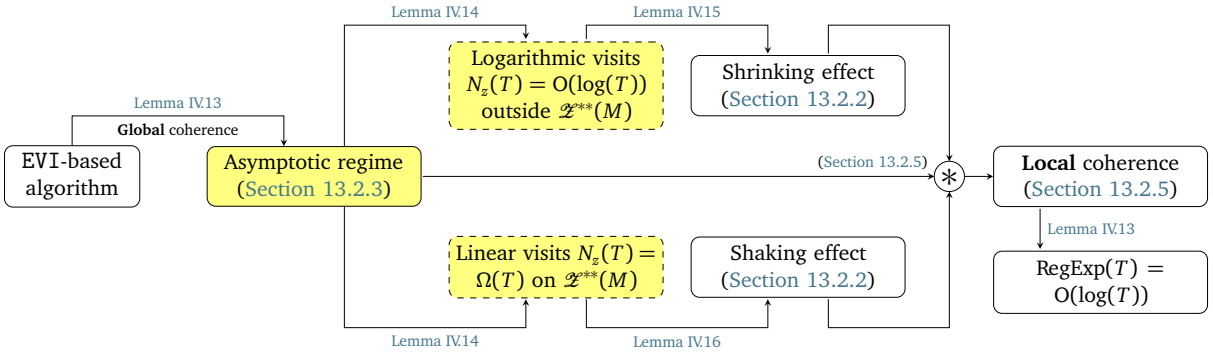
Conclude by combining (IV.12) with Section 13.A.2 (**STEP 4**) and Section 13.A.3 (**STEP 3**).

13.B The asymptotic regime of (VM): Proof of Lemma IV.14

In this section, we provide a proof of Lemma IV.14: *Let $M\in\mathscr{M}$ a non-degenerate model. Assume that the algorithm running is* EVI*-based with a confidence region $\mathscr{M}(t)\equiv\mathscr{M}_{\delta(t)}(t)$ as given by Section 7.A.2 such that Assumption 5 is satisfied, managing episodes with f-(VM) with arbitrary $f\in(0,1]$. There exists $\lambda>0$ s.t.:*

$$\begin{aligned}&\forall z\notin\mathscr{Z}_{**}(M),\quad \mathbf{P}^M(\exists T,\forall t\geq T:N_z(t)<\lambda\log(t))=1,\text{ and}\\&\forall z\in\mathscr{Z}_{**}(M),\quad \mathbf{P}^M\left(\exists T,\forall t\geq T:N_z(t)>\tfrac{1}{\lambda}t\right)=1.\end{aligned}$$

The proof is relies on the coherence lemma (Lemma IV.13). We show that the confidence regions are such that, if a sub-optimal policy is played, one of the pairs responsible for its optimistic gain must be sub-sampled. This provides a "global" coherence property, see (**STEP 1**), this



is used to upper bound the asymptotic visit rates of sub-optimal pairs, see (**STEP 2**) and (IV.29). Invoking the coherence lemma requires Assumption 5 to hold and M to be non-degenerate. We deduce in parallel that non optimal pairs are visited at most logarithmically often in (**STEP 3**) with (IV.31), and that optimal pairs are visited at least linearly often in (**STEP 4**) with (IV.32).

(**STEP 1**) *There exists a sequence of adapted events* (F_t) *satisfying* $\mathbf{P}(\exists T, \forall t \geq T : F_t) = 1$ *and a function* $\varphi : \mathbf{N} \to \mathbf{R}_+$ *with* $\varphi(t) = O(t)$ *s.t. the algorithm is* $((F_t), \lfloor \log(T) \rfloor, T, \varphi)$*-coherent.*

Proof. Introduce the good event $E_t := \{M \in \mathscr{M}(t)\}$. By design of the confidence region (see Section 7.A.2), $\mathbf{P}(\exists T, \forall t \geq T : E_t) = 1$. Let $T \geq 1$ and set $T_0 := \lfloor \log(T) \rfloor$. Pick $t \in [T_0, T]$ and let $[t_k, t_{k+1})$ the unique episode it falls in. We denote $\pi \equiv \pi^k$ for short and assume that π is sub-optimal from $S_t t$, i.e.,

$$g^*(S_t; M) > g^\pi(S_t; M). \tag{IV.26}$$

So there exists a class of pairs $\mathscr{Z}'$ which is recurrent under π, with $\mathscr{Z}' \subseteq \mathrm{Reach}(\pi, S_t)$ and such that $g^\pi(s; M) < g^*(s; M)$ for every $s \in \mathscr{S}(\mathscr{Z}')$. Let $s_0 \in \mathscr{S}(\mathscr{Z}')$.

Denote $\Delta_g := \min\{g^*(s; M) - g^\pi(s; M) : \pi \in \Pi, s \in \mathscr{S}, g^*(s;, M) > g^\pi(s; M)\} > 0$ the minimal gain gap in M. Because π is output by EVI (Algorithm 2) at time t_k, it is optimistically optimal at time t_k and $g^*(S_t; \mathscr{M}(t_k)) = g(S_t; \tilde{r}_\pi, \tilde{p}_\pi)$ for some $\tilde{r}_\pi \in \prod_s \mathscr{R}_{s,\pi(s)}(t_k)$ and $\tilde{p}_\pi \in \prod_s \mathscr{P}_{s,\pi(s)}(t_k)$. Furthermore, on E_t, we have $D(\mathscr{M}(t)) \leq D(M)$ hence every policy returned by EVI (Algorithm 2) has optimistic bias span at most $D(M)$ and its optimistic gain has null span. We have, on E_{t_k},

$$\begin{aligned}
\Delta_g &\leq g^*(s_0; M) - g^\pi(s_0; M) \\
&\overset{(\dagger)}{\leq} g^{\pi_{t_k}}(s_0; \mathscr{M}(t_k)) - g^\pi(s_0; M) \\
&\overset{(\ddagger)}{\leq} \|\tilde{r} - r\|_{\infty, \mathrm{Reach}(\pi, s_0)} + \tfrac{1}{2} D(M) \|\tilde{p} - p\|_{1, \mathrm{Reach}(\pi, s_0)}.
\end{aligned}$$

In (†), we have used that, on E_{t_k}, $g^*(s_0; M) \leq g^*(s_0; \mathscr{M}(t_k)) = g^*(S_{t_k}; \mathscr{M}(t_k)) = g^\pi(S_{t_k}; \mathscr{M}(t_k))$. In (‡), we first invoke a gain deviation inequality (Theorem II.1), then rely on the fact that by Assumption 5, the optimistic gain of π computed by EVI only depends on pairs that are reachable from s_0 under π on M. One of the two terms of the RHS of the above equation must be at least $\frac{1}{2}\Delta_g$. For instance, $D(M)\|\tilde{p} - p\|_{1, \mathrm{Reach}(\pi, s_0)} \geq \Delta_g$. We have:

$$\begin{aligned}
\Delta_g &\leq D(M)\big(\|\tilde{p} - \hat{p}\|_{1, \mathrm{Reach}(\pi, s_0)} + \|\hat{p} - p\|_{1, \mathrm{Reach}(\pi, s_0)}\big) \\
&= D(M)\Big(\min_{z \in \mathrm{Reach}(\pi, s_0)} \|\tilde{p}_z - \hat{p}_z(t_k)\|_1 + \min_{z \in \mathrm{Reach}(\pi, s_0)} \|\hat{p}_z - p_z\|_1\Big).
\end{aligned} \tag{IV.27}$$

Knowing that $\tilde{p}_z \in \mathscr{P}_z(t_k)$ and that $\mathscr{P}_z(-)$ either (**C1**) directly using Weissman's inequality (Lemma I.23), or (**C2**) using an empirical Bernstein's inequality (Lemma I.24) or (**C3**) using empirical likelihood inequalities (Lemma I.25), there exists a constant such that $N_z(t)\|\tilde{p}_z - \hat{p}_z(t)\|_1^2 \leq$

$C\log(t)$. Under (**C1**) this follows by construction; Under (**C2**), this is a consequence of Pinsker's inequality; This is slightly less easy to see under (**C3**). If $\mathscr{P}_z(t)$ is built using an empirical Bernstein's inequality, then it is obtained as a region of the form:

$$\prod_{s\in\mathscr{S}}\left\{\tilde{p}_{z,s}\in[0,1]:\left|\tilde{p}_{z,s}-\hat{p}_{z,s}(t)\right|\leq\alpha\sqrt{\tfrac{\log(t)}{N_z(t)}}+\beta\tfrac{\log(t)}{N_z(t)}\right\}\cap\mathscr{P}(\mathscr{S})$$

that is approximated by inequality in ℓ_1-norm by using $(\alpha+\beta^2)\sqrt{\log(t)/n}\wedge 1\geq(\alpha\sqrt{\log(t)/n}+\beta\log(t)/n)\wedge 1$. Overall, there exists a constant $C>0$ such that whether $\mathscr{P}_z(t)$ is built out of (**C1-3**), we have:

$$\mathscr{P}_z(t)\subseteq\left\{\tilde{p}_z\in\mathscr{P}(\mathscr{S}):N_z(t)\|\tilde{p}_z-\hat{p}_z(t)\|_1^2\leq C\log(t)\right\}=:\mathscr{P}'_z(t).$$

Up to increasing C, we can further assume that $\mathbf{P}(\exists T\geq 1,\forall t\geq 1:p_z\in\mathscr{P}'_z(t))=1$. Injecting this in Equation (IV.27), we see that on the asymptotically almost sure event $F^p_{t_k}:=\left\{\forall z,p_z\in\mathscr{P}'_z(t)\right\}$, we have:

$$\Delta_g\leq 2D(M)\min_{z\in\mathrm{Reach}(\pi,s_0)}\sqrt{\frac{C\log(t_k)}{N_z(t_k)}}\overset{(\dagger)}{\leq}2D(M)\min_{z\in\mathrm{Reach}(\pi,s_0)}\sqrt{\frac{2C\log(2t)}{N_z(t)}}\tag{IV.28}$$

where (†) uses that the (VM) guarantees $N_z(t_{k+1})\leq 2N_z(t_k)$ and $t_{k+1}\leq 2t_k$. Solving (IV.28) in $N_z(t)$, we find a condition of the form $N_z(t)\leq C'\log(t)$.

The same rationale can be used to handle the case where $\|\tilde{r}-r\|_{\infty,\mathrm{Reach}(\pi,s_0)}\geq\frac{1}{2}\Delta_g$, dealing with the design of another asymptotically almost sure event $F^r_t:=\left\{\forall z,r_z\in\mathscr{R}'_z(t)\right\}$ and ending with the same kind of upper-bound on $N_z(t)$. In the end, setting $F_t:=\bigcap_{t'=\lfloor t/2\rfloor}^t F^r_{t'}\cap F^p_{t'}\cap E_{t'}$ and $\varphi(T_0)=C'\log(t)$, we see that the algorithm is $((F_t),T_0,T,\varphi)$-coherent. □

(**STEP 2**) *There exists $C>0$ such that:*

$$\mathbf{P}\big(\exists T,\forall t\geq T,\forall z\in\mathscr{Z}^-(M):N_z(t)\leq C\log(t)\big).\tag{IV.29}$$

Proof. Since M is non-degenerate, coherence can be converted to regret guarantees (Lemma IV.13): From (**STEP 1**) follows that there exists constants $C_1,C_2>0$ such that:

$$\forall T\geq 1,\quad\mathbf{P}\left(\mathrm{Reg}(\log(T),T)\geq C_1+C_2\log(T)\text{ and }\bigcap_{t=\lfloor\log(T)\rfloor}^{T}F_t\right)\leq T^{-2}.\tag{IV.30}$$

Since $N_z(T)\leq N_z(T_0)+\Delta_z^{-1}\mathrm{Reg}(T_0,T)$, the condition $\mathrm{Reg}(\log(T),T)\leq C_1+C_2\log(T)$ is converted to $N_z(T)\leq C'_1+C'_2\log(T)$ for all $z\in\mathscr{Z}^-(M)$. We have:

$$\begin{aligned}(*)&:=\mathbf{P}\big(\forall T,\exists t\geq T,\exists z\in\mathscr{Z}^-(M):N_z(t)>C'_1+C'_2\log(t)\big)\\&\overset{(\dagger)}{=}\mathbf{P}\left(\forall T,\exists t\geq T,\exists z\in\mathscr{Z}^-(M):N_z(t)>C'_1+C'_2\log(t)\text{ and }\bigcap_{t=\lfloor\log(T)\rfloor}^{T}F_t\right)\\&\leq\mathbf{P}\left(\forall T,\exists t\geq T,\exists z\in\mathscr{Z}^-(M):\mathrm{Reg}(\log(T),T)>C_1+C_2\log(T)\text{ and }\bigcap_{t=\lfloor\log(T)\rfloor}^{T}F_t\right)\\&=\lim_{T\to\infty}\sum_{t\geq T}\sum_{z\in\mathscr{Z}^-(M)}\mathbf{P}\left(\mathrm{Reg}(\log(T),T)>C_1+C_2\log(T)\text{ and }\bigcap_{t=\lfloor\log(T)\rfloor}^{T}F_t\right)\\&\overset{(\ddagger)}{\leq}SA\lim_{T\to\infty}\tfrac{1}{T}=0.\end{aligned}$$

In the above, (†) follows by $\mathbf{P}(\limsup F_t)=1$ and (‡) by (IV.30). Up to assuming t large enough, we eventually have $C'_2\log(T)\geq C'_1$ hence the constant term can be ignored. □

(**STEP 3**) *There exists $C > 0$ such that:*

$$\mathbf{P}(\exists T, \forall t \geq T, \forall z \notin \mathscr{Z}^{**}(M) : N_z(t) \leq C \log(t)). \tag{IV.31}$$

Proof. Because M is non-degenerate, $\mathscr{Z}^*(M)$ defines a unique policy that we denote π^*. Introduce the reward function $f(z) := \mathbf{1}(z \notin \mathscr{Z}^{**}(M))$. Let g^f, h^f and Δ^f the gain, bias and gap functions of π^* in M endowed with the reward function f. Remark that $g^f(s) = 0$, that $h^f(s) = 0$ for $(s, \pi^*(s)) \in \mathscr{Z}^{**}(M)$ and that $\Delta^f(z) = 0$ for $z \in \mathscr{Z}^*(M)$. Denote $H_f := \mathrm{sp}(h^f) \vee \max_z |\Delta^f(z)|$. Therefore:

$$\begin{aligned}
\sum_{z \notin \mathscr{Z}^{**}(M)} N_z(T) &= \sum_{t=1}^{T} f(Z_t) \\
&= \sum_{t=1}^{T} \Big(\big(e_{S_t} - p(Z_t) \big) h^f - \Delta_f(Z_t) \Big) \\
&\leq H^f + \sum_{t=1}^{T} \mathbf{1}(Z_t \notin \mathscr{Z}^{**}(M)) \big(e_{S_{t+1}} - p(Z_t) \big) h^f + H^f \sum_{z \in \mathscr{Z}^-(M)} N_z(T) \\
&\overset{(\dagger)}{\leq} H^f \left(1 + 2 \sqrt{\sum_{z \notin \mathscr{Z}^{**}(M)} N_z(T) \log(T)} + \sum_{z \in \mathscr{Z}^-(M)} N_z(T) \right) \\
&\overset{(\ddagger)}{\leq} H^f \left(1 + 2 \sqrt{\sum_{z \notin \mathscr{Z}^{**}(M)} N_z(T) \log(T)} + SAC \log(T) \right)
\end{aligned}$$

where (†) holds with probability $1 - T^{-2}$ by Azuma-Hoeffding's inequality (Lemma I.22), and (†) holds on the asymptotically almost sure event $(\forall z \in \mathscr{Z}^-(M), N_z(T) \leq C \log(T))$ (see (IV.29)). This is an equation of the form $n \leq \alpha + \beta \sqrt{n}$ that implies in particular $n \leq 2(\alpha + \beta^2)$. In the end, we get:

$$\mathbf{P}\left(\forall T, \exists t \geq T : \sum_{z \notin \mathscr{Z}^{**}(M)} N_z(t) \leq 2H^f (1 + SAC \log(T) + 4 \log(T)) \right) = 1.$$

This concludes the proof. □

(**STEP 4**) *There exists $c > 0$ such that:*

$$\mathbf{P}(\exists T, \forall t \geq T, \forall z \in \mathscr{Z}^{**}(M) : N_z(t) \geq ct) = 1. \tag{IV.32}$$

Proof. This is established with a similar technique than (IV.31) in (**STEP 3**). By non-degeneracy of M, $\mathscr{Z}^*(M)$ defines a unique policy that we denote π^*. Fix $z_0 \in \mathscr{Z}^{**}(M)$ and introduce the reward function $f(z) = \mathbf{1}(z = z_0)$. Remark that $g^f(s) = c > 0$ for all $s \in \mathscr{S}$ and that $\Delta^f(z) = 0$ for all $z \in \mathscr{Z}^*(M)$. Let $H^f := \mathrm{sp}(h^f) \vee \max_z |\Delta^f(z)|$. We have:

$$\begin{aligned}
N_{z_0}(T) &:= \sum_{t=1}^{T} f(Z_t) \\
&= cT + \sum_{t=1}^{T} \Big(\big(e_{S_t} - p(Z_t) \big) h^f - \Delta_f(Z_t) \Big) \\
&\geq cT - \sum_{t=1}^{T} \mathbf{1}(Z_t \in \mathscr{Z}^-(M)) \big(e_{S_{t+1}} - p(Z_t) \big) h^f - H^f \sum_{z \in \mathscr{Z}^-(M)} N_z(T)
\end{aligned}$$

$$\geq cT - 2\sqrt{H^f SAC}\cdot\log(T) - H^f SAC\log(T) \sim cT$$

where the last inequality holds with probability $1-T^{-2}$ on the asymptotically almost sure event $(\forall z\in\mathscr{Z}^{-}(M): N_z(T)\leq C\log(T))$ given by (IV.29). We conclude accordingly. □

13.C The shrinking effect: Proof of Lemma IV.15

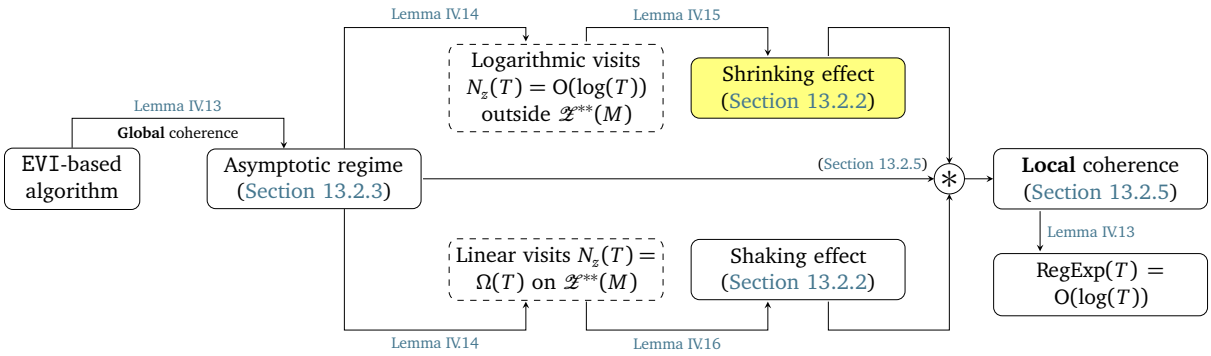
In this section, we provide a proof of Lemma IV.15: *Let $(t_{k(i)})$ the enumeration of exploration episodes, and let $T\geq 1$. Fix $\lambda, z\in\mathscr{Z}>0$. For all $\delta>0$, we can find $\epsilon, m, C>0$ such that:*

$$\textit{kernel}: \quad \mathbf{P}\left(F_{t_{k(i)}} \text{ and } \left[\exists t\in[t_{k(i)}, t_{k(i)}+T]: \begin{array}{c}\mathscr{P}_z(t)\not\subseteq\mathscr{P}_z(t_{k(i)-1})\\ \text{and } N_z(t)>N_z(t_{k(i)})+C\log\left(\frac{T}{\delta}\right)\end{array}\right]\right)\leq\delta,$$

$$\textit{reward}: \quad \mathbf{P}\left(F_{t_{k(i)}} \text{ and } \left[\exists t\in[t_{k(i)}, t_{k(i)}+T]: \begin{array}{c}\sup\mathscr{R}_z(t)>\sup\mathscr{R}_z(t_{k(i)-1})-\frac{N_z(t)-N_z(t_{k(i)})}{(t_{k(i)})^{1/3}}\\ \text{and } N_z(t)>N_z(t_{k(i)})+C\log\left(\frac{T}{\delta}\right)\end{array}\right]\right)\leq\delta$$

where $F_{t_{k(i)}} := (N_z(t_{k(i)})<\frac{1}{\lambda}\log(t_{k(i)}), \left|\hat{r}_z(t_{k(i)-1})-r_z\right|<\epsilon, \left\|\hat{p}_z(t_{k(i)-1})-p_z\right\|<\epsilon, t_{k(i)}>m)$.

The result is established separately for confidence regions constructed out of (**C1**) Weissman's inequality, (**C2**) empirical Bernstein inequalities and (**C3**) Empirical likelihood inequalities (see Section 7.A.2 for explicit formulas).



13.C.1 Weissman-type confidence regions

Establishing the shrinking phenomenon on kernels and rewards follows a similar line for rewards and kernels. Because they are a little bit harder, the analysis of the shrinking behavior of the kernels' confidence regions is detailed, and we explain how to adapt the proof to the rewards' confidence regions.

Lemma IV.20 (Shrinking Kernels, Weissman region). *Assume that confidence regions are built out of Weissman's inequality* (**C1**)*, i.e., $\mathscr{P}_z(t):=\{\tilde{p}_z\in\mathscr{P}(\mathscr{S}): N_z(t)\|\tilde{p}_z-\hat{p}_z(t)\|_1^2\leq C_0\log(C_1 t)\}$ with $C_0, C_1>0$. Fix $\lambda>0, z\in\mathscr{Z}$. For all $\delta>0$, we can find $\epsilon, M, C>0$ such that:*

$$\mathbf{P}\left(F_{t_{k(i)}} \text{ and } \left[\exists t\in[t_{k(i)}, t_{k(i)}+T]: \begin{array}{c}\mathscr{P}_z(t)\not\subseteq\mathscr{P}_z(t_{k(i)-1})\\ \text{and } N_z(t)>N_z(t_{k(i)})+C\log\left(\frac{T}{\delta}\right)\end{array}\right]\right)\leq\delta$$

where $F_t := (N_z(t_{k(i)})<\frac{1}{\lambda}\log(t_{k(i)}), |\hat{p}_z(t_{k(i)-1})-p_z|<\epsilon, t_{k(i)}>m)$.

Proof. Write $N_z(t,t') := N_z(t') - N_z(t)$ the number of times $z \in \mathscr{Z}$ is visited between the times t and t'. The empirical transition at time t is expressed as:

$$\begin{aligned}\hat{p}_z(t) &= \hat{p}_z(t_{k(i)-1}) + \frac{1}{N_z(t)}\left(N_z(t_{k(i)-1},t)\big(p_z - \hat{p}_z(t_{k(i)-1})\big) + \sum_{i=t_{k(i)-1}}^{t-1} \mathbf{1}(Z_i = z)\big(e_{S_{i+1}} - p_z\big)\right)\\ &=: \hat{p}_z(t_{k(i)-1}) + w_z(t_{k(i)-1},t). \end{aligned} \tag{IV.33}$$

On the event F_t, we have $\|w_z(t_{k(i)-1},t)\|_1 \le \frac{1}{N_z(t)}(N_z(t_{k(i)-1},t)\epsilon + \|\sum_i \mathbf{1}(Z_i = z)(e_{S_{i+1}} - p)\|_1)$, consisting in two terms. The first term is an error a priori, while the second is a the norm of a martingale which is the sum of $N_z(t_{k(i)-1},t)$ terms. Because in the target probability of Lemma IV.20, we invoke $F_{t_{k(i)}}$ that is $H_{t_{k(i)}}$-measurable, any visit of z prior to $t_{k(i)}$ in the martingale part of $w_z(t_{k(i)-1},t)$ must be casted out. Thankfully, on $F_{t_{k(i)}}$, we have:

$$N_z(t_{k(i)}) \le \lfloor \big(1 + f(t_{k(i)-1})\big)N_z(t_{k(i)})\big)\rfloor + 1 \le N_z(t_{k(i)-1}) + 1 + \lfloor \tfrac{1}{\lambda} f(t_{k(i)-1})\log(t_{k(i)})\rfloor$$

which is equal to $N_z(t_{k(i)-1}) + 1$ since $f(t) = \mathrm{o}\big(\frac{1}{\log(t)}\big)$, provided that $t_{k(i)} \ge m$ is large enough. Accordingly, we have $N_z(t_{k(i)-1}, t_{k(i)}) \le 1$ on $F_{t_{k(i)}}$. So, on $F_{t_{k(i)}}$, we have:

$$\big\|w_z(t_{k(i)-1},t)\big\|_1 \le \frac{1}{N_z(t)}\left(2 + N_z(t_{k(i)},t)\epsilon + \left\|\sum_{i=t_{k(i)}}^{t-1} \mathbf{1}(Z_i = z)\big(e_{S_{i+1}} - p_z\big)\right\|_1\right). \tag{IV.34}$$

By applying Weissman's inequality (Lemma I.23), the martingale can then be bounded as follows:

$$\mathbf{P}\left(\forall t \in [t_{k(i)}, t_{k(i)} + T],\ \left\|\sum_{i=t_{k(i)}}^{t-1} \mathbf{1}(Z_i = z)\big(e_{S_{i+1}} - p_z\big)\right\|_1 \ge \sqrt{SN_z(t_{k(i)},t)\log\big(\tfrac{T}{\delta}\big)}\right) \le \delta. \tag{IV.35}$$

Pick $\tilde{p}_z \in \mathscr{P}_z(t)$. Following (IV.35), we derive conditions on $N_z(t_{k(i)},t)$ such that $\tilde{p}_z \in \mathscr{P}_z(t_{k(i)-1})$ with high probability. We have:

$$\begin{aligned}(*) &= \sqrt{N_z(t_{k(i)-1})}\big\|\tilde{p}_z - \hat{p}_z(t_{k(i)-1})\big\|_1\\ &\le \sqrt{N_z(t_{k(i)-1})}\big\|\hat{p}_z(t) - \hat{p}_z(t_{k(i)-1})\big\|_1 + \sqrt{N_z(t_{k(i)-1})}\|\tilde{p}_z - \hat{p}_z(t)\|_1\\ &\le \sqrt{N_z(t_{k(i)-1})}\big\|w_z(t_{k(i)-1},t)\big\|_1 + \sqrt{\frac{N_z(t_{k(i)-1})}{N_z(t)}\cdot C_1\log(C_2 t)}\\ &\le \sqrt{N_z(t_{k(i)-1})}\big\|w_z(t_{k(i)-1},t)\big\|_1 + \sqrt{\frac{N_z(t) - N_z(t_{k(i)-1},t)}{N_z(t)}\cdot C_1\log(C_2 t_{k(i)-1})} + \mathrm{o}\big(\tfrac{1}{t}\big)\\ &\overset{(\dagger)}{\le} \sqrt{C_1\log(C_2 t_{k(i)-1})} - \frac{N_z(t_{k(i)-1},t)\sqrt{N_z(t_{k(i)-1})}}{2N_z(t)}\left(\sqrt{\frac{C_1\log(C_2 t_{k(i)-1})}{N_z(t_{k(i)-1})}} - 2\epsilon - \frac{4}{N_z(t_{k(i)},t)}\right)\\ &\qquad + \frac{\sqrt{SN_z(t_{k(i)-1})N_z(t_{k(i)},t)\log\big(\frac{T}{\delta}\big)}}{N_z(t)} + \mathrm{o}\big(\tfrac{1}{t_{k(i)-1}}\big)\end{aligned}$$

where (†) is obtained by a combination of (IV.34) and (IV.35), hence hold uniformly on t with probability $1-\delta$. On $F_{t_{k(i)}}$, we have $N_z(t_{k(i)-1}) \le \frac{1}{\lambda}\log(t_{k(i)-1})$ hence $N_z(t_{k(i)-1}) \le \frac{1}{\lambda}\log(C_2 t_{k(i)-1})$ since $C_2 \ge 1$. Using this in the equation above and simplifying terms a bit, we upper-bound $(*)$ by:

$$\sqrt{C_1\log(C_2 t_{k(i)-1})} + \frac{\sqrt{N_z(t_{k(i)-1})}}{N_z(t)}\Big(2 - N_z(t_{k(i)-1},t)\big(\tfrac{1}{2}\sqrt{\lambda C_1} - \epsilon\big) + \sqrt{N_z(t_{k(i)-1},t)\cdot S\log\big(\tfrac{T}{\delta}\big)}\Big)$$

The right-term is a deviation that up to the normalization $\sqrt{N_z(t_{k(i)-1})}/N_z(t)$, contains a term scaling as $-N_z(t_{k(i)-1},t)$ and another in $\sqrt{N_z(t_{k(i)-1},t)}$. The first is expected to be dominant, and this is indeed the case if $\epsilon < \frac{1}{6}\sqrt{\lambda C_1}$ and $\sqrt{N_z(t_{k(i)-1},t)\log(T/\delta} < \frac{1}{6}N_z(t_{k(i)-1},t)\sqrt{\lambda C_1}$. The second requirement leads to the sufficient condition:

$$N_z(t_{k(i)-1},t) \geq \tfrac{36S}{C_1\lambda}\log\big(\tfrac{T}{\delta}\big) =: C\log\big(\tfrac{T}{\delta}\big). \tag{IV.36}$$

Combining (IV.36), (IV.35) and the above upper-bound on $(*)$, provided that $N_z(t_{k(i)-1},t) \geq C\log(\frac{T}{\delta})$ and that $\epsilon > 0$ is chosen small enough, it holds on $F_{t_{k(i)}}$ with probability $1-\delta$ that, for all $t \in [t_{k(i)}, t_{k(i)}+T]$,

$$\sqrt{N_z(t_{k(i)-1})}\big\|\tilde{p}_z - \hat{p}_z(t_{k(i)-1})\big\|_1 \leq \sqrt{C_1\log(C_2 t_{k(i)-1})} - \frac{N_z(t_{k(i)-1},t)\sqrt{N_z(t_{k(i)-1})C_1\lambda}}{6N_z(t)} + o\big(\tfrac{1}{t_{k(i)-1}}\big) \tag{IV.37}$$

To conclude, observe that the $o(\frac{1}{T})$ is negligible in front of the negative term provided that t is large in front of T. We conclude that the condition (IV.37) is enough to guarantee $\tilde{p}_z \in \mathscr{P}_z(t_{k(i)-1})$, hence finishing the proof. □

Lemma IV.21 (Shrinking Rewards, Weissman region). *Assume that confidence regions are built out of Weissman's inequality* (**C1**)*, i.e.,* $\mathscr{R}_z(t) := \{\tilde{r}_z \geq 0 : N_z(t)|\tilde{r}_z - \hat{r}_z(t)|^2 \leq C_0\log(C_1 t)\}$ *with* $C_0, C_1 > 0$*. Fix* $\lambda > 0, z \in \mathscr{Z}$*. For all* $\delta > 0$*, we can find* $\epsilon, M, C > 0$ *such that:*

$$\mathbf{P}\left(F_{t_{k(i)}} \text{ and } \left[\exists t \in [t_{k(i)}, t_{k(i)}+T] : \begin{array}{c}\sup\mathscr{R}_z(t) > \sup\mathscr{R}_z(t_{k(i)-1}) - \frac{N_z(t)-N_z(t_{k(i)})}{C\log(t_{k(i)})} \\ \text{and } N_z(t) > N_z(t_{k(i)}) + C\log\big(\frac{T}{\delta}\big)\end{array}\right]\right) \leq \delta$$

where $F_t := (N_z(t_{k(i)}) < \frac{1}{\lambda}\log(t_{k(i)}), |\hat{r}_z(t_{k(i)-1}) - r_z| < \epsilon, t_{k(i)} > m)$.

Proof. The proof is exactly the same up to the analogue of (IV.37):

$$\sqrt{N_z(t_{k(i)-1})}\big\|\hat{r}_z(t_{k(i)-1}) - \tilde{r}_z\big\|_1 \leq \sqrt{C_1\log(C_2 t_{k(i)-1})} - \frac{N_z(t_{k(i)-1},t)\sqrt{N_z(t_{k(i)-1})C_1\lambda}}{6N_z(t_{k(i)-1})} + o\big(\tfrac{1}{t_{k(i)-1}}\big).$$

For the same reasons, the $o(\frac{1}{t_{k(i)-1}})$ is negligible. But also, because $N_z(t_{k(i)-1}) \leq \lambda\log(t_{k(i)-1})$, provided that t is large in front of T, we are guaranteed that $N_z(t)\sqrt{1/(\lambda C_1)} \leq (\lambda\log(t_{k(i)}) + T)\sqrt{1/(\lambda C_1)} \leq C_0\log(t_{k(i)})$ for some constant C_0, provided that $t_{k(i)}$ is large enough. Hence, we obtain that on the concentration event specified by (IV.35) and on $F_{t_{k(i)-1}}$,

$$\big\|\hat{r}_z(t_{k(i)-1}) - \tilde{r}_z\big\|_1 \leq \sqrt{\frac{C_1\log(C_2 t_{k(i)-1})}{N_z(t_{k(i)-1})}} - \frac{N_z(t_{k(i)-1},t)}{6C_0\log(t_{k(i)})}$$

Further pick $C \geq 6C_0$. This concludes the proof. □

13.C.2 About empirical Bernstein and empirical likelihood confidence regions

Lemma IV.20 can be adapted to Bernstein-type confidence regions (II.7) and empirical likelihood confidence regions (II.6), see Section 7.A.2 for the descriptions of such regions with explicit constants. These proofs follow a similar line than Lemma IV.20 but the computations are region-specific. The details can be found in the original paper Boone and Gaujal (2024).

13.D The shaking effect: Proof of Lemma IV.16

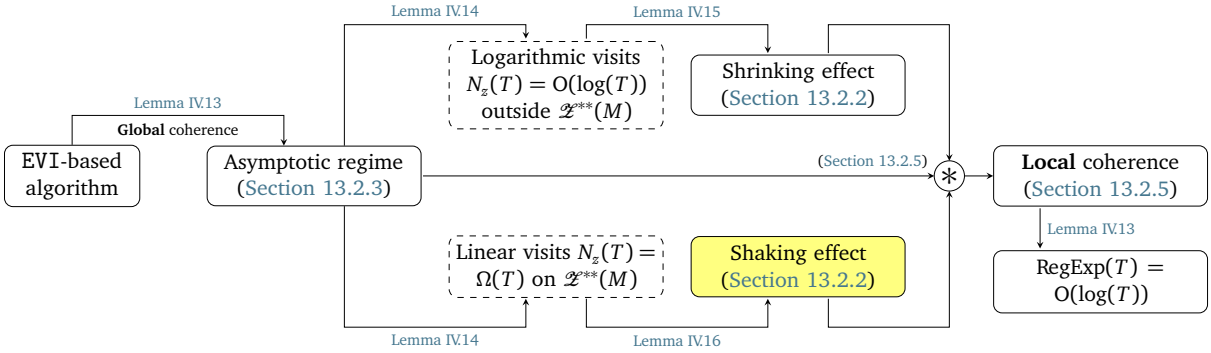
In this section, we provide a proof of Lemma IV.16: *Let $(t_{k(i)})$ the enumeration of exploration episodes, and let $T \geq 1$. Fix $\lambda, z \in \mathscr{Z}$ and for two sets $\mathscr{U}, \mathscr{V} \subseteq \mathbf{R}^n$, denote $d_{\mathrm{H}}(\mathscr{U}, \mathscr{V})$ the Hausdorff distance induced by the one-norm. We can find $c, m > 0$ such that:*

$$\text{(kernels)} \quad F_{t_{k(i)}} \supseteq \left(\forall t \in [t_{k(i)}, t_{k(i)} + T] : d_{\mathrm{H}}(\mathscr{P}_z(t), \mathscr{P}_z(t_{k(i)-1})) \leq \sqrt{\frac{c \log(t)}{t}} \right),$$

$$\text{(rewards)} \quad F_{t_{k(i)}} \supseteq \left(\forall t \in [t_{k(i)}, t_{k(i)} + T] : d_{\mathrm{H}}(\mathscr{R}_z(t), \mathscr{R}_z(t_{k(i)-1})) \leq \sqrt{\frac{c \log(t)}{t}} \right)$$

where $F_{t_{k(i)}} := (N_z(t_{k(i)-1}) > \lambda t_{k(i)-1}, t_{k(i)} > m) \cap (\forall t \in [t_{k(i)-1}, t_{k(i)}], M \in \mathscr{M}_{\delta(t)}(t))$.

Similarly to Lemma IV.15, the result is established separately for confidence regions constructed out of (**C1**) Weissman's inequality, (**C2**) empirical Bernstein inequalities and (**C3**) Empirical likelihood inequalities (see Section 7.A.2 for explicit formulas).



13.D.1 Weissman-type confidence regions

Again, the shaking phenomenon on kernels and rewards follows for similar reasons; Proving it for rewards consists in establishing the shaking property in dimension one, while proving it for kernels consits in establishing the shaking property in dimension $|\mathscr{S}|$. We accordingly to the proof of the result for kernels.

Lemma IV.22 (Shaking Kernels, Weissman region)**.** *Assume that confidence regions are built out of Weissman's inequality* (**C1**)*, i.e., $\mathscr{P}_z(t) := \{\tilde{p}_z \in \mathscr{P}(\mathscr{S}) : N_z(t) \|\tilde{p}_z - \hat{p}_z(t)\|_1^2 \leq C_0 \log(C_1 t)\}$ with $C_0, C_1 > 0$. Fix $\lambda, z \in \mathscr{Z}$ and for two sets $\mathscr{U}, \mathscr{V} \subseteq \mathbf{R}^n$, denote $d_{\mathrm{H}}(\mathscr{U}, \mathscr{V})$ the Hausdorff distance induced by the one-norm. We can find $c, m > 0$ such that:*

$$\text{(kernels)} \quad F_{t_{k(i)}} \subseteq \left(\forall t \in [t_{k(i)}, t_{k(i)} + T] : d_{\mathrm{H}}(\mathscr{P}_z(t), \mathscr{P}_z(t_{k(i)-1})) \leq \sqrt{\frac{c \log(t)}{t}} \right),$$

$$\text{(rewards)} \quad F_{t_{k(i)}} \subseteq \left(\forall t \in [t_{k(i)}, t_{k(i)} + T] : d_{\mathrm{H}}(\mathscr{R}_z(t), \mathscr{R}_z(t_{k(i)-1})) \leq \sqrt{\frac{c \log(t)}{t}} \right)$$

where $F_{t_{k(i)}} := (N_z(t_{k(i)-1}) > \lambda t_{k(i)-1}, t_{k(i)} > m) \cap (\forall t \in [t_{k(i)-1}, t_{k(i)}], M \in \mathscr{M}_{\delta(t)}(t))$.

Proof. Recall that $\delta(t) = \frac{1}{t}$. By construction of $\mathscr{M}_{\delta(t)}(t)$, for all $t \in \{t_{k(i)-1}, \ldots, t_{k(i)}\}$, we have

$$\|p_z - \hat{p}_z(t)\|_1 \leq \sqrt{\frac{S \log(2SA(1 + N_z(t)))}{N_z(t)}} \leq \sqrt{\frac{S \log(2SAt_{k(i)})}{\lambda t_{k(i)-1}}}. \tag{IV.38}$$

The vanishing multiplicative condition (VM) guarantees that $t_{k(i)} \leq 2t_{k(i)-1}$ provided that $t_{k(i)}$ is large enough, hene providing $N_z(t_{k(i)}) \geq \frac{1}{2}\lambda t_{k(i)}$. Moreover, if $t_{k(i)}$ is large in front of T, then $\hat{p}(t)$ moves by $O(\frac{T}{N_z(t_{k(i)})}) = O(\frac{T}{t_{k(i)}})$ during the time-segment $\{t_{k(i)}, \dots, t_{k(i)}+T-1\}$, which is negligible in front of $\sqrt{\log(t_{k(i)})/t_{k(i)}}$. We accordingly extend (IV.38) to $t \in \{t_{k(i)-1}, \dots, t_{k(i)} + T - 1\}$ with:

$$\|p_z - \hat{p}_z(t)\|_1 \leq \sqrt{\frac{2S\log(2SAt_{k(i)})}{\lambda t_{k(i)}}} + o\left(\sqrt{\frac{\log(t_{k(i)})}{t_{k(i)}}}\right) = \Theta\left(\sqrt{\frac{\log(t)}{t}}\right). \tag{IV.39}$$

The result is therefore obtained by estimating the Hausdorff distance between ℓ_1-ball of radius $\Theta(\sqrt{\log(t)/t})$ and with centers at distance $\Theta(\sqrt{\log(t)/t})$. □

13.D.2 About empirical Bernstein-type and empirical likelihodd confidence regions

Similarly to the shrinking phenomenon, Lemma IV.22 can be adapted to Bernstein-type confidence regions (II.7) and empirical likelihood confidence regions (II.6), see Section 7.A.2 for the descriptions of such regions with explicit constants. These proofs follow a similar line than Lemma IV.20 but the computations are region-specific. The details can be found in the original paper Boone and Gaujal (2024).

Chapter 14

Beyond the regret of exploration: The Sliding Regret

In this last chapter, we go beyond the study of the local regret at exploration times (Definition IV.1). The question is essentialized by studying the problem from the lens of **multi-armed bandits** and more particularly **two-armed bandits**. The question is the following: The celebrated UCB algorithm Auer (2002) is perhaps the simpler EVI-based algorithm (Algorithm II.2) possible. It is episode-less by renewing its policy at each round, so it should not suffer of the linear regret of exploration induced by Equation (DT), and should even have better guarantees that (PT) and (VM). And yet, when running UCB, we observe that its first order regret is bumpy because UCB plays sub-optimal actions many times in a row for arbitrarily large times (see Figure 14.1.1).

Although many learning algorithms for multi-armed bandits exist today, over a single run, their first order regret seems to follow one out of two tendencies. The first order regret of classical algorithms can be classified into two categories: **smooth** and **bumpy**.

14.1 Smooth and bumpy pseudo-regret curves

> **About notations.** This chapter is dedicated to stochastic bandits, a setting that is slightly different from Markov decision processes. We slightly modify the notations of the manuscript to adopt a style that is closer to the multi-armed bandits' standards.

We start by adapting notations.

A stochastic multi-armed bandit is a state-less Markov decision process, meaning that the pair space can be put in the form $\mathscr{Z} := \{1\} \times \{1, \dots, K\}$ where K denotes the number of arms of the bandit. The actions picked by the learner are written $A_1, A_2, \dots$ and the achieved rewards $R_1, R_2, \dots$. Again, the objective of the learner is to maximize the aggregate rewards $\sum_{t=1}^{T} R_t$, or equivalently to minimize the **regret**. In the sequel, we focus on **two-armed bandits** with **Bernoulli rewards** which is arguably the simplest settings of all, yet this setting is already a rich ground to study the local behavior of efficient learners. This results below can easily be generalized to multi-armed bandits (more than two) with single parameter exponential distributions for rewards. We use the style of notation of Honda and Takemura (2015). The distribution of arm a is denoted F_a, is a Bernoulli distribution $\mathrm{F}_a = \mathrm{B}(\mu_a)$ where μ_a is the mean of the arm. We denote $\mathbf{P}_{\mathrm{F}}(-)$ and $\mathbf{E}_{\mathrm{F}}[-]$ the associated probability and expectation operators, and whenever the distributions on arm are clear in the context, the subscript F is dropped; Remark that F accounts for what we previously wrote M in the manuscript. We further assume, up to permutations of arms, that $0 < \mu_2 < \mu_1 < 1$, thus both arms have interior mean rewards,

arm 1 is optimal and arm 2 is suboptimal. Symbolically, we have $g^*(M) = \mu_1$ hence the regret is $\mathrm{Reg}(T) := T\mu_1 - \sum_{t=1}^{T} R_t$. In the multi-armed bandit literature, the first order regret is rather called the **pseudo-regret**:

$$T\mu_1 - \sum_{t=1}^{T} \mu_{A_t} \tag{IV.1}$$

and the mean arm gap $\Delta := \mu_1 - \mu_2$ is the Bellman gap of action 2. During a run of an algorithm and for every arm $a = 1, 2$, we keep track of the number of visits with $N_a(t) := \sum_{i=1}^{t-1} \mathbf{1}(A_i = a)$. We will have $S_{a,n}$ and $\hat{\mu}_{a,n}$ denote the number of successes (when the reward equals one) and empirical mean of arm a after n draws of it. $S_a(t) := S_{a,N_a(t)}$ and $\hat{\mu}_a(t) := \hat{\mu}_{a,N_a(t)}$ will denote the associated number of successes and empirical mean at time t.

The multi-armed bandit problem consists in the design of regret efficient algorithms for multi-armed bandits and is simply a reinforcement learning problem on an average reward Markov decision with a single state. The whole discussion of 2 is therefore perfectly suited for this setting as well – this is not the first time of the manuscript that multi-armed bandits are mentioned. In the stochastic setting, three formulations coexist: the minimax formulation (Section 2.3), the model dependent formulation (Section 2.4) and the Bayesian formulation. Despite being my favorite, the latter was entirely skipped in this manuscript but the reader can read the very well-written and pretty complete introduction of (Lattimore and Szepesvári, 2020, §34-36) to the subject. The local regret considerations of this chapters are **model dependent** in nature, hence we will focus on the model dependent setting. Lower bounds of achievable expected regret are known (see Lai and Robbins (1985) and Section 2.4) and achieved by multiple methods, for example Thompson Sampling (Kaufmann et al. (2012)), `MED` (Honda and Takemura (2010)), `IMED` (Honda and Takemura (2015)), `KLUCB` (Garivier and Cappé (2011); Maillard et al. (2011)) or `MOSS` (Audibert et al. (2009)).

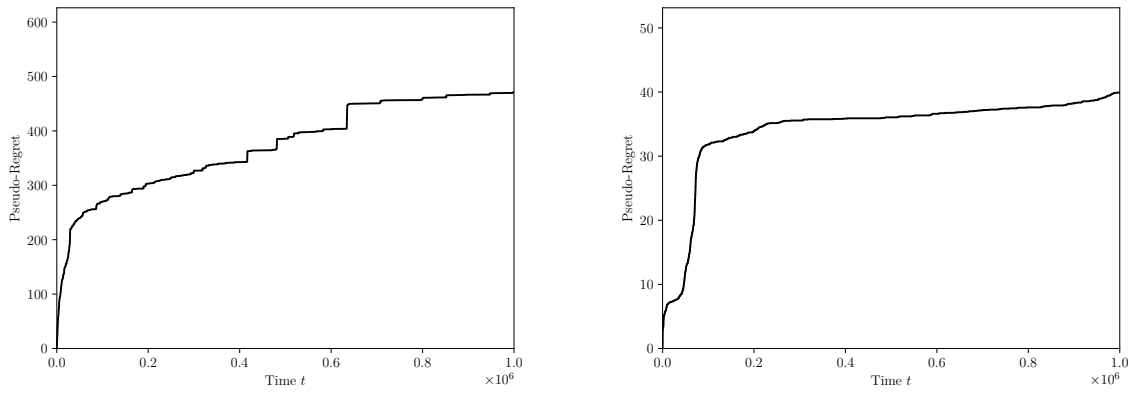


Figure 14.1.1: Typical one-shot pseudo-regret of `UCB` (left) and Thompson Sampling (right). The model is a two-armed bandit with Bernoulli rewards $\mathrm{B}(0.85), \mathrm{B}(0.8)$.

When applied to real-world tasks, what usually matters is the performance over a single run. Yet, although many of the previously mentioned methods are asymptotically optimal in expectation, their trajectory behaviors differ significatively. This is illustrated in Figure 14.1.1, plotting the typical pseudo-regrets of two popular algorithms: `UCB` by Auer (2002) and Thompson Sampling (`TS`) by Thompson (1933).

The difference is striking.

`UCB` has bumpy pseudo-regret and alternates between periods of time when it pulls the best and a suboptimal arm, meaning that it repeatedly pulls a bad arm several times in a row. In opposition, the pseudo-regret of Thompson Sampling is smooth and the algorithm seems to

pull suboptimal arms sporadically over time. These two trajectory portraits are in fact the two representatives of most existing algorithms for stochastic bandits. UCB showcases the typical one-shot pseudo-regret of index policies while TS illustrates the ones of randomized policies.

Outline of the chapter. The goal is the chapter is to explain the phenomenon reported in Figure 14.1.1. To simplify the discussion, the results are established for two-arm Bernoulli bandits. To measure the asymptotic bumpiness of the pseudo-regret is introduced the **sliding regret** (Definition IV.7), given by the worst pseudo-regret on time-windows of fixed length sliding to infinity; It is stronger than the previous introduced **regret of exploration** (Definition IV.2) because it covers the asymptotic time horizon continuously. Our first result, Theorem IV.24, provides a general condition to guarantee that a given policy have small sliding regret, later used to show that Thompson Sampling and MED have optimal sliding regret. Our second result, Theorem IV.37, states that all index policies have linear sliding regret provided that the index meets some regularity conditions. An **index policy** (see (Lattimore and Szepesvári, 2020, 35.4)) is an algorithm that, out of its current observations, associates a real-valued index to each arm then picks the arm with maximal index. Our result covers all classical index policies in the literature, such as UCB (Auer (2002)), UCB-V (Audibert et al. (2009)), MOSS (Audibert and Bubeck (2009)), KLUCB (Garivier and Cappé (2011); Maillard et al. (2011)), IMED (Honda and Takemura (2015)) as well as their variants.

The study of the sliding regret of index policies indicates that such algorithms have a tendency to pick suboptimal arms several times in a row at exploration episodes. What happens at these critical time-instants is what makes the sliding regret of index policies linear, as the probability to pick a suboptimal arm T times in a row starting from an exploration episode t is positive and does not vanish with t. We further show that this behavior is not negligible in average, because the **regret of exploration** (Definition IV.2) is shown to be optimal for Thompson Sampling and MED but sub-optimal for classical index policies.

14.2 Sliding regret and behavioral robustness to local histories

The presence of bumps observed in Figure 14.1.1 is related to the slope of the pseudo-regret, which is given by the pseudo-regret difference between two points in time. This consists in its truncation to a given time-window. Accordingly, to study the local behavior of the pseudo-regret, we study its truncation to time-windows of fixed length sliding to infinity.

Definition IV.7. *The asymptotic sliding regret (or **sliding regret** for short) is given by*

$$\mathrm{SliReg}(T) := \limsup_{t\to\infty} \left(T\mu^* - \sum_{i=1}^{T} \mu_{A_{t+i}} \right). \tag{IV.2}$$

Informally, the sliding regret measures the worst **local** pseudo regret over the trajectory. It is a non-negative quantity that measures the presence and the amplitude of the local changes of the pseudo-regret in the asymptotic regime. It is a new learning metric and as we will see, no-regret algorithms in the literature present two tendencies: those with small sliding regret and high sliding regret, embodied by Thompson Sampling and UCB respectively. The sliding regret can easily be lower and upper bounded by $\Delta^* := \mu_1 - \mu_2$ and $T\Delta^* = T(\mu_1 - \mu_2)$, as shown by the proposition below.

Proposition IV.23. *Consider an algorithm that, for all distribution on arms, have sublinear expected regret. Then it has sliding regret bounded as:*

$$\mu_1 - \mu_2 \leq \mathrm{SliReg}(T) \leq T(\mu_1 - \mu_2). \tag{IV.3}$$

This is a direct consequence of Proposition 14.A.1, established in Section 14.A.

There is a world between the lower and the upper bound and yet, the lower bound is usually achieved with randomized methods (such as TS) while the upper bound is reached with index policies (such as UCB). But associating small sliding regret guarantees with randomization is slightly misleading; this is rather a question of how the policy behaves depending on its recent history.

14.2.1 Behavioral robustness to local histories

In this section, we provide a general condition to obtain time independent upper bounds on the sliding regret with Theorem IV.24. This theorem states that if, regardless of the recent history (e.g., a bad arm has been pulled), the probability of picking a suboptimal action remains small, then the sliding regret is small. This is the property that we refer to as the behavioral robustness to local histories. It is precisely the property that UCB does not satisfy and that will lead to suboptimal sliding regret later on.

Theorem IV.24. *Let π a policy such that there exists a sequence of events $(E_t : t \geq 1)$ with $\mathbf{P}(\exists t, \forall s \geq T : E_s) = 1$, that satisfies:*

$$\exists d > 0, \forall i \geq 1 : \quad \mathbf{P}(A_{t+i} \neq 1 | H_{t:t+i}, E_t) = \mathrm{O}\left(\frac{1}{t^d}\right) \tag{IV.4}$$

where $H_{t:t+i}$ is the truncated history $(A_{t+j}, R_{t+j})_{0 \leq j < i}$. Then $\mathrm{SliReg}(\pi; T) \leq \lfloor \frac{1}{d} \rfloor (\mu_1 - \mu_2)$.

Proof. Let $n > \frac{1}{d}$ an integer. We show that $\mathbf{P}(\forall t, \exists s \geq t : \mathrm{Reg}(s; s+T) \geq n(\mu_1 - \mu_2)) = 0$. Remark that if $\mathrm{Reg}(s; s+T) \geq n(\mu_1 - \mu_2)$, there exists a set $I \subseteq \{0, \dots, T-1\}$ of size n such that, for all $i \in I$, $A_{s+i} = 2$. Denote Λ_n the collection of subsets of $\{0, \dots, T-1\}$ of size n and fix $I \in \Lambda_n$ whose elements are denoted $i_1 < i_2 < \dots < i_n$. We have:

$$\begin{aligned}
\mathbf{P}(\forall i \in I, A_{t+i} = 2; E_t) &= \mathbf{P}(A_{t+i_n} = 2 \mid E_t, (\forall i \in I \setminus \{i_n\}, A_{t+i} = 2)) \mathbf{P}(\forall i \in I \setminus \{i_n\}, A_{t+i} = 2; E_t) \\
&= \mathrm{O}\left(\frac{1}{t^d}\right) \cdot \mathbf{P}(\forall i \in I \setminus \{i_n\}, A_{t+i} = 2; E_t) \\
&= \dots \\
&= \mathrm{O}\left(\frac{1}{t^{nd}}\right) \mathbf{P}(E_t) = \mathrm{O}\left(\frac{1}{t^{nd}}\right).
\end{aligned}$$

Because E_s satisfies $\mathbf{P}(\liminf E_s) = 1$, check that for all sequence of events (F_s), $\mathbf{P}(\forall t, \exists s \geq t : F_s) = \mathbf{P}(\forall t, \exists s \geq t : E_s, F_s)$. We complete the proof with:

$$\begin{aligned}
\mathbf{P}(\forall t, \exists s \geq t : \mathrm{Reg}(s; s+T) \geq n(\mu_1 - \mu_2)) &= \lim_{t \to \infty} \mathbf{P}(\exists s \geq t : \mathrm{Reg}(s; s+T) \geq n(\mu_1 - \mu_2)) \\
&= \lim_{t \to \infty} \mathbf{P}(\exists s \geq t : \mathrm{Reg}(s; s+T) \geq n(\mu_1 - \mu_2), E_s) \\
&\leq \lim_{t \to \infty} \sum_{s \geq t} \mathbf{P}(\mathrm{Reg}(s; s+T) \geq n(\mu_1 - \mu_2), E_s) \\
&\leq \lim_{t \to \infty} \sum_{s \geq t} \mathbf{P}(\exists I \in \Lambda_n, \forall i \in I : A_{s+i} = 2; E_s)
\end{aligned}$$

$$\leq \lim_{t\to\infty} \sum_{s\geq t} T^n \, \mathrm{O}\left(\frac{1}{s^{nd}}\right) = 0$$

because $nd > 1$. So $\mathrm{SliReg}(T) \leq (n-1)(\mu_1 - \mu_2)$. □

In order to apply the theorem, one has to find the right sequence of events E_t such that (IV.4) is satisfied. This event usually characterizes what we will refer to as the **asymptotic regime** of an algorithm, consisting in concentration guarantees for the empirical data of the algorithm as well as convergence of the visit rate of the suboptimal arm. A complete example is provided with Thompson Sampling.

14.2.2 Application: Thompson Sampling

Thompson sampling (TS) Thompson (1933) is a Bayesian algorithm that, at time t, samples estimates of the arms' values from its posterior distribution and picks the arm with highest estimate. In the chosen Bernoulli setting, when the initial prior of TS is a tensor product of uniform distributions over $[0,1]$, the posteriors are Beta distributions and TS's estimates are sampled as:

$$\theta_a(t) \sim \mathrm{Beta}(1 + S_a(t), 1 + N_a(t) - S_a(t)). \tag{IV.5}$$

The expected regret of TS is pretty well understood. In the frequentist formulation of the multi-armed bandit problem, the regret is $\mathrm{O}(\log(T))$ (Agrawal and Goyal (2012)) and the multiplicative coefficient is the best possible, making TS an asymptotically optimal algorithm, see Agrawal and Goyal (2012); Kaufmann et al. (2012); Accordingly, TS achieves the model dependent lower bound of Lai and Robbins (1985) (Theorem I.15). We additionally show that its sliding regret is optimal.

Theorem IV.25. *Thompson Sampling has optimal sliding regret* $\mathrm{SliReg}(\mathrm{TS}; T) = \mu_1 - \mu_2$.

The complete proof is pretty tedious and deferred to the appendix. We outline the proof below.

Proof sketch of Theorem IV.25. The goal is to invoke Theorem IV.24, and this is achieved by characterizing the asymptotic behavior of TS, consisting in estimates of the sampling rates $\mathbf{P}(A_t = a)$ of the algorithm, estimates of the visit rates $N_a(t)$ as well as convergence of its empirical data.

Because the expected regret is sublinear, all arms are visited infinitely often, hence posteriors concentrate around the true means μ_1, μ_2, meaning that $\hat{\mu}_a(t)$ eventually converges to μ_a for $a = 1, 2$. A second known property of the asymptotic regime is that for some $b > 0$, $\sum \mathbf{P}(N_1(t) \leq t^b) < \infty$, see (Kaufmann et al., 2012, Proposition 1). So by Borel-Cantelli's lemma, $\mathbf{P}(\liminf(N_1(t) > t^b)) = 1$. Together with a combination of the Beta-Bernoulli trick (Agrawal and Goyal (2012)) and Sanov' Theorem, the sampling rates of Thompson Sampling are bounded as follows: For all $\epsilon > 0$, there exists a sequence of events (F_t^ϵ) with $\mathbf{P}(\liminf F_t^\epsilon) = 1$ such that

$$e^{-(1+c(\epsilon))N_2(t)\mathrm{kl}(\mu_2,\mu_1)} \leq \mathbf{P}(A_t = 2 \mid F_t^\epsilon) \leq e^{-(1-c(\epsilon))N_2(t)\mathrm{kl}(\mu_2,\mu_1)}, \tag{IV.6}$$

where $c(\epsilon)$ is a $o(1)$ when ϵ vanishes. We use (IV.6) to show that $N_2(t) \sim \log(t)/\mathrm{kl}(\mu_2, \mu_1)$. More precisely, we show that for all $\epsilon > 0$, the event

$$E_t^\epsilon := (\forall a, |\hat{\mu}_a(t) - \mu_a| < \epsilon) \cap \left(|N_2(t) - \tfrac{\log(t)}{\mathrm{kl}(\mu_2,\mu_1)}| < \epsilon \cdot \tfrac{\log(t)}{\mathrm{kl}(\mu_2,\mu_1)}\right)$$

holds eventually (the limit inferior is almost-sure). On this event and relying on (IV.6), we establish that

$$\mathbf{P}(A_t = 2 \mid E_t^\epsilon, H_{t:t+i}) = \mathrm{O}\Big(t^{-1+\underset{\epsilon\to 0}{\mathrm{o}}(1)}\Big).$$

Applying Theorem IV.24, we obtain $\mathrm{SliReg}(T) \leq \lfloor \frac{1}{1+\mathrm{o}(1)} \rfloor (\mu_1 - \mu_2)$. Making ϵ go to zero, we obtain optimal sliding regret guarantees for TS. □

14.2.3 Application: MED

MED (Honda and Takemura (2010)) is a randomized algorithm that, at time t, samples the arm a with probability proportional to $\exp(-N_a(t)\mathrm{kl}(\hat{\mu}_a(t), \hat{\mu}^*(t)))$. MED is known to have asymptotically optimal expected regret (refer to the original paper). As the empirical estimates converge, the sampling rate of the arm $a = 2$ is approximately $\exp(-N_2(t)\mathrm{kl}(\mu_2, \mu_1))$, which is essentially the same as Thompson Sampling's in the asymptotic regime. Therefore, the analysis of its sliding regret is similar to Thompson Sampling's.

Theorem IV.26. MED *has optimal sliding regret* $\mathrm{SliReg}(\mathtt{MED}; T) = \mu_1 - \mu_2$.

14.3 The bumpy regret of UCB

In this section, we show that unlike TS and MED, UCB does not have good sliding regret guarantees. In fact, the sliding regret of UCB is the worst possible as shown with Theorem IV.29.

The UCB algorithm from Auer (2002) is an index algorithm rooted in the *optimism-in-face-of-uncertainty* principle, that can be traced back at least to Lai and Robbins (1985). At time t, it picks the arm maximizing the index

$$\hat{\mu}_a(t) + \sqrt{\frac{2\log(t)}{N_a(t)}} \tag{IV.7}$$

which is $+\infty$ if $N_a(t) = 0$ by convention. Expected regret guarantees in $\mathrm{O}(\log(T))$ can be found in the original paper. This algorithm is the basis of all EVI-based algorithms (Algorithm II.2) and has been thoroughly investigated. This is why, to build intuition on how index algorithms typically behave, we dedicate this section to the analysis of the almost sure regime of UCB.

Thankfully, the almost-sure behavior of UCB at infinity is well-behaved and easy to describe. Eventually $\hat{\mu}_a(t)$ converges to μ_a and the visit rates of arms are such that the index of both arms (IV.7) are approximately equal. In fact, $N_1(t) \sim t$ and $N_2(t) \sim \frac{2}{(\mu_1-\mu_2)^2}\log(t)$ when time goes to infinity, see Proposition IV.27.

Proposition IV.27. *For all $\epsilon > 0$ and when running* UCB*, both of the following hold:*

(1) $\mathbf{P}(\exists t, \forall s \geq t : \forall a, |\mu_a(s) - \mu_a(s)| < \epsilon) = 1$;

(2) $\mathbf{P}\Big(\exists t, \forall s \geq t : \Big|N_2(t) - 2\Big(\frac{1}{\mu_1-\mu_2}\Big)^2 \log(t)\Big| < \epsilon \cdot 2\Big(\frac{1}{\mu_1-\mu_2}\Big)^2 \log(t)\Big) = 1.$

The proof of Proposition IV.27 is provided in Section 14.C.

14.3.1 The sliding regret of UCB

The analysis is driven by the behavior observed in Figure 14.1.1. UCB pulls every arm infinitely often, and every time it does pick the suboptimal arm, the probability that it picks it again in the

next round is high. Intuitively speaking, this happens because when UCB picks the suboptimal arm $a = 2$ and receives full reward $R_t = 1$, the empirical estimate $\hat{\mu}_2(t)$ increases enough so that UCB "thinks" that it has been sub-sampled. In other words, in the asymptotic regime of UCB, if a suboptimal arm provides promising rewards, UCB will keep pulling it to "make sure" that this arm's estimate is not wrongly estimated. This means that the central condition of Theorem IV.24 is not met by UCB. The time instants when UCB starts pulling suboptimal arms are called **exploration episodes**, and corresponds to the **exploration times** introduced by Definition IV.1, and are formally given by the increasing sequence of stopping times:

$$\tau_1 := \inf\{t : A_t = 2\}, \quad \tau_{k+1} := \inf\{t > \tau_k : A_t = 2 \wedge A_{t-1} = 1\}. \tag{IV.8}$$

Since all arms are pulled infinitely often, all these are almost surely finite.

Lemma IV.28. *Consider running* UCB*, and fix $T > 0$. There exists a sequence of events indexed by exploration episodes (E_{τ_k}) with $\mathbf{P}(\liminf_k E_{\tau_k}) = 1$, such that, for all sequence $(U_t : t \geq 1)$ of $\sigma(H_t)$-measurable events:*

$$\mathbf{P}\big(\forall i < T : A_{\tau_k+i} = 2 \mid E_{\tau_k}, U_{\tau_k}\big) \geq \mu_2^T.$$

The additional sequence (U_t) informs that the above lower bound is resilient to pollution of the history, and we could also write $\mathbf{P}(\forall i < T : A_{\tau_k+i} = 2 | E_{\tau_k}, H_{\tau_k}) \geq \mu_2^T$. The event E_{τ_k} is mostly about the concentrations of empirical means $\hat{\mu}_a(t)$ and of visit rates, given by Proposition IV.27. The main line of the proof is to estimate the evolution of UCB's index (IV.7) with respect to $\hat{\mu}_a(t)$, t and $N_a(t)$ for $a = 1, 2$, and to show that UCB keeps picking the suboptimal arm while it provides optimal reward. It follows that the sliding regret is linear.

Theorem IV.29. UCB *has the worst possible sliding regret* $\mathrm{SliReg}(T) = T(\mu_1 - \mu_2)$.

Proof. Since $\tau_{k+T} > \tau_k + 2T$, the events $(\exists i < T : A_{\tau_k+i} \neq 2)$ and $(\exists i < T : A_{\tau_{k+T}} \neq 2)$ do not overlap. Denote $F_{\tau_{\ell T}} := (\exists i < T : A_{\tau_{\ell T}+i} \neq 2)$ and let E_{τ_k} the event given by Lemma IV.28 for a fixed $\epsilon < \mu_2$. Observe that $\mathbf{P}(\forall \ell \geq k : E_{\tau_{\ell T}} \cap F_{\tau_{\ell T}})$ can be put in the form:

$$\prod_{\ell \geq k} \mathbf{P}\Big(\exists i < T : A_{\tau_{\ell T}+i} \neq 2 \mid E_{\tau_{\ell T}}, U_{\tau_{(\ell-1)T}+T}\Big)\mathbf{P}\Big(E_{\tau_{\ell T}} \mid U_{\tau_{(\ell-1)T}+T}\Big)$$

where $U_{\tau_{(\ell-1)T}+T} := \bigcap_{m \leq \ell-1}(E_{\tau_{mT}} \cap F_{\tau_{mT}})$ is a $\sigma(H_{\tau_{(\ell-1)T}+T})$-measurable event. Applying Lemma IV.28, we obtain:

$$\mathbf{P}(\forall \ell \geq k : E_{\tau_{\ell T}} \cap F_{\tau_{\ell T}}) \leq \prod_{\ell \geq k}\big(1 - (\mu_2 - \epsilon)^T\big) = 0.$$

It follows that:

$$\mathbf{P}(\forall l \geq k : F_{\tau_{\ell T}}) \leq \mathbf{P}\big(\forall l \geq k : (E_{\tau_{\ell T}})^{\complement}\big) + \mathbf{P}(\forall \ell \geq k : E_{\tau_{\ell T}} \cap F_{\tau_{\ell T}}) = \mathbf{P}\big(\forall l \geq k : (E_{\tau_{\ell T}})^{\complement}\big).$$

But since $\mathbf{P}(\liminf_k E_{\tau_{kT}}) = 1$, the above RHS goes to zero as $k \to \infty$. Therefore, we obtain $\mathbf{P}(\forall k, \exists \ell \geq k : \forall i < T, A_{\tau_\ell+i} = 2) = 1$, proving $\mathrm{SliReg}(\mathrm{UCB}; T) = T(\mu_1 - \mu_2)$. □

With the same proof techniques than Lemma IV.28, the above result can be further refined. When UCB receives full reward from the suboptimal arm, the associated index increases significatively so that, not only UCB will pick the suboptimal arm again in the next round, but it will also pick it in the next round, independently of the observed reward. Roughly speaking, if UCB receives many promising rewards in a row for the suboptimal arm, the associated index is polluted and UCB will blindly pick it again many times in succession, independently of the feedback.

Proposition IV.30. *Fix $T > 0$ and assume that we are running* UCB. *There exists an increasing sequence of almost-surely finite stopping times $(\sigma_k : k \geq 1)$ s.t.,*

$$\mathbf{P}(\mathrm{Reg}(\sigma_k; \sigma_k + T) \geq (\mu_1 - \mu_2)T) = 1.$$

For the construction of (σ_k), refer to Section 14.C.3.

Regarding Lemma IV.28, the lower bound of for $\mathbf{P}(\forall i < T : A_{\tau_k+i} = 2)$ is decreasing exponentially fast with T. Even though Theorem IV.29 states that the pseudo-regret of UCB makes arbitrarily large jumps infinitely often, how rare are these large jumps? While it is known that the infinite monkey eventually writes the complete works of William Shakespeare, the expected time that the animal requires to eventually write the first sentence of *Romeo and Juliet* is stupidly large.

14.3.2 The regret of exploration of UCB

If UCB has a tendency to pick the suboptimal arm many times at exploration episodes $(\tau_k : k \geq 1)$, see (IV.8), how significant is this tendency? We now investigate the expected regret starting from τ_k using the regret of exploration (Definition IV.2). In multi-armed bandits, the regret of exploration can equivalently be written as follows.

Definition IV.8. *The **regret of exploration** of an algorithm is the quantity:*

$$\mathrm{RegExp}(T) := \limsup_{k\to\infty} \mathbf{E}[\mathrm{Reg}(\tau_k; \tau_k + T)]. \tag{IV.9}$$

As soon as an algorithm visits every arm infinitely often (e.g., if consistent), the regret of exploration is well-defined, although the notion of **exploration episode** is less natural for randomized algorithms such as TS or MED than it is for index algorithms like UCB. The regret of exploration is an alternative measure to the sliding regret, also quantifying the tendency of an algorithm to aggregate suboptimal play. The two are linked as follows.

Proposition IV.31. *For every consistent algorithm,* $\mathrm{RegExp}(T) \leq \mathbf{E}[\mathrm{SliReg}(T)]$.

Proof. Since $\tau_k < \tau_{k+1}$, we have $\tau_k \geq k$. Therefore:

$$\begin{aligned}\mathrm{RegExp}(T) := \inf_k \sup_{\ell\geq k} \mathbf{E}[\mathrm{Reg}(\tau_\ell; \tau_\ell + T)] &\leq \inf_t \sup_{s\geq t} \mathbf{E}[\mathrm{Reg}(s; s+T)] \\ &\leq \inf_t \mathbf{E}\Big[\sup_{s\geq t} \mathrm{Reg}(s; s+T)\Big].\end{aligned}$$

By definition, $\mathrm{Reg}(s; s+T) \in [0, T(\mu_1 - \mu_2)]$ almost surely, so is bounded. By the Bounded Convergence Theorem, $\inf_t \mathbf{E}\big[\sup_{s\geq t} \mathrm{Reg}(s; s+T)\big] = \mathbf{E}[\inf_t \sup_{s\geq t} \mathrm{Reg}(s; s+T)]$. We readily obtain: $\mathrm{RegExp}(T) \leq \mathbf{E}[\mathrm{SliReg}(T)]$. □

Combined with Theorem IV.25, this shows that Thompson Sampling has optimal regret of exploration. The same goes for MED, since MED also has sliding regret $\mu_1 - \mu_2$.

Corollary IV.32. *Thompson Sampling and* MED *have optimal regret of exploration, that is, for $\pi =$* TS *or* MED, $\mathrm{RegExp}(\pi; T) = \mu_1 - \mu_2$.

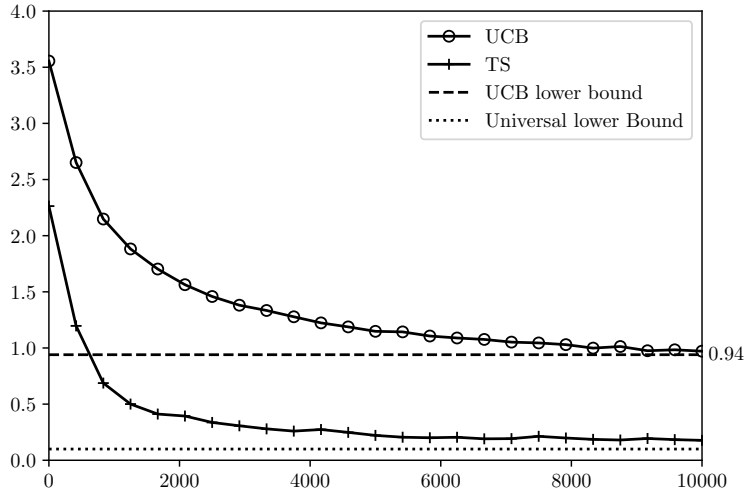


Figure 14.3.1: The approximate of $t \mapsto \mathrm{RegExp}'(t;100)$ for UCB and TS executed on the two-arm bandit $(\mathrm{B}(0.9), \mathrm{B}(0.8))$, averaged over 10k runs, with $W = 200$ (see below).

The regret of exploration of UCB is shown to be lower bounded by $C(T)(\mu_1 - \mu_2)$ where $C(T)$ is bounded away from 1. We show that at exploration episodes and in the asymptotic regime, UCB behaves like a random walk with a negative drift, and that the regret exploration is related to the reaching time to $\mathbf{R}_-$. The proof is found in Section 14.C.4.

Theorem IV.33. *Let $(X_t : t \geq 1)$ a sequence of i.i.d. random variables with distribution $\mathrm{B}(\mu_2)$. Let σ_T the stopping time $T \wedge \inf\{t \geq 1 : -\frac{\mu_1-\mu_2}{2} + \frac{1}{t}\sum_{i=1}^{t}(X_t - \mu_2) \leq 0\}$. For all $T \geq 1$, we have $\mathrm{RegExp}(\mathrm{UCB}; T) \geq (\mu_1 - \mu_2)\mathbf{E}[\sigma_T]$.*

How tight is this result? How close is $\mathbf{E}[\mathrm{Reg}(\tau_k; \tau_k + T)]$ to $(\mu_1 - \mu_2)\mathbf{E}[\sigma_T]$ in practice? To proceed, we estimate as a function of t the expected regret at exploration episodes near t, consisting in $\mathrm{RegExp}'(t;T) := \mathbf{E}[\mathrm{Reg}(t; t+T) | \exists k, t = \tau_k]$. To estimate this function, we repeatedly run the algorithm to obtain a family S of samples of $(\tau_k, \mathrm{Reg}(\tau_k; \tau_k + T))$. Then, we approximate $\mathrm{RegExp}'(t;T)$ as the averages of y for $(x,y) \in S$ such that $|x - t| < W$ where W is a parameter of the approximation. As shown in Figure 14.3.1, we indeed confirm that in practice, the expected regret during an exploration episode seems to converge to the anticipated lower bound $(\mu_1 - \mu_2)\mathbf{E}[\sigma_T]$ reasonably quickly.

14.4 General index algorithms

The behavior reported in Figure 14.1.1 and analyzed in the previous section is not specific to UCB. In this section, we generalize the analysis of UCB to most index policies of the literature. We provide a set of conditions under which an index policy has linear sliding regret, see Theorem IV.37. My original paper Boone (2023), from which this chapter is adapted, provides a collection of conditions that an algorithm have to satisfy in order to generalize the proof of UCB; Actually, this whole section is nothing less than a heavy abstractification of the proof techniques of Section 14.3 and perhaps fails to properly convey the general intuition.

The general intuition is that, ignoring second order terms, all index algorithms I know of Audibert and Bubeck (2009); Audibert et al. (2009); Auer (2002); Garivier and Cappé (2011);

Algorithm	Original Index	Reworked index	$I_{\max}$
UCB	$\hat{\mu}_a(t) + \sqrt{\frac{2\log(t)}{N_a(t)}}$	-	∞
MOSS	$\hat{\mu}_a(t) + \sqrt{\frac{\log(t/KN_a(t))}{N_a(t)}}$	-	∞
KLUCB	$\max\{\mu : N_a(t)\mathrm{kl}(\hat{\mu}_a(t), \mu) \leq \log(t)\}$	-	1
IMED	$-N_a(t)\mathrm{kl}(\hat{\mu}_a(t), \hat{\mu}^*(t)) - \log N_a(t)$	$\frac{\log(t)}{N_a(t)\mathrm{kl}(\hat{\mu}_a(t),\hat{\mu}^*(t))+\log N_a(t)}$	∞

Table 14.4.1: Examples of indexes.

Honda and Takemura (2015); Lattimore (2018); Maillard et al. (2011); Thompson (1933) (in)directly track a statistical lower bound. UCB, MOSS and AdaUCB track the lower-bound for Gaussian distributions, KLUCB and IMED track the lower bound for general distributions, UCB-V track the lower-bound for exponential distributions. Morally, these lower bounds are strongly convex in the arms' means hence when sampling a sub-optimal arm a and obtaining a very good reward, the required visit count of a estimated from the empirical data increases by strictly more than 1, hence the algorithm believes that a is sub-visited and will pull the arm right away.

The theory below arguably applies to a broader class of index policies – but the class of index policies that do not follow a lower bound is not populated.

14.4.1 Index policies and generalizing UCB's analysis

An **index policy** is an algorithm that, out of past observations, associates to every arm a numerical value called the **index** of the arm, and pulls the arm with maximal index. In the sequel, we consider indexes of the form

$$I(\hat{\mu}_a(t), \hat{\mu}_{-a}(t), N_a(t), t) \in [0, I_{\max}] \tag{IV.10}$$

where $I_{\max} \in (0, +\infty]$ is the maximal value that the index can reach (possibly infinite), $\hat{\mu}_a(t)$ the empirical value of the considered arm, $\hat{\mu}_{-a}(t)$ the collection of the empirical values of other arms, $N_a(t)$ the current number of visits of the arm and t the time. Accordingly, at time t, the algorithm picks $A_t \in \arg\max_a I(\hat{\mu}_a(t), \hat{\mu}_{-a}(t), N_a(t), t)$. Remark that the ordering of $\hat{\mu}_a(t)$ and $\hat{\mu}_{-a}(t)$ is important because $I(\hat{\mu}_1(t), \hat{\mu}_2(t), N_1(t), t)$ refers to the index of arm $a = 1$ while $I(\hat{\mu}_2(t), \hat{\mu}_1(t), N_2(t), t)$ refers to the index of arm $a = 2$; we will write $I_a(t)$ and $I_a(\hat{\mu}(t), N_a(t), t)$ for simplicity.

Our goal is to generalize Theorem IV.29 and Theorem IV.33 to general index policies. Our final result is summarized with Theorem IV.37. Of course, it is impossible to grasp all index policies within a single result, so the index has to meet regularity conditions for our result to be applicable. We design a set of nine conditions (**A 1-9**). All of them are met by classical existing indexes.

The argument mostly follows the lines of UCB's; Hence the question is whether what are the properties that $I(-)$ must satisfy so that the ideas behind the local analysis of UCB still applies. The steps are as follows: (1) all arms are visited infinitely often; (2) visit rates converge; (3) at the asymptotic regime, if a draw of the bad arm yields maximal reward, it will be drawn again immediately; and (4) the third property is enough so that the index algorithm is subjected to poisoning. By poisoning, we mean that if the bad arm provides maximal reward several times in a row, then whatever happens thereafter, the algorithm will keep picking the bad arm a few times in a row.

14.4.2 Asymptotic regimes of algorithms

Most of the regularity conditions that we require on $I(-)$ can be expressed in terms of continuity with respect to the topology of coordinate-wise equivalence of sequences. This topology appears naturally. As times goes on, one may expect that $(\hat{\mu}_1(t), \hat{\mu}_2(t), N_1(t), N_2(t))$ gets closer an closer to $(\mu_1, \mu_2, n_1(t), n_2(t))$ where $n_1(t)$ and $n_2(t)$ are the deterministic visit rates of arms. In order to approximate $I(\hat{\mu}_2(t), \hat{\mu}_1(t), N_2(t), t)$ by $I(\mu_2, \mu_1, n_2(t), t)$ for instance, we need $I(-)$ to act continuously on equivalent sequences.

Definition IV.9 (Asymptotic Topology)**.** *Consider a sequence $(x_1(n), \dots, x_d(n))$ of $\mathbf{R}^d$. A set $U \subseteq \mathbf{N} \to \mathbf{R}^d$ is said **open at** x if there exists $\epsilon > 0$ such that it contains all $y : \mathbf{N} \to \mathbf{R}^d$ satisfying:*

$$\exists N, \forall n > N, \forall i : \quad |x_i(n) - y_i(n)| < \epsilon |x_i(n)|.$$

This topology that we obtain is the topology of coordinate-wise equivalence of sequences. For instance, if $d = 1$, we have $x(n) \sim y(n)$ if, and only if y belongs to all the neighborhoods of x; Hence we write $x \sim y$ if y belongs to every neighborhood of x From now on, we endow the set of sequences of $\mathbf{R}^d$ with this topology.

Regarding the literature, it is fairly reasonable to have an index satisfying the following properties. (1) Monotonicity: the index is increasing in μ_a, decreasing in μ_{-a}, decreasing in N_a and increasing in t. (2) Unplayed arms see their index growing enough so that all arms are pulled infinitely often. (3) Convergence: an arm which is being pulled linearly often (e.g., an optimal arm) have converging index. Most index algorithms in the literature can be reworked so that these properties are satisfied, see Table 14.4.1.

Assumptions 1. The first required set of assumptions is the following.

(A 1) (Monotonicity) *The index $I_a(-)$ is increasing in μ_a, decreasing in μ_{-a}, decreasing in N_a and increasing in t.*

(A 2) (Diverging in t) *For all fixed $n \geq 1$ and $\nu_2 \in [0, 1]$, in the neighborhood of (ν_2, μ_1, n, t), we have $I_2(t) \to I_{\max}$.*

(A 3) (Convergence) *In the neighborhood of (μ_1, μ_2, t, t), $I_1(t)$ converges to some positive $I(\mu_1, \mu_2) < I_{\max}$.*

Lemma IV.34. *Assume that $I(-)$ satisfies (**A 1-3**). Then, for $a = 1, 2$, $\hat{\mu}_a(t) \to \mu_a$ a.s.*

Equivalently, $t \mapsto (\hat{\mu}_1(t), \hat{\mu}_2(t))$ is in every neighborhood of $t \mapsto (\mu_1, \mu_2)$, i.e., the two sequences are topologically indistinguishable. In practice, when running `UCB`, or `KLUCB` or `IMED`, the arms' numbers of visits are such that all indexes are equal. Because the index of the optimal arm converges to $I(\mu_1, \mu_2)$, $N_2(t)$ must be such that $I(\hat{\mu}_2(t), \hat{\mu}_1(t), N_2(t), t)$ is approximately $I(\mu_1, \mu_2)$, and the inverse must be continuous in $\hat{\mu}_2(t), \hat{\mu}_1(t), I(\mu_1, \mu_2)$. This leads to the condition (**A 4**). It is completed with (**A 5**), stating that the derivative of the inverse is not null. The two combined guarantee that $N_2(t) \sim n_2(t)$ for some deterministic $n_2(t)$. The last condition (**A 6**), which is a formulation of the no-regret property, makes sure that $N_1(t) \sim t$ once $N_2(t) \sim n_2(t)$.

Assumptions 2. Convergence of visit rates $N_a(t)$.

(A 4) (Continuous inverse in n) *Denote $f_{\nu_1, \nu_2, x}(t) := [I_2^{-1}(\nu_2, \nu_1, -, t)](x)$ the partial inverse in the number of visits for arm $a = 2$ and let $n_2(t) := f_{\mu_1, \mu_2, I(\mu_1, \mu_2)}(t)$. The map $(t \mapsto (\nu_1(t), \nu_2(t), x(t))) \mapsto (t \mapsto f_{\nu_1(t), \nu_2(t), x(t)}(t))$ is continuous in a neighborhood of $(\mu_1, \mu_2, I(\mu_1, \mu_2))$.*

(A 5) (Asymptotic monotonicity in n) *There is a non-negative definite function ℓ such that for $\epsilon > 0$ and in a neighborhood of $t \mapsto (\mu_1, \mu_2)$,*

$$I(\nu_2, \nu_1, (1+\epsilon)n_2(t), t) \leq (1-\ell(\epsilon))I(\nu_2, \nu_1, n_2(t), t),$$

and similarly, $I(\nu_2, \nu_1, (1-\epsilon)n_2(t), t) \geq (1+\ell(\epsilon))I(\nu_2, \nu_1, n_2(t), t)$.

(A 6) (No-Regret) *$n_2(t)$ is sublinear in t, $n_2(t) \to \infty$ and $n_2(t) \sim n_2(at)$ (for all $a > 0$) when $t \to \infty$.*

Lemma IV.35. *If $I(-)$ satisfies (**A 1-6**), then $(\hat{\mu}_1(t), \hat{\mu}_2(t), N_1(t), N_2(t)) \sim (\mu_1, \mu_2, t, n_2(t))$ a.s. The sequence $t \mapsto (\mu_1, \mu_2, t, n_2(t))$ will be called the **asymptotic regime**.*

14.4.3 Local behavior in the asymptotic regime of index policies

To analyze the local evolution of indexes in the asymptotic regime, we assume that for every arm, $I_a(t+h) - I_a(t)$ can be approximated by its Taylor expansion, and that this Taylor expansion depends continuously on the parameters $\hat{\mu}_a(t), N_a(t)$ and t. This is expressed by (**A 7**). (**A 9**) states that not all terms vary at the same speed; Namely, that the partial derivatives of $I_1(t)$ are negligible, and that in the Taylor expansion of $I_2(t+h) - I_2(t)$, the term $\partial_t I_2(t)$ can be neglected. Lastly, (**A 8**) states that the evolution of $I_2(t)$ relatively to $N_2(t)$ and $\hat{\mu}_2(t)$ are comparable, and the evolution relatively to $\hat{\mu}_2(t)$ is large enough in front of the one relatively to $N_2(t)$. This guarantees that if the suboptimal arm $a = 2$ is pulled and yield maximal reward $R_t = 1$, it will be pulled in the next round.

Assumptions 3. Local properties of $I_a(t)$ in the asymptotic regime.

(A 7) (Taylor expansion) *In a neighborhood of the asymptotic regime (say (ν_1, ν_2, m_1, m_2) in a neighborhood of (μ_1, μ_2, n_1, n_2)), for all fixed $h \geq 1$ and all arm a, we have:*

$$\begin{aligned} I_a(t+h) - I_a(t) \sim\ & (\nu_a(t+h) - \nu_a(t)) \cdot \partial_{\mu_a} I_a(t) \\ & + (\nu_{-a}(t+h) - \nu_{-a}(t)) \cdot \partial_{\mu_{-a}} I_a(t) \\ & + (m_a(t+h) - m_a(t)) \cdot \partial_n I_a(t) \\ & + h \cdot \partial_t I_a(t). \end{aligned}$$

(A 8) (ρ-optimism condition) *There is a constant $\rho \in [0,1)$, such that in a neighborhood of the asymptotic regime, $\partial_n I_2(t) \sim -\frac{\rho(1-\mu_2)}{m_2(t)} \partial_{\mu_2} I_2(t)$.*

(A 9) (Negligible derivatives) *In a neighborhood of the asymptotic regime, both $\partial_t I_1(t)$ and $\partial_t I_2(t)$ are $o(\partial_n I_2(t))$; And $\partial_{\mu_2} I_1(t) = o(\partial_{\mu_2} I_2(t))$.*

Lemma IV.36. *Let $I(-)$ an index satisfying (**A 1-9**). Fix $T > 0$. There exists a sequence of events indexed by exploration episodes (E_{τ_k}) with $\mathbf{P}(\liminf_k E_{\tau_k}) = 1$, such that, for all sequence $(U_t : t \geq 1)$ of $\sigma(H_t)$-measurable events:*

$$\mathbf{P}\big(\forall i < T : A_{\tau_k + i} = 2 \mid E_{\tau_k}, U_{\tau_k}\big) \geq \mu_2^T.$$

Proof. By Lemma IV.35, we know that $(\hat{\mu}_1(t), \hat{\mu}_2(t), N_1(t), N_2(t))$ goes to the asymptotic regime $(\mu_1, \mu_2, t, n_2(t))$ almost surely, so (**A 7-9**) can be instantiated to the random quantities. Suppose that t is large enough and is such that over the time-range $\{t, \dots, t+h-1\}$, we have $A_s = 2$. Then we can write:

$$(I_2(t+h) - I_1(t)) - (I_2(t) - I_1(t))$$

$$\sim \frac{\sum_{i=0}^{h-1}(R_{t+i}-\mu_2)}{N_2(t)}\big(\partial_{\mu_2}I_2(t)-\partial_{\mu_2}I_1(t)\big)+h\partial_n I_2(t)+h(\partial_t I_2(t)-\partial I_1(t)) \qquad \textbf{(A 7)}$$

$$\sim \frac{\sum_{i=0}^{h-1}(R_{t+i}-\mu_2)}{N/_2(t)}\partial_{\mu_2}I_2(t)+h\partial_n I_2(t)+h(\partial_t I_2(t)-\partial I_1(t)) \qquad \textbf{(A 9)}$$

$$\gtrsim \frac{\sum_{i=0}^{h-1}(R_{t+i}-\mu_2)-\rho(1-\mu_2)h}{N_2(t)}\partial_{\mu_2}I_2(t)+h(\partial_t I_2(t)-\partial I_1(t)). \qquad \textbf{(A 8)}$$

Assume that $R_{t+i}=1$ for all $i\in\{0,\dots,h-1\}$. We get

$$\begin{aligned}(I_2(t+h)-I_1(t))-(I_2(t)-I_1(t)) &\gtrsim \frac{(1-\rho)(1-\mu_2)h}{N_2(t)}\partial_{\mu_2}I_2(t)+h(\partial_t I_2(t)-\partial I_1(t))\\ &\sim \frac{(1-\rho)(1-\mu_2)h}{N_2(t)}\partial_{\mu_2}I_2(t). \qquad \textbf{(A 9)}\end{aligned}$$

Since $A_t=2$, we have $I_2(t)-I_1(t)\geq 0$. By **(A 1)**, $\partial_{\mu_2}(I_2(t))>0$ so $I_2(t+h)-I_1(t+h)>0$, hence $A_{t+h}=2$. We have established that, in the asymptotic regime, if $A_s=2$ for $s\in\{t,\dots,t+h-1\}$ with $R_s=1$, then $A_{t+h}=2$ as well. This means that the index policy essentially behaves like UCB: If the bad arm only yields optimal rewards, it is repeatedly pulled. □

It means that Lemma IV.28 extends to general indexes satisfying **(A 1-9)**. Therefore, and with the same proof, so does Theorem IV.29: Index policies pull the bad arm for arbitrary long time-window infinitely often. Theorem IV.33 also generalizes, and the regret of an index policy at exploration episodes can be predicted. It locally behaves like a random walk. The accuracy of the prediction is experimentally measured in Figure 14.3.1.

Theorem IV.37. *Let $I(-)$ an index satisfying **(A 1-9)**. Then:*

(1) Sliding Regret: $\mathrm{SliReg}(I;T)=(\mu_1-\mu_2)T$.

(2) Regret of Exploration: *Let $(X_t : t\geq 1)$ a sequence of i.i.d. random variables with distribution $\mathrm{B}(\mu_2)$. Let $\sigma_T := T\wedge\inf\{t\geq 1 : -\rho(1-\mu_2)+\frac{1}{t}\sum_{i=1}^{t}(X_t-\mu_2)\leq 0\}$. We have* $\mathrm{RegExp}(I;T)\geq(\mu_1-\mu_2)\mathbf{E}[\sigma_T]$.

14.4.4 Examples and experiments

Checking that an index satisfies the requirements **(A 1-9)** is mostly computations. Example 1 details the checking process for IMED. More examples are provided in Table 14.4.2.

Example 1 (IMED). IMED *from Honda and Takemura (2015) picks the arm maximizing:*

$$I_a(t) := \frac{\log(t)}{N_a(t)\mathrm{kl}(\hat{\mu}_a(t),\hat{\mu}^*(t))+\log N_a(t)}.$$

*We have $I_{\max}=\infty$, and **(A 1-3)** are obvious. When the arm $a=1$ is pulled linearly often, we have $I_1(t)=\log(t)/\log N_1(t)\sim 1$, so $I(\mu_1,\mu_2)=1$. We see that $n_2(t):=\frac{\log(t)}{\mathrm{kl}(\mu_2,\mu_1)}$, that depends continuously on μ_2,μ_1 so that **(A 4)** is satisfied. **(A 5-6)** also immediately follow. The last conditions are the ones that need more work, but they result from straight forward computations. Asymptotically, for $\hat{\mu}_2\equiv\hat{\mu}_2(t)<\hat{\mu}_1(t)\equiv\hat{\mu}_1$, we get:*

$$I_1(t+h)-I_1(t)\sim\frac{h}{t}+\frac{N_1(t+h)-N_1(t)}{N_1(t)\log N_1(t)}\sim\frac{h}{t},$$

Algorithm	Index	$n_2(t)$	$\partial_{\mu_2} I_2$	$\partial_n I_2$
UCB	$\hat{\mu}_a(t)+\sqrt{\frac{2\log(t)}{N_a(t)}}$	$\frac{2\log(t)}{(\mu_1-\mu_2)^2}$	1	$-\frac{\mu_1-\mu_2}{2n_2(t)}$
MOSS	$\hat{\mu}_a(t)+\sqrt{\frac{\log\left(\frac{t}{2N_a(t)}\right)}{N_a(t)}}$	$\frac{\log(t)}{(\mu_1-\mu_2)^2}$	1	$-\frac{\mu_1-\mu_2}{2n_2(t)}$
UCB-V	$\hat{\mu}_a(t)+\sqrt{\frac{2\hat{\mu}_a(t)(1-\hat{\mu}_a(t))\log(t)}{N_a(t)}}+\frac{3c\log(t)}{N_a(t)}$	$\frac{2\mu_2(1-\mu_2)}{(\mu_1-\mu_2)^2}\left(1+\sqrt{1+\frac{6c(\mu_1-\mu_2)}{\mu_2(1-\mu_2)}}\right)^2\log(t)$	$(*)$	$(**)$
KLUCB	$\max\{\mu : N_a(t)\mathrm{kl}(\hat{\mu}_a(t),\mu)\le\log(t)\}$	$\frac{\log(t)}{\mathrm{kl}(\mu_2,\mu_1)}$	$\frac{\log\left(\frac{\mu_1(1-\mu_2)}{\mu_2(1-\mu_1)}\right)}{\frac{\mu_1-\mu_2}{\mu_1(1-\mu_1)}}$	$-\frac{\mathrm{kl}(\mu_2,\mu_1)}{n_2(t)\frac{\mu_1-\mu_2}{\mu_1(1-\mu_1)}}$
IMED	$\frac{\log(t)}{N_a(t)\mathrm{kl}(\hat{\mu}_a(t),\hat{\mu}^*(t))+\log N_a(t)}$	$\frac{\log(t)}{\mathrm{kl}(\mu_2,\mu_1)}$	$\frac{\log\left(\frac{\mu_1(1-\mu_2)}{\mu_2(1-\mu_1)}\right)}{\mathrm{kl}(\mu_2,\mu_1)}$	$-\frac{1}{n_2(t)}$

Table 14.4.2: Examples of asymptotic regimes. The missing entries of UCB-V are $(*)$ $\partial\mu_2 I_2 := 1+(1-2\mu_2)\sqrt{\frac{\log(t)}{2n_2(t)\mu_2(1-\mu_2)}}$ and $(**)$ $\partial_n I_2 := -\frac{\log(t)}{n_2(t)}\left(\sqrt{\frac{\mu_2(1-\mu_2)\log(t)}{2n_2(t)}}+\frac{6c\log(t)}{n_2(t)}\right)$. In this array, we can group algorithms in three families of algorithms with similar asymptotic regimes and identical ratios $n_2(t)\partial_{\mu_2}I_2/\partial_n I_2$, known to account for the regret of exploration: UCB and MOSS, UCB-V, KLUCB and IMED.

$$I_2(t+h)-I_2(t)\sim\frac{(\hat{\mu}_2(t+h)-\hat{\mu}_2(t))\log\left(\frac{\hat{\mu}_1(1-\hat{\mu}_2)}{\hat{\mu}_2(1-\hat{\mu}_1)}\right)}{\mathrm{kl}(\hat{\mu}_2,\hat{\mu}_1)}-\frac{N_2(t+h)-N_2(t)}{\frac{\log(t)}{\mathrm{kl}(\hat{\mu}_2,\hat{\mu}_1)}}+\frac{h}{t\log(t)}.$$

*These two Taylor expansions are continuous in $\hat{\mu}_2$ and $\hat{\mu}_1$, so that **(A 5)** is satisfied. **(A 9)** follows directly and **(A 8)** can be checked numerically. Following Theorem IV.37,*

$$\mathrm{SliReg}(\mathtt{IMED};T)=(\mu_1-\mu_2)T$$

and its regret of exploration can be predicted via the random walk specified in Theorem IV.37.2.

Example 2 (Experiments)**.** *We extend the experiment of Figure 14.3.1 to other index policies. We estimate the function* $\mathrm{RegExp}'(t;T):=\mathbf{E}[\mathrm{Reg}(t;t+T)|\exists k, t=\tau_k]$ *as a function of* t*. To estimate this function, we repeatedly run the algorithm to obtain a family* S *of samples of* $(\tau_k,\mathrm{Reg}(\tau_k;\tau_k+T))$*. Then, we approximate* $\mathrm{RegExp}'(t;T)$ *as the averages of* y *for* $(x,y)\in S$ *such that* $|x-t|<W$ *where* W *is a parameter of the approximation. In the experiments, we take* $W=128$ *and* $T=100$*.*

We overall observe a convergence to the predicted theoretical value (Theorem IV.37.2). Observing the precise rate of convergence of $\mathrm{RegExp}'(t;T)$ *as a function of* t *is rather difficult, especially for* IMED *and* KLUCB*, because these algorithms are very aggressive and rarely pick the suboptimal arm, meaning that there a only a few exploration episodes during a run. The amount of data required to accurately estimate the curve increases exponentially with* t*. Nonetheless, it seems that* $\mathrm{RegExp}'(t;T)$ *is slightly below the theoretical* $(\mu_1-\mu_2)\mathbf{E}[\sigma_T]$ *sometimes, see* IMED *for instance. This is due to two things. First, although we eventually have* $|N_2(t)-n_2(t)|<\epsilon n_2(t)$ *with* ϵ *as small as desired, for* $t=10000$*, the correct* ϵ *may remain large. For instance, in* IMED*'s index, the term* $\log n_2(t)$ *cannot be neglected in front of* $n_2(t)\mathrm{kl}(\mu_2,\mu_1)$ *even when* $t=10000$*, implying that* $N_2(t)$ *and* $n_2(t)$ *are of the same order but still a bit far away. Second, the analysis assumes that the partial derivatives of the index stay approximately the same over* $[t;t+T]$*, which is quite imprecise when* t *isn't large enough in front of* T*.*

14.5 Future directions

The big take-away of this chapter is that the locally bad behavior of UCB, and more generally index policies, is hard-coded in their design. By blindly allocating visits according to a lower bound, they will have burst of sub-optimal play when this lower bound suddenly jumps and

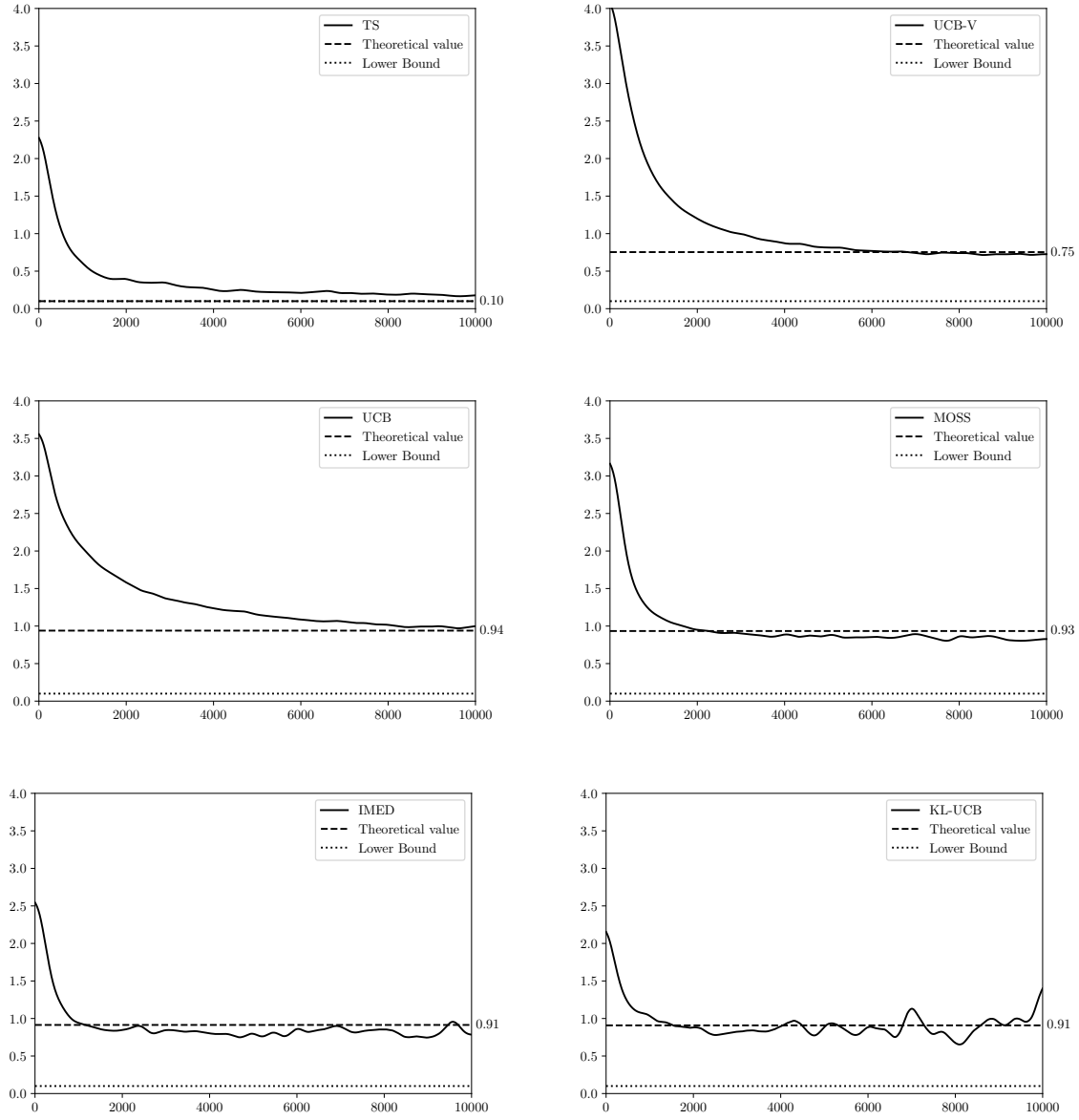


Figure 14.4.1: Estimated regret of exploration for various algorithms on the two-arm Bernoulli bandit $(\mathtt{B}(0.9), \mathtt{B}(0.8))$. The performance of `TS` and `UCB` from Figure 14.3.1 are compared to `UCB-V`, `MOSS`, `IMED` and `KLUCB`. The lower bound (dotted) of RegExp(—) is 0.1. The theoretical value is reported to the right.

requires for pulls of sub-optimal actions. By being randomized, or maybe lazy, methods such as TS or MED are immune to this behavior.

Many things are still to be done.

First, a few results above are incomplete. For instance, Theorem IV.33 lower-bounds the regret of exploration of UCB by a random walk. Is there equality? In Chapter 13, we have presented a technique showing that $\mathrm{RegExp}(T) = \mathrm{O}(\log(T))$ and this technique is probably applicable to UCB, but the return time to $\mathbf{R}_-$ of the random walk is more likely $\mathrm{O}(1)$ because of the negative drift. I conjecture that, for UCB, $\mathrm{RegExp}(T) = \mathrm{O}(1)$ and can be characterized by the return time of the random walk of Theorem IV.33 up to a infinitesimal randomization of the means μ_1, μ_2.

Second, a possible direction is to go even beyond the sliding regret. The motivation is EXP3 Auer et al. (1995). Not only the proof techniques presented in this chapter (Theorem IV.24 and asymptotical regimes) fail to apply to EXP3, but its behavior seems unique. It sits in-between the smooth behavior of TS and the bumpy behavior of UCB. Actually, the expected visit counts of EXP3 seem to differ from their almost sure values that don't even seem to exist in the first place, meaning that EXP3 is heavily unstable. This is not that much of a surprise, because EXP3 has been designed to tackle **adversarial bandits**. Yet, any attempt at understanding the behavior of EXP3 is likely to introduce interesting proof techniques.

The third perspective is obviously to extend this work beyond stochastic bandits. While our proofs can be adapted to cover multi-arm bandits with non-Bernoulli reward distributions, they are specific to stochastic bandits. However, I believe that it is too early to pursue this direction. First, the EVI-based algorithms cannot go beyond of regret of exploration guarantees because of Proposition IV.30. Second, the literature of regret efficient algorithms in the model dependent setting for average reward Markov decision processes is far from stabilized. The only algorithm to achieve the lower bound is ECoE (Algorithm III.2), was born in the end of July 2024 and has no implementable version yet, hence cannot be fairly called "an algorithm." Of course, the literature on recurrent models is a bit more stable and one may try to analyze IMED-RL Pesquerel and Maillard (2022). However, especially regarding the minor structure of recurrent models and Proposition III.8, recurrent models resemble bandits too much and nothing much new is to discover there.

Appendix of Chapter 14

14.A Almost-sure properties of consistent algorithms

This section is dedicated to the proof of the following result.

Proposition 14.A.1. *Consider a policy such that whatever the distributions* F *on arms, the expected regret grows linearly, i.e.,* $\mathbf{E}_{\mathrm{F}}[\mathrm{Reg}(T)] = \mathrm{o}(T)$. *Then all arms are visited infinitely often, that is,* $\mathbf{P}(\forall n, \exists t : N_a(t) \geq n) = 1$.

Proof. (**STEP 1**) Assume on the contrary that, for some distributions F on arms and for some arm a, $\mathbf{P}_{\mathrm{F}}(\forall t : N_a(t) < n) > 0$ where $n \geq 1$. Because the expected regret is sublinear, a has to be suboptimal. Let F' any distribution on arms making a the unique optimal arm, and such that $\mathrm{F}(a') = \mathrm{F}'(a')$ for all $a' \neq a$. Denote the likelihood-ratio of the observations $(A_1, R_1, \ldots, A_{t-1}, R_{t-1})$ as

$$L_t \equiv L(A_1, R_1, \ldots, A_{t-1}, R_{t-1}) := \sum_b \sum_{s=1}^{t-1} \mathbf{1}(A_s = b) \frac{f_b(R_s)}{f'_b(R_s)}$$

Denoting $\mathscr{F}_t := \sigma(A_1, R_1, \ldots, A_{t-1}, R_{t-1})$, it is known (see Kaufmann et al. 2016, Lemma 18) that if E is a $\mathscr{F}_t$-measurable event, then

$$\mathbf{P}_{\mathrm{F}'}(E) = \mathbf{E}_{\mathrm{F}}[\mathbf{1}(E) \exp(-L_t)].$$

(**STEP 2**) Because F is non-degenerate with $0 < \mu_a < \mu^* < 1$, we can assume that $\mu'_a < 1$ and that there exists $c > 0$ such that for $r \in \{0, 1\}$, we have $\frac{1}{c} \leq \exp(f_a(r)/f'_a(r)) \leq c$. Then,

$$\begin{aligned}
\mathbf{P}_{\mathrm{F}'}(\forall t : N_a(t) < n) &= \lim_{t\to\infty} \mathbf{P}_{\mathrm{F}'}(N_a(t) < n) \\
&= \lim_{t\to\infty} \mathbf{E}_{\mathrm{F}}\left[\mathbf{1}(N_a(t) < n) \exp\left(-\sum_b \sum_{s=1}^{t-1} \mathbf{1}(A_s = b) \frac{f_b(R_s)}{f'_b(R_s)} \right) \right] \\
&= \lim_{t\to\infty} \mathbf{E}_{\mathrm{F}}\left[\mathbf{1}(N_a(t) < n) \prod_{s=1}^{t-1} \exp\left(-\mathbf{1}(A_s = b) \frac{f_a(R_s)}{f'_a(R_s)} \right) \right] \\
&\geq \lim_{t\to\infty} \mathbf{E}_{\mathrm{F}}\left[\mathbf{1}(N_a(t) < n) \left(\frac{1}{c} \right)^{N_a(t)} \right] \\
&\geq c^{-n} > 0.
\end{aligned}$$

(**STEP 3**) Let $\Delta' > 0$ the gap between the optimal arm and the best suboptimal arm under F'. We get

$$\mathbf{E}_{\mathrm{F}'}[N_a(T)] \leq n\mathbf{P}_{\mathrm{F}'}(N_a(T) < n) + T\mathbf{P}_{\mathrm{F}'}(N_a(T) \geq n) \leq n + T(1 - c^{-n}).$$

So:

$$\begin{aligned}\mathbf{E}_{\mathrm{F}'}[\mathrm{Reg}(T)] &\geq \Delta'(T - \mathbf{E}_{\mathrm{F}'}[N_a(T)]) \\ &\geq \Delta'(c^{-n}T - n) = \Omega(T).\end{aligned}$$

This comes in contradiction with $\mathbf{E}_{\mathrm{F}'}[\mathrm{Reg}(T)] = o(T)$. □

14.B Analysis of Thompson Sampling

14.B.1 Preliminaries: Sanov's Theorem

Our analysis of Thompson Sampling relies on a quantitative version of Sanov's Theorem.

Lemma 14.B.1 (Sanov's Theorem). *Let $q \in (0,1)$ and $(X_n : n \geq 1)$ a family of i.i.d. random variables with distribution $\mathrm{B}(q)$. Let $S_n := X_1 + \ldots + X_n$ and denote $\mathbf{P}_q(S_n \in \ldots)$ the induced probability distribution. Then, for $\epsilon > \frac{1}{n}$,*

$$\frac{1}{n+1}e^{-n\mathrm{kl}(q-\epsilon-\frac{1}{n},q)} \leq \mathbf{P}_q(S_n \leq n(q-\epsilon)) \leq ne^{-n\mathrm{kl}(q-\epsilon,q)}, \tag{14.B.1}$$

$$\frac{1}{n+1}e^{-n\mathrm{kl}\left(q+\epsilon+\frac{1}{n},q\right)} \leq \mathbf{P}_q(S_n \geq n(q+\epsilon)) \leq ne^{-n\mathrm{kl}(q+\epsilon,q)}. \tag{14.B.2}$$

Proof. Naming these inequalities "Sanov's Theorem" is a bit of an overstatement but is nonetheless very close to the original. The proof is classic, but we write it below for the paper to be self-contained.

(**STEP 1**) We start by a combinatorial lemma. Write $h(p) := -p\log(p) - (1-p)\log(1-p)$ the Shannon entropy. For all n and $k \in \{0,\ldots,n\}$, we have

$$\frac{e^{nh\left(\frac{k}{n}\right)}}{n+1} \leq \binom{n}{k} \leq e^{nh\left(\frac{k}{n}\right)}.$$

To establish this, remark that $1 = \sum_\ell \binom{n}{\ell}(\frac{k}{n})^\ell(1-\frac{k}{n})^{n-\ell}$. The term for $\ell = k$ is equal to $e^{-nh(k/n)}$. In particular, we have $1 \geq \binom{n}{k}e^{-nh(k/n)}$, giving the upper bound above. But also, since the term for $\ell = k$ is the largest of the sum, we get $1 \leq (n+1)\binom{n}{k}e^{-nh(k/n)}$, leading to the lower bound.

(**STEP 2**) Let $k \in \{0,\ldots,n\}$. We have

$$\begin{aligned}\mathbf{P}_q(S_n = k) &= \binom{n}{k}q^k(1-q)^{n-k} \\ &= \binom{n}{k}\left(q^{\frac{k}{n}}(1-q)^{1-\frac{k}{n}}\right)^n \\ &= \binom{n}{k}e^{-nh\left(\frac{k}{n}\right)} \cdot e^{-n\mathrm{kl}\left(\frac{k}{n},q\right)}.\end{aligned}$$

We therefore obtain

$$\frac{e^{-n\mathrm{kl}\left(\frac{k}{n},q\right)}}{n+1} \leq \mathbf{P}_q(S_n = k) \leq e^{-n\mathrm{kl}\left(\frac{k}{n},q\right)}.$$

(**STEP 3**) We establish the bounds for $\mathbf{P}_q(S_n \leq n(q-\epsilon))$, see (14.B.1).

$$\mathbf{P}_q(S_n \leq n(q-\epsilon)) = \sum_{k=0}^{\lfloor n(q-\epsilon)\rfloor} \mathbf{P}_q(S_n = k).$$

For the upper bound, remark that when over $\{0,\dots,\lfloor n(q-\epsilon)\rfloor\}$, the function $k \mapsto e^{-n\mathrm{kl}\left(\frac{k}{n},q\right)}$ is decreasing, thus we get

$$\mathbf{P}_q(S_n \le n(q-\epsilon)) \le (n-\lfloor n(q-\epsilon)\rfloor)e^{-n\mathrm{kl}\left(\frac{\lfloor n(q-\epsilon)\rfloor}{n},q\right)} \le ne^{-n(\mathrm{kl}(q-\epsilon,q))}.$$

Rearranging terms provides the upper bound part in Sanov's Theorem. For the lower bound, check that

$$\mathbf{P}_q(S_n \le n(q-\epsilon)) \ge \frac{e^{-n\mathrm{kl}\left(\frac{\lfloor n(q-\epsilon)\rfloor}{n},q\right)}}{n+1} \ge \frac{e^{-n\mathrm{kl}(q-\epsilon-\frac{1}{n},q)}}{n+1}.$$

The bounds for $\mathbf{P}_q(S_n \ge n(q+\epsilon))$, see (14.B.2), are established similarly. □

14.B.2 The almost-sure asymptotic behavior of Thompson Sampling

Starting from this section, we assume throughout that the bandit model is $(\mathrm{B}(\mu_1),\mathrm{B}(\mu_2))$ with non-degenerate means $0<\mu_2<\mu_1<1$. We first bound the sampling rates of TS.

Lemma 14.B.2. *There exist a positive definite function $c:\mathbf{R}_+\to\mathbf{R}$ and a family $(n_\epsilon:\epsilon>0)$ such that for all $\epsilon>0$, there exists a sequence of events (G_t^ϵ) with $\liminf G_t^\epsilon$ a.s., and such that:*

$$e^{-N_2(t)(1+c(\epsilon))\mathrm{kl}(\mu_2,\mu_1)} \le \mathbf{E}\big[\mathbf{1}(A_t=2)\mid G_t^\epsilon, N_2(t), N_2(t)\ge n_\epsilon\big] \le e^{-N_2(t)(1-c(\epsilon))\mathrm{kl}(\mu_2,\mu_1)}.$$

Proof. (**STEP 1**) Denote $F^{\mathrm{Beta}}_{\alpha,\beta}$ (respectively $F^{\mathrm{Bin}}_{n,p}$) the c.d.f. of a Beta distribution $\mathrm{Beta}(\alpha,\beta)$ (respectively a Binomial distribution $\mathrm{Binom}(n,p)$). Using the Beta-Binomial trick and Lemma 14.B.1, we obtain:

$$\begin{aligned}
\mathbf{E}[\mathbf{1}(\theta_a(t)\le\hat\mu_a(t)-\epsilon)\mid\hat\mu_a(t),N_a(t)] &= F^{\mathrm{Beta}}_{1+S_a(t),1+N_a(t)-S_a(t)}(\hat\mu_a(t)-\epsilon)\\
&\overset{(*)}{=} 1-F^{\mathrm{Bin}}_{N_a(t)+1,\hat\mu_a(t)-\epsilon}(S_a(t))\\
&\overset{(\dagger)}{\le} \mathbf{E}\Big[(N_a(t)+1)e^{-(N_a(t)+1)\mathrm{kl}\left(\frac{S_a(t)}{N_a(t)+1},\hat\mu_a(t)-\epsilon\right)}\mid\hat\mu_a(t),N_a(t)\Big]\\
&= (N_a(t)+1)e^{-(N_a(t)+1)\mathrm{kl}\left(\frac{S_a(t)}{N_a(t)+1},\hat\mu_a(t)-\epsilon\right)}.
\end{aligned}$$

We can similarly derive a bound for $\mathbf{1}(\theta_a(t)\ge\hat\mu_a(t)+\epsilon)$, showing that:

$$\mathbf{E}[\mathbf{1}(\theta_a(t)\ge\hat\mu_a(t)+\epsilon)\mid\hat\mu_a(t),N_a(t)] \ge \frac{e^{-(N_a(t)+1)\mathrm{kl}\left(\frac{S_a(t)}{N_a(t)+1},\hat\mu_a(t)+\epsilon\right)}}{N_a(t)+2}.$$

where $(*)$ follows by Beta-Binomial Trick and $(\dagger)$ follows from Sanov's Theorem (Lemma 14.B.1).

(**STEP 2**) Introduce the events $F_t := (|\hat\mu_1(t)-\mu_1|<\epsilon/3)$ and $E_t := (N_1(t)>t^b)$ that are such that both $\liminf F_t$ and $\liminf E_t$ are almost-sure (see Kaufmann et al. 2012, Proposition 1 for E_t). We have

$$\begin{aligned}
\mathbf{P}(\forall t,\exists s\ge t:\theta_1(s)\le\mu_1-\epsilon) &= \lim_{t\to\infty}\mathbf{P}\big(\exists s\ge t:\theta_1(s)\le\mu_1-\epsilon,F_s^\epsilon,E_s\big)\\
&\le \lim_{t\to\infty}\sum_{s\ge t}\mathbf{P}\big(\theta_1(s)\le\hat\mu_1(s)-\tfrac{2\epsilon}{3},F_s^\epsilon,E_s\big)\\
&\le \lim_{t\to\infty}\sum_{s\ge t}\mathbf{E}\big[\mathbf{1}(\theta_1(s)\le\hat\mu_1(s)-\tfrac{2\epsilon}{3})\mid F_s^\epsilon,E_s,N_1(s),\hat\mu_1(s)\big]\\
&\le \lim_{t\to\infty}\sum_{s\ge t}(t^b+1)e^{-(t^b+1)\mathrm{kl}\left(\mu_1-\frac{\epsilon}{3},\mu_1-\frac{2\epsilon}{3}\right)}\\
&= 0.
\end{aligned}$$

Accordingly, $\mathbf{P}(\exists t, \forall s \geq t : \theta_1(s) > \mu_1 - \epsilon) = 1$. Similarly one can show that $\mathbf{P}(\exists t, \forall s \geq t : \theta_1(s) < \mu_1 + \epsilon) = 1$.

(**STEP 3**) Following (**STEP 2**), we see that in the asymptotic regime, the suboptimal arm $a = 2$ cannot be picked unless $\theta_2(t) \geq \mu_1 - \epsilon$. And conversely, if $\theta_2(t) \geq \mu_1 + \epsilon$, then arm 2 is pulled. Introduce the asymptotically almost sure event:

$$G_t^\epsilon := (|\theta_1(t) - \mu_1| < \epsilon) \cap (|\hat{\mu}_2(t) - \mu_2| < \epsilon).$$

The probability to pick $A_t = 2$ conditionally on G_t^ϵ is bounded accordingly:

$$\frac{e^{-(N_2(t)+1)\mathrm{kl}(\mu_2-\epsilon,\mu_1+\epsilon)}}{N_2(t)+2} \leq \mathbf{E}\big[\mathbf{1}(A_t = 2) \mid G_t^\epsilon, N_2(t)\big] \leq (N_2(t)+1)e^{-(N_2(t)+1)\mathrm{kl}(\mu_2+\epsilon,\mu_1-\epsilon)}.$$

Therefore, there exists a positive definite function $c(\epsilon)$ such that, when $N_2(t)$ is large enough relatively to ϵ, say $N_2(t) \geq n_\epsilon$, we have:

$$e^{-N_2(t)(1+c(\epsilon))\mathrm{kl}(\mu_2,\mu_1)} \leq \mathbf{E}\big[\mathbf{1}(A_t = 2) \mid G_t^\epsilon, N_2(t), N_2(t) \geq n_\epsilon\big] \leq e^{-N_2(t)(1-c(\epsilon))\mathrm{kl}(\mu_2,\mu_1)},$$

establishing the claim. □

The second result of this section provides a precise description of TS's visit rates at infinity. The visit rate of the suboptimal, $N_2(t)$, will be later called the *asymptotic regime* of TS.

Lemma 14.B.3. *For all $\delta > 0$,*

$$\mathbf{P}\left(\exists t, \forall s \geq t : \left|N_2(s) - \frac{\log(s)}{\mathrm{kl}(\mu_2,\mu_1)}\right| \leq \delta \cdot \frac{\log(s)}{\mathrm{kl}(\mu_2,\mu_1)}\right) = 1.$$

Proof. Let $\delta > 0$. Denote $c \equiv c(\epsilon)$ and $k_0 \equiv \mathrm{kl}(\mu_2,\mu_1)$ for short, and choose ϵ small enough so that $\frac{1}{1+2c(\epsilon)} > 1-\delta$ and $3c(\epsilon) < \delta$.

(**STEP 1**) We show that the event $(N_2(t) \geq \frac{\log(t)}{(1+2c)k_0})$ holds eventually. We proceed by considering the complementary event and denote $\lambda_t := \frac{\log(t)}{(1+2c)k_0}$. From Lemma 14.B.2 we also know that

$$\mathbf{E}\big[\mathbf{1}(A_s \neq 2) \mid G_s^\epsilon, n_\epsilon \leq N_2(s) \leq \lambda_t\big] \leq 1 - e^{-\frac{(1+c)\log(t)}{1+2c}}.$$

Denote $E_s^\epsilon := G_s^\epsilon \cap (n_\epsilon \leq N_2(s) \leq \lambda_s)$ for short. Remark that if $N_2(s) < \lambda_s$, then the arm $a = 2$ has been sampled less than λ_s times over $[\frac{1}{2}s, s]$ with $N_2(u) < \lambda_s$ each time. Said differently, there exists Λ a subset of $[\frac{1}{2}s, s]$ with at least $\frac{s}{2} - \lambda_s$ elements (we write $\Lambda \subseteq_{s/2-\lambda_s} [\frac{1}{2}s, s]$) such that for all $i \in \Lambda$, $A_i = 1$. Therefore, where F_j^ϵ below is a shorthand for $(A_j = 1, N_2(j) < \lambda_s, E_j^\epsilon)$, we have:

$$\begin{aligned}
\mathbf{P}(\forall t, \exists s \geq t : N_2(s) < \lambda_s) &= \lim_{t\to\infty} \mathbf{P}(\exists s \geq t : N_2(s) < \lambda_s) \\
&\leq \lim_{t\to\infty} \mathbf{P}\big(\exists s \geq t, \exists \Lambda \subseteq_{s/2-\lambda_s} [\tfrac{1}{2}s, s], \forall i \in \Lambda : A_i = 1, N_2(i) < \lambda_s\big) \\
&= \lim_{t\to\infty} \mathbf{P}\big(\exists s \geq t, \exists \Lambda \subseteq_{s/2-\lambda_s} [\tfrac{1}{2}s, s], \forall i \in \Lambda : A_i = 1, N_2(i) < \lambda_s, E_i^\epsilon\big) \\
&= \lim_{t\to\infty} \sum_{s\geq t} \sum_{I \subseteq_{s/2-\lambda_s} [\frac{1}{2}s,s]} \prod_{i\in I} \mathbf{P}\Big(A_i = 1, N_2(i) < \lambda_s, E_i^\epsilon \mid \forall j < i \in I : F_j^\epsilon\Big) \\
&\leq \lim_{t\to\infty} \sum_{s\geq t} \sum_{I \subseteq_{s/2-\lambda_s} [\frac{1}{2}s,s]} \prod_{i\in I} \mathbf{P}\Big(A_i = 1 \mid N_2(i) < \lambda_s, E_i^\epsilon, (\forall j < i \in I : F_j^\epsilon)\Big) \\
&\leq \lim_{t\to\infty} \sum_{s\geq t} \sum_{I \subseteq_{s/2-\lambda_s} [\frac{1}{2}s,s]} \Big(1 - e^{-\frac{(1+c)\log(t)}{1+2c}}\Big)^{\frac{t}{2} - \frac{\log(t)}{(1+2c)k_0}}
\end{aligned}$$

$$\leq \lim_{t\to\infty} \sum_{s\geq t} \binom{t/2}{\frac{\log(t)}{(1+2c)k_0}} \left(1-e^{-\frac{(1+c)\log(t)}{1+2c}}\right)^{\frac{t}{2}-\frac{\log(t)}{(1+2c)k_0}}.$$

Using standard equivalents, the summand happens to be asymptotically upper bounded by

$$e^{C\log^2(t)} \cdot e^{-Ct^{1-\frac{1+c}{1+2c}}} \lesssim e^{-C't^{\frac{c}{1+2c}}}$$

where $C > C' > 0$. This term has finite sum. We conclude has follows:

$$\mathbf{P}\Big(\forall t, \exists s \geq t : N_2(t) < \tfrac{\log(s)}{(1+2c)k_0}\Big) \leq \lim_{t\to\infty} \sum_{s\geq t} e^{-C's^{\frac{c}{1+2c}}} = 0.$$

(**STEP 2**) We show that the event $(N_2(t) \leq \frac{(1+3c)\log(t)}{k_0})$ holds eventually. Again, we consider the complementary event and denote $\lambda_t := \frac{1}{k_0}\log(t)$. By Lemma 14.B.2,

$$\mathbf{E}\big[\,\mathbf{1}(A_s = 2) \mid G_s^\epsilon, n_\epsilon \leq N_2(s), (1+2c)\lambda_t < N_2(s)\big] \leq e^{-(1+c-2c^2)\log(t)}.$$

Let $E_s^\epsilon := (G_s^\epsilon) \cap (n_\epsilon \leq N_2(s)) \cap ((1+2c)\lambda_t < N_2(s))$ for short. Note that if $N_2(t) > (1+3c)\lambda_t$, then it has been sampled at least $c\lambda_t$) times with $N_2(s) > (1+2c)\lambda_t$ over the time interval $[(1+2c)\lambda_t, t]$. So, there exists Λ a subset of $[(1+2c)\lambda_t, t]$ of size at most $c\lambda_t$ (we write $\Lambda \subseteq_{c\lambda_t} [(1+2c)\lambda_t, t]$) such that for all $i \in \Lambda$, $A_i = 2$. Therefore, and where F_j^ϵ below is a shorthand for $(A_j = 2, N_2(j) > (1+2c)\lambda_j, E_j^\epsilon)$, we have

$$\begin{aligned}
(-) &:= \mathbf{P}(\forall t, \exists s \geq t : N_2(s) > (1+3c)\lambda_s) \\
&= \lim_{t\to\infty} \mathbf{P}(\exists s \geq t : N_2(s) > (1+3c)\lambda_s) \\
&\leq \lim_{t\to\infty} \mathbf{P}\Big(\exists s \geq t, \exists \Lambda \subseteq_{c\lambda_s} [(1+2c)\lambda_s, s], \forall i \in \Lambda : A_i = 2, N_2(i) > (1+2c)\lambda_s\Big) \\
&= \lim_{t\to\infty} \mathbf{P}\Big(\exists s \geq t, \exists \Lambda \subseteq_{c\lambda_s} [(1+2c)\lambda_s, s], \forall i \in \Lambda : A_i = 2, N_2(i) > (1+2c)\lambda_s, E_i^\epsilon\Big) \\
&= \lim_{t\to\infty} \sum_{s\geq t} \sum_{I \subseteq_{c\lambda_s} [(1+2c)\lambda_s, s]} \prod_{i\in I} \mathbf{P}\Big(A_i = 2, N_2(i) > (1+2c)\lambda_s, E_i^\epsilon \mid \forall j < i \in I : F_j^\epsilon\Big) \\
&= \lim_{t\to\infty} \sum_{s\geq t} \sum_{I \subseteq_{c\lambda_s} [(1+2c)\lambda_s, s]} \prod_{i\in I} \mathbf{P}\Big(A_i = 2 \mid N_2(i) > (1+2c)\lambda_s, E_i^\epsilon, (\forall j < i \in I : F_j^\epsilon)\Big) \\
&\leq \lim_{t\to\infty} \sum_{s\geq t} \binom{s}{\frac{c\log(s)}{k_0}} e^{-(1+c-2c^2)\log(s)\cdot\frac{c}{k_0}\log(s)}.
\end{aligned}$$

Using standard equivalents, the summand is asymptotically upper bounded by

$$e^{\left(\frac{c}{k_0}+o(1)\right)\log^2(t)} \cdot e^{-(1+c-2c^2)\frac{c}{k_0}\log^2(t)} = e^{-(c-2c^2+o(1))\frac{c}{k_0}\log^2(t)}.$$

Again, this has finite sum. We conclude:

$$\mathbf{P}(\forall t, \exists s \geq t : N_2(s) > (1+3c)\lambda_s) = 0.$$

This concludes the proof. □

14.B.3 Proof of Theorem IV.25

Proof of Theorem IV.25. We conclude that Thompson Sampling has optimal sliding regret. Fix $T \geq 1$. Combining Lemma 14.B.2 and Lemma 14.B.3, we see that for all $\epsilon > 0$, there exists a sequence of events (E_t^ϵ) with $\mathbf{P}(\liminf E_t^\epsilon) = 1$, and such that:

$$\mathbf{P}\big(A_t = 2 \mid E_t^\epsilon\big) \leq e^{-(1-\epsilon)\log(t)} = \frac{1}{t^{1-\epsilon}}.$$

Since all arms are visited infinitely often, eventually T is negligible in front of $N_2(t)$, meaning that for all partial history $H_{t:t+h} = h_{t:t+h}$ over $[t, t+h]$ (with $h \leq T$), we will have

$$\mathbf{P}\big(A_{t+h} = 2 \mid E_t^\epsilon, H_{t:t+h} = h_{t:t+h}\big) \leq \frac{1}{t^{1-\epsilon}}.$$

Conclude with Theorem IV.24. □

14.C Analysis of UCB

14.C.1 The asymptotic regime of UCB

Proposition IV.27. *For all $\epsilon > 0$ and when running* UCB*, both of the following hold:*

(1) $\mathbf{P}(\exists t, \forall s \geq t : \forall a, |\mu_a(s) - \mu_a(s)| < \epsilon) = 1$;

(2) $\mathbf{P}\Big(\exists t, \forall s \geq t : \Big|N_2(t) - 2\Big(\frac{1}{\mu_1-\mu_2}\Big)^2 \log(t)\Big| < \epsilon \cdot 2\Big(\frac{1}{\mu_1-\mu_2}\Big)^2 \log(t)\Big) = 1.$

Proof of Proposition IV.27. Because UCB has sublinear expected regret, all arms are visited infinitely often by Proposition 14.A.1, hence by the Strong Law of Large numbers, the empirical estimates of every arm converge to their true means. This proves Proposition IV.27.1. We will denote $E_t^\epsilon := (\forall a : |\hat{\mu}_a(t) - \mu_a| < \epsilon)$. We now focus on the proof of Proposition IV.27.2. Denote $\lambda_t := \frac{2}{(\mu_1-\mu_2)^2}\log(t)$ the theoretical visit rate of arm 2

(**STEP 1**) Let $\epsilon > 0$. We show that the event $(N_2(t) > (1-\epsilon)\lambda_t)$ holds eventually. As usual, we proceed by considering the complementary event. Let $\delta > 0$. Remark that if the arm $a = 2$ has been visited less than $(1-\epsilon)\lambda_s$ times, then the other arm $a = 1$ must have been pulled within the time range $\{s - \lambda_s - 1, s\}$, hence within $[\frac{1}{2}s, s]$ provided that s is large enough. Since $\lambda_s = o(s)$, we can in addition assume that when $a = 1$ is pulled, $N_1(s) \geq \frac{1}{2}s$. Accordingly, and denoting $F_u^\epsilon := (N_2(u) \leq (1-\epsilon)\lambda_s) \cap (N_1(u) \geq \frac{1}{2}s)$ for short, we have:

$$\begin{aligned}
(-) &:= \mathbf{P}(\forall t, \exists s \geq t : N_2(s) \leq (1-\epsilon)\lambda_s) \\
&= \lim_{t\to\infty} \mathbf{P}(\exists s \geq t : N_2(s) \leq (1-\epsilon)\lambda_s,) \\
&\leq \lim_{t\to\infty} \mathbf{P}\big(\exists s \geq t, \exists u \in [\tfrac{1}{2}s, s] : N_2(u) \leq (1-\epsilon)\lambda_s, N_1(u) \geq \tfrac{1}{2}s, A_u = 1\big) \\
&= \lim_{t\to\infty} \mathbf{P}\left(\exists s \geq t, \exists u \in [\tfrac{1}{2}s, s] : F_u^\epsilon, E_u^\delta, \hat{\mu}_2(u) + \sqrt{\frac{2\log(u)}{N_2(u)}} \leq \hat{\mu}_1(u) + \sqrt{\frac{2\log(u)}{N_1(u)}}\right) \\
&\leq \lim_{t\to\infty} \mathbf{P}\left(\exists s \geq t, \exists u \in [\tfrac{1}{2}s, s] : \mu_2 - \delta + \sqrt{\frac{2\log(\frac{1}{2}s)}{\frac{2(1-\epsilon)}{(\mu_1-\mu_2)^2}\log(s)}} \leq \mu_1 + \delta + \sqrt{\frac{2\log(s)}{\frac{1}{2}s}}\right) \\
&\leq \lim_{t\to\infty} \mathbf{1}\left(\frac{\mu_1-\mu_2}{\sqrt{1-\epsilon}} \cdot \sqrt{\frac{\log(\frac{1}{2}t)}{\log(t)}} \leq \mu_1 - \mu_2 + 3\delta\right).
\end{aligned}$$

In the above, $\delta > 0$ can be chosen arbitrarily small. Since $\sqrt{1-\epsilon} < 1$, we see that by choosing δ small regarding ϵ, we obtain $\mathbf{P}(\forall t, \exists s \geq t : N_2(s) \leq (1-\epsilon)\lambda_s) = 0$.

(**STEP 2**) Let $\epsilon > 0$. We now show that the event $(N_2(t) < (1+\epsilon)\lambda_t)$ holds eventually. Let $\delta > 0$. Observe that if $N_2(s) \geq (1+\epsilon)\lambda_s$, then arm 2 must have been pulled within the time range $\{(1+\epsilon)\lambda_s, \ldots, s\}$ with $N_2(u) \geq (1+\epsilon)\lambda_s$. Following this idea, we obtain:

$$(-) := \mathbf{P}(\forall t, \exists s \geq t : N_2(s) \geq (1+\epsilon)\lambda_s)$$

$$\begin{aligned}
&\leq \lim_{t\to\infty} \mathbf{P}(\exists s \geq t, \exists u \in [(1+\epsilon)\lambda_s, s] : N_2(u) \geq (1+\epsilon)\lambda_s, A_u = 2) \\
&= \lim_{t\to\infty} \mathbf{P}\big(\exists s \geq t, \exists u \in [(1+\epsilon)\lambda_s, s] : E_u^\delta, N_2(u) \geq (1+\epsilon)\lambda_s, A_u = 2\big) \\
&\leq \lim_{t\to\infty} \mathbf{P}\left(\exists s \geq t, \exists u \in [(1+\epsilon)\lambda_s, s] : \mu_2 + \delta + (\mu_1 - \mu_2)\sqrt{\frac{\log(u)}{(1+\epsilon)\log(s)}} \geq \mu_1 - \delta\right) \\
&\leq \lim_{t\to\infty} \mathbf{1}(\frac{\mu_1 - \mu_2}{\sqrt{1+\epsilon}} \geq \mu_1 - \mu_2 - 2\delta).
\end{aligned}$$

The above indicator is asymptotically 0 when δ is small enough. □

14.C.2 The sliding regret of UCB: Proof of Lemma IV.28

As given by Proposition IV.27, the asymptotic regime is denoted

$$E_t^\epsilon := (\forall a : |\hat{\mu}_a(t) - \mu_a| < \epsilon) \cap \Big(N_2(t) = \tfrac{2\cdot(1\pm\epsilon)}{(\mu_1-\mu_2)^2}\log(t)\Big)$$

Denote $I_a(t) := \hat{\mu}_a(t) + \sqrt{2\log(t)/N_a(t)}$ UCB's index of arm a.

Lemma 14.C.1. *Let $G_{t:t+h} := (\forall i < h, A_{t+i} = 2)$. For all $\epsilon > 0$ and $h \geq 1$, there exists $\delta, T > 0$ such that, for all $t > T$ and on $E_t^\delta \cap G_{t:t+h}$, we have:*

$$\left|(I_1(t+h) - I_2(t+h)) - \left(I_1(t) - I_2(t) - \frac{\sum_{i<h}(R_{t+i} - \mu_2 - \frac{\mu_1-\mu_2}{2})}{N_2(t)}\right)\right| \leq \frac{2h\epsilon}{N_2(t)}.$$

Proof. Fix $\epsilon > 0$. (**STEP 1**) The time variations of UCB's indexes are given by:

$$I_a(t+h) - I_a(t) = (\hat{\mu}_a(t+h) - \hat{\mu}_a(t)) + \left(\sqrt{\frac{2\log(t+h)}{N_a(t+h)}} - \sqrt{\frac{2\log(t)}{N_a(t)}}\right)$$

This is split into two terms. There is the variation of the empirical estimate, and the variation of the optimistic bonus. Considering arm 1, since $N_1(t+h) = N_1(t)$ on $G_{t:t+h}$, we get, when $t \to \infty$,

$$\begin{aligned}
I_1(t+h) - I_1(t) &= \sqrt{\frac{2}{N_1(t)}}\Big(\sqrt{\log(t+h)} - \sqrt{\log(t)}\Big) \\
&\sim \frac{\frac{h}{t}}{\sqrt{2\log(t)N_1(t)}} = \mathrm{o}\left(\frac{h}{t\sqrt{N_1(t)}}\right).
\end{aligned}$$

This will appear to be negligible in comparison to $I_2(t+h) - I_2(t)$.

(**STEP 2**) We know bound the variations of the empirical estimates of arm 2. We have:

$$\hat{\mu}_2(t+h) - \hat{\mu}_2(t) = \frac{\sum_{i<h} R_{t+i} - h\hat{\mu}_2(t)}{N_2(t)+h} = \frac{\sum_{i<h}(R_{t+i} - \mu_2)}{N_2(t)+h} + \frac{h(\mu_2 - \hat{\mu}_2(t))}{N_2(t)+h}.$$

Because arms are visited infinitely often, we have $\mu_2 - \mu_2(t) < \epsilon$ eventually, with $\epsilon > 0$ fixed. Since, $N_2(t) + h \sim N_2(t)$, hence, when $t \to \infty$ and for $\delta > 0$ small regarding ϵ, on $E_t^\delta \cap G_{t:t+h}$, we have:

$$\left|\hat{\mu}_2(t+h) - \hat{\mu}_2(t) - \frac{\sum_{i<h} R_{t+i} - \mu_2}{N_2(t)}\right| \leq \frac{h\epsilon}{N_2(t)} \quad \text{a.s.}$$

(**STEP 3**) We now bound the variation of the optimistic bonus of arm 2,

$$\sqrt{\frac{2\log(t+h)}{N_2(t+h)}}-\sqrt{\frac{2\log(t)}{N_2(t)}}=\sqrt{\frac{2\log(t+h)}{N_2(t+h)}}-\sqrt{\frac{2\log(t)}{N_2(t+h)}}+\sqrt{\frac{2\log(t)}{N_2(t+h)}}-\sqrt{\frac{2\log(t)}{N_2(t)}}$$
$$\sim o\Big(\frac{h}{tN_2(t)}\Big)+\frac{h}{2N_2(t)}\sqrt{\frac{2\log(t)}{N_2(t)}}.$$

Provided that $\delta>0$ is small enough, we have on E_t^δ:

$$\left|\sqrt{\frac{2\log(t+h)}{N_2(t+h)}}-\sqrt{\frac{2\log(t)}{N_2(t)}}-\frac{h(\mu_1-\mu_2)}{2N_2(t)}\right|\leq\frac{h\epsilon}{N_2(t)}$$

(**STEP 4**) All together, on $E_t^\delta\cap G_{t:t+h}$, we have:

$$I_1(t+h)-I_2(t+h)=I_1(t)-I_2(t)-\frac{\sum_{i<h}(R_{t+i}-\mu_2-\frac{\mu_1-\mu_2}{2})}{N_2(t)}\pm\frac{2h\epsilon}{N_2(t)}.$$

This proves the claim. □

We prove Lemma IV.28 as a corollary below.

Lemma IV.28. *Consider running* UCB*, and fix $h>0$. There exists a sequence of events indexed by exploration episodes (E_{τ_k}) with $\mathbf{P}(\liminf_k E_{\tau_k})=1$, such that, for all sequence $(U_t:t\geq1)$ of $\sigma(H_t)$-measurable events:*

$$\mathbf{P}\big(\forall i<h:A_{\tau_k+i}=2\mid E_{\tau_k},U_{\tau_k}\big)\geq\mu_2^h.$$

Proof. Fix $h>0$ and let $\delta(h),T(h)>0$ as given by Lemma 14.C.1 for some arbitrary $\epsilon>0$. Let $G'_{t:t+h}:=(\forall i<h,R_{t+i}=1)$, stating that every arm pull over $[t,t+h)$ provides full reward. On $E_t^{\delta(h)}\cap G_{t:t+h}\cap G'_{t:t+h}$ with $t\geq T(h)$, we have:

$$I_1(t+h)\leq I_2(t+h)+(I_1(t)-I_2(t))+\frac{\sum_{i<h}\big(\mu_2+\frac{\mu_1-\mu_2}{2}+\epsilon-1\big)}{N_2(t)}.$$

For $t\equiv\tau_k$ an exploration episode with τ_k, as $I_1(\tau_k)\leq I_2(\tau_k)$ (by definition), we obtain:

$$I_1(\tau_k+h)\leq I_2(\tau_k+h)+\frac{\sum_{i<h}\big(\mu_2+\frac{\mu_1-\mu_2}{2}+\epsilon-1\big)}{N_2(\tau_k)}.$$

We see that taking $\epsilon<\frac{\mu_1-\mu_2}{2}$, the summand is always negative, and as a consequence, $I_1(\tau_k+h)\leq I_2(\tau_k+h)$. So on $\mathbf{1}(\tau_k>T(h))\cap E_{\tau_k}^{\delta(h)}$, if every pull of the suboptimal arm $a=2$ over $[\tau_k,\tau_k+h)$ provides a full reward $R_t=1$, then $A_{\tau_k+h}=1$. More formally:

$$\mathbf{1}(\tau_k\geq T(h))\cap E_{\tau_k}^{\delta(h)}\cap G_{\tau_k:\tau_k+h}\cap G'_{\tau_k:\tau_k+h}=\mathbf{1}(\tau_k\geq T(h))\cap E_{\tau_k}^{\delta(h)}\cap G_{\tau_k:\tau_k+h+1}\cap G'_{\tau_k:\tau_k+h}$$

Now choose $\delta:=\min_{i\leq h}\delta(i)$ and $T:=\max_{i\leq h}T(i)$. The event $E_{\tau_k}:=E_{\tau_k}^\delta\cap\mathbf{1}(\tau_k\geq T)$ is $\sigma(H_{\tau_k})$-measurable, and we see that:

$$(-):=\mathbf{P}(\forall i<h:A_{\tau_k+i}=2\mid E_{\tau_k},U_{\tau_k})$$

$$\begin{aligned}
&\geq \mathbf{P}(\forall i < h : A_{\tau_k+i} = 2, R_{\tau_k+i} = 1 \mid E_{\tau_k}, U_{\tau_k}) \\
&= \prod_{i<h} \mathbf{P}\big(R_{\tau_k+i} = 1 \mid A_{\tau_k+i} = 2\big)\mathbf{P}\Big(A_{\tau_k+i} = 2 \mid G_{\tau_k:\tau_k+i}, G'_{\tau_k:\tau_k+i}, E_{\tau_k}, U_{\tau_k}\Big) \\
&= \prod_{i<h} \mathbf{P}\big(R_{\tau_k+i} = 1 \mid A_{\tau_k+i} = 2\big) \\
&= \mu_2^h.
\end{aligned}$$

Moreover, because $\tau_k < \tau_{k+1}$, the event $(\tau_k > T)$ is eventually true as $k \to \infty$, meaning that $\mathbf{P}(\liminf_k E_{\tau_k}) = 1$. This establishes the claim. □

14.C.3 Waiting for UCB to fail: Proof of Proposition IV.30

We recall the statement below.

Proposition IV.30. *Fix $h > 0$ and assume that we are running* UCB. *There exists an increasing sequence of almost-surely finite stopping times $(\sigma_k : k \geq 1)$ s.t.,*

$$\mathbf{P}(\mathrm{Reg}(\sigma_k; \sigma_k + h) \geq (\mu_1 - \mu_2)h) = 1.$$

Proof. Fix $h \geq 0$. Let $\ell > \lceil \frac{\mu_1 h}{1-\mu_1} \rceil$ and $\epsilon < \frac{\mu_1 - \mu_2}{2}$. Consider τ_k an exploration episode. Assume that $F_{\tau_k:\tau_k+\ell} := (\forall i < \ell : A_{\tau_k+i} = 2, R_{\tau_k+i} = 1)$ holds, which is of probability at least μ_2^ℓ on the event E_{τ_k} given by Lemma IV.28. From Lemma 14.C.1, almost surely, we have:

$$I_1(\tau_k + \ell + i) \leq I_2(\tau_k + \ell + i) - \frac{\ell(1-\mu_1)}{N_2(t)} + \frac{h\mu_1}{N_2(t)} < I_2(\tau_k + \ell + i).$$

Thus $E_{\tau_k} \cap F_{\tau_k:\tau_k+\ell} \subseteq (\forall i < h : A_{\tau_k+\ell+i} = 2)$ almost surely, and in particular $\mathbf{P}(\mathrm{Reg}(\tau_k + \ell; \tau_k + \ell + h) \geq (\mu_1 - \mu_2)h \mid E_{\tau_k}, F_{\tau_k:\tau_k+\ell}) = 1$. Since $\mathbf{P}(F_{\tau_k:\tau_k+\ell} \mid E_{\tau_k}) \geq \mu_2^\ell$ by Lemma IV.28, we deduce by Borel-Cantelli's Lemma that $\mathbf{P}(\limsup_k (E_{\tau_k} \cap F_{\tau_k;\tau_k+\ell})) = 1$. Hence, define

$$\sigma_1 := \inf\big\{\tau_k + \ell : E_{\tau_k} \cap F_{\tau_k:\tau_k+\ell}\big\}, \quad \sigma_{n+1} := \inf\big\{\tau_k + \ell > \sigma_n : E_{\tau_k} \cap F_{\tau_k:\tau_k+\ell}\big\}.$$

We see that σ_n is a stopping time. Moreover, we have $\mathbf{P}(\sigma_n < \infty)$ and $\mathbf{P}(\mathrm{Reg}(\sigma_n; \sigma_n + h) \geq (\mu_1 - \mu_2)h) = 1$ by construction. □

14.C.4 The regret of exploration of UCB: Proof of Theorem IV.33

This section is dedicated to a proof of:

Theorem IV.33. *Let $(X_t : t \geq 1)$ a sequence of i.i.d. random variables with distribution* B(μ_2). *Let σ_T the stopping time $T \wedge \inf\big\{t \geq 1 : -\frac{\mu_1-\mu_2}{2} + \frac{1}{t}\sum_{i=1}^t (X_t - \mu_2) \leq 0\big\}$. For all $T \geq 1$, we have* $\mathrm{RegExp}(\mathrm{UCB}; T) = \lim_{k\to\infty} \mathbf{E}[\mathrm{Reg}(\tau_k; \tau_k + T)] \geq (\mu_1 - \mu_2)\mathbf{E}[\sigma_T]$.

Again, as given by Proposition IV.27, the asymptotic regime is denoted

$$E_t^\epsilon := (\forall a : |\hat{\mu}_a(t) - \mu_a| < \epsilon) \cap \Big(N_2(t) = \tfrac{2\cdot(1\pm\epsilon)}{(\mu_1-\mu_2)^2}\log(t)\Big)$$

Denote $I_a(t) := \hat{\mu}_a(t) + \sqrt{2\log(t)/N_a(t)}$ UCB's index of arm a, similarly to previous sections. We begin by establishing a variant of Lemma 14.C.1 for time-periods when only the optimal arm is being pulled.

Lemma 14.C.2. *Let $F_{t:t+h} := (\forall i < h, A_{t+i} = 1)$. Fix arbitrary $\epsilon > 0$ and $h \geq 1$. There exists $\delta, T > 0$ such that, whenever $t > T$ and for all $\ell < h$, on $E_t^\delta \cap F_{t:t+\ell}$ we have:*

$$\left|(I_1(t+\ell) - I_2(t+\ell)) - \left(I_1(t) - I_2(t) + \frac{\sum_{i<\ell}(R_{t+i} - \mu_1)}{t}\right)\right| \leq \frac{2\ell\epsilon}{t}.$$

Proof. The proof is essentially similar to the one of Lemma 14.C.1: Approximate $I_a(t+\ell) - I_a(t)$ using equivalents in the asymptotic regime. Using that $N_1(t) \sim t$ and $N_2(t) \sim \frac{2}{(\mu_1-\mu_2)^2}\log(t)$, we find that the dominant term in the variations of $I_1(t+\ell) - I_2(t+\ell)$ with respect to ℓ is the one coming from the variations of the best arm's empirical estimate.

$$I_1(t+\ell) - I_1(t) = \frac{\sum_{i<\ell}(R_{t+i} - \mu_1)}{t} + o\left(\frac{\ell}{t}\right)$$
$$I_2(t+\ell) - I_2(t) = \frac{\ell(\mu_1 - \mu_2)}{2t\log(t)} + o\left(\frac{\ell}{t\log(t)}\right).$$

Quantifying the equivalents with $\epsilon > 0$, we obtain the statement of Lemma 14.C.2. □

Proof of Theorem IV.33. Recall that $(X_t : t \geq 1)$ denotes a sequence of i.i.d. random variables with distribution $B(\mu_2)$. Fix $h \geq 1$ and denote the *exploitation episodes* (τ'_k) as:

$$\tau'_k := \inf\{t > \tau_k : A_t = 1, A_{t-1} = 2\}.$$

It is obvious that $\mathbf{E}[\mathrm{Reg}(\tau_k; \tau_k + h)] \leq (\mu_1 - \mu_2)\mathbf{E}[\min(\tau'_k - \tau_k, h)]$, hence we are ought to bound $\mathbf{E}[\min(\tau'_k - \tau_k, h)]$ which is related to the expected duration of the k-th exploration episode clipped to $[0, h]$. From Lemma 14.C.2 follows that at the beginning of an exploration episode τ_k and on $E^\epsilon_{\tau_k}$ for ϵ (resp. τ_k) small enough (resp. large enough), we have:

$$0 \leq I_2(\tau_k) - I_1(\tau_k) \leq \frac{2}{\tau_k}.$$

Furthermore, if $A_{\tau_k+\ell} = 1$, then $I_2(\tau_k + \ell) - I_1(\tau_k + \ell) \leq 0$, so combined with Lemma 14.C.1 and denoting $\nu_0 := \frac{\mu_1+\mu_2}{2}$, it implies that

$$\sum_{i<\ell}(R_{\tau_k+i} - \nu_0) \leq 2\ell\epsilon + \frac{2N_2(\tau_k)}{\tau_k}$$

where $R_{\tau_k+i} \sim B(\mu_2)$. Provided that τ_k is large enough (i.e., that k is large enough), this in particular implies that $\sum_{i<\ell}(R_{\tau_k+i} - \nu_0) \leq 3h\epsilon$. Since ϵ can be chosen arbitrarily close to 0, we deduce that, for all $\epsilon > 0$,

$$\liminf_{k\to\infty} \mathbf{E}[\mathrm{Reg}(\tau_k; \tau_k + h)] \geq (\mu_1 - \mu_2)\mathbf{E}\left[\inf\left\{t \leq h : \sum_{i<t}(X_t - \nu_0) \leq \epsilon\right\}\right]$$

Since $\sum_{i<t}(X_t - \nu_0)$ takes finitely many values when $t \leq h$, we have:

$$\inf_{\epsilon>0} \mathbf{E}\left[\inf\left\{t \leq h : \sum_{i<t}(X_t - \nu_0) \leq \epsilon\right\}\right] = \mathbf{E}\left[\inf\left\{t \leq h : \sum_{i<t}(X_t - \nu_0) \leq 0\right\}\right]$$

This proves the result. □

14.D General index theory

We write $X_a(t)$ any data relative to the arm a, and $X_{-a}(t)$ any data relative to the other arm. The index of arm a is thus denoted I_a, while I_{-a} denotes the one of the other arm.

14.D.1 Proof of Lemma IV.34

Lemma IV.34. *Assume that $I(-)$ satisfies (**A 1-3**). Then, for $a = 1, 2$, $\hat{\mu}_a(t) \to \mu_a$ a.s.*

Proof of Lemma IV.34. (**STEP 1**) We start by showing that both arms are visited infinitely often, that is, for $a = 1, 2$ and for a fixed arbitrary n, $\mathbf{P}(\exists t : N_a(t) \geq n) = 1$. By the Strong Law of Large Number (SLLN, or just time-uniform concentration inequalities), the result will follow. Consider the complementary event. Remark that if $N_a(t) < n$, there must be $s \in \{t-n, \dots, t\}$ such that $A_s \neq a$. So, we have, for $\delta > 0$ small enough,

$$
\begin{aligned}
(-) &:= \mathbf{P}(\forall t : N_a(t) < n) \\
&\leq \mathbf{P}(\forall t, \exists s \geq t-n : N_a(s) < n, A_s \neq a) \\
&= \lim_{t\to\infty} \mathbf{P}(\exists s \geq t-n : N_a(s) < n, I_a(\hat{\mu}(s), N_a(s), s) \leq I_{-a}(\hat{\mu}(s), N_{-a}(s), s)) \\
&= \lim_{t\to\infty} \mathbf{P}(\exists s \geq t-n : I_a(\hat{\mu}(s), n, s) \leq I_{-a}(\hat{\mu}(s), s-n, s), N_{-a}(s) \geq s-n) && \textbf{(A 1)} \\
&\leq \lim_{t\to\infty} \mathbf{P}(\exists s \geq t-n : I_a(\hat{\mu}(s), n, s) \leq I_{-a}(\hat{\mu}(s), s-n, s), |\hat{\mu}_{-a}(s) - \mu_{-a}| \leq \delta) && \text{(SLLN)} \\
&\leq \lim_{t\to\infty} \mathbf{P}(\exists s \geq t-n : I_a(0, \mu_{-a} - \delta, n, s) \leq I_{-a}(\mu_{-a} + \delta, 1, s-n, s)) && \textbf{(A 1)} \\
&\leq \lim_{t\to\infty} \mathbf{P}\Big(\exists s \geq t-n : I_a(0, \mu_{-a} - \delta, n, s) \leq I(\mu_1, 1) + \operatorname*{o}_{\delta\to 0}(1)\Big) && \textbf{(A 3)} \\
&= 0. && \textbf{(A 2)}
\end{aligned}
$$

(**STEP 2**) Since all arms are pulled infinitely often, the empirical estimates must converge to the mean values by the Strong Law of Large Numbers. □

14.D.2 Proof of Lemma IV.35

Lemma IV.35. *If $I(-)$ satisfies (**A 1-6**), then $(\hat{\mu}_1(t), \hat{\mu}_2(t), N_1(t), N_2(t)) \sim (\mu_1, \mu_2, t, n_2(t))$ a.s. The sequence $t \mapsto (\mu_1, \mu_2, t, n_2(t))$ will be called the* asymptotic regime.

Proof of Lemma IV.35. Since $n_2(t)$ is sublinear, we only have to show the property on $N_2(t)$ and everything will follow.

(**STEP 1**) Let $\epsilon > 0$ and focus on $a = 2$ the suboptimal arm. For conciseness, denote $c := 1 - \epsilon$. Similarly to the previous point, remark that if $N_2(s) < cn_2(s)$, there must be some $u \in \{s - cn_2(s), \dots, s\}$ when $A_u \neq 2$. Let $F_t^\delta := (\forall a, |\hat{\mu}_a(t) - \mu_a| < \delta)$ the concentration event, proved to hold eventually, as given by Lemma IV.34. We then have:

$$
\begin{aligned}
(-) &:= \mathbf{P}(\forall t, \exists s \geq t : N_2(s) < cn_2(s)) \\
&\leq \lim_{t\to\infty} \mathbf{P}(\exists s \geq t, \exists u \geq s - cn_2(s) : N_2(s) \leq cn_2(s), I_2(u) \leq I_1(u)) \\
&\leq \lim_{t\to\infty} \mathbf{P}\big(\exists s \geq t, \exists u \geq s - cn_2(s) : N_2(s) \leq cn_2(s), I_2(u) \leq I_1(u), F_s^\delta\big) && \text{(Lem. IV.34)} \\
&\leq \lim_{t\to\infty} \mathbf{P}\left(\exists s \geq t, \exists u \geq s - cn_2(s) : \begin{array}{c} I(\mu_2 - \delta, \mu_1 + \delta, cn_2(s), u) \\ \leq \\ I(\mu_1 + \delta, \mu_2 - \delta, u - cn_2(s), u) \end{array}\right) && \textbf{(A 1)} \\
&\leq \lim_{t\to\infty} \mathbf{P}\left(\exists s \geq t, \exists u \geq s - cn_2(s) : \begin{array}{c} I(\mu_2 - \delta, \mu_1 + \delta, cn_2(s), u) \\ \leq \Big(1 + \operatorname*{o}_{\delta\to 0}(1) + \operatorname*{o}_{t\to\infty}(1)\Big) I(\mu_1, \mu_2) \end{array}\right) && \textbf{(A 3,6)} \\
&\leq \lim_{t\to\infty} \mathbf{P}\left(\exists s \geq t, \exists u \geq s - cn_2(s) : \begin{array}{c} I(\mu_2 - \delta, \mu_1 + \delta, (1-\epsilon)n_2(s), u) \\ \leq (1 + o(1)) I(\mu_2 - \delta, \mu_1 + \delta, n_2(s), s) \end{array}\right) && \textbf{(A 4)}
\end{aligned}
$$

$$\leq \lim_{t\to\infty} \mathbf{P}\Big(\exists s \geq t, \exists u \geq s - cn_2(s) : \begin{array}{c}(1+\ell(\epsilon))I(\mu_2-\delta,\mu_1+\delta,n_2(s),s)\\ \leq (1+\mathrm{o}(1))I(\mu_2-\delta,\mu_1+\delta,n_2(s),s)\end{array}\Big) \qquad \textbf{(A 5)}$$

$$\leq \lim_{t\to\infty} \mathbf{P}\Big(\exists s \geq t, \exists u \geq s - cn_2(s) : 1+\ell(\epsilon) \leq 1 + \operatorname*{o}_{\delta\to 0}(1) + \operatorname*{o}_{t\to\infty}(1)\Big)$$

$$= 0. \qquad (\delta \to 0)$$

(**STEP 2**) Let $\epsilon > 0$ and focus on $a = 2$ the suboptimal arm. This time, denote $c := 1+\epsilon$. The analysis is mostly similar, but the initial decomposition starts differently. Remark that if $N_2(s) > cn_2(s)$, then there must be $u \in \{cn_2(s), \dots, s\}$ such that $A_u = 2$. So,

$$\begin{aligned}
(-) &:= \mathbf{P}(\forall t, \exists s \geq t : N_2(s) \geq cn_2(s)) \\
&\leq \lim_{t\to\infty} \mathbf{P}(\exists s \geq t, \exists u \in [cn_2(s), s] : N_2(u) \geq cn_2(s), A_u = 2) \\
&\leq \lim_{t\to\infty} \mathbf{P}(\exists s \geq t, \exists u \in [cn_2(s), s] : N_2(u) \geq cn_2(s), I_2(u) \geq I_1(u)) \\
&\leq \lim_{t\to\infty} \mathbf{P}(\exists s \geq t, \exists u \in [cn_2(s), s] : I_2(\hat{\mu}(u), cn_2(u), u) \geq I_1(\hat{\mu}(u), u, u)) && \textbf{(A 1)} \\
&\leq \lim_{t\to\infty} \mathbf{P}\Big(\exists s \geq t, \exists u \in [cn_2(s), s] : I_2(\hat{\mu}(u), cn_2(u), u) \geq I_1(\hat{\mu}(u), u, u), F_u^\delta\Big) \\
&\leq \lim_{t\to\infty} \mathbf{P}\Big(\exists s \geq t, \exists u \in [cn_2(s), s] : I_2(\hat{\mu}(u), cn_2(u), u) \geq (1+\mathrm{o}(1))I(\mu_1, \mu_2), F_u^\delta\Big) && \textbf{(A 3)} \\
&\leq \lim_{t\to\infty} \mathbf{P}\left(\exists s \geq t, \exists u \in [cn_2(s), s] : F_u^\delta, \begin{array}{c} I_2(\hat{\mu}(u), (1+\epsilon)n_2(u), u) \\ \geq \\ (1+\mathrm{o}(1))I_2(\hat{\mu}(u), n_2(u), u)\end{array}\right) && \textbf{(A 4)} \\
&\leq \lim_{t\to\infty} \mathbf{P}\Big(\exists s \geq t, \exists u \in [cn_2(s), s] : 1-\ell(\epsilon) \geq 1 + \operatorname*{o}_{\delta\to 0}(1) + \operatorname*{o}_{t\to\infty}(1)\Big) && \textbf{(A 5)} \\
&= 0. && (\delta \to 0)
\end{aligned}$$

So, $t \mapsto N_2(t)$ converges to $t \mapsto n_2(t)$ almost-surely for the asymptotic topology. □

Conclusion, Past and Future Works

Every document eventually ends. Yet so much could still be said. At the end of every part, I have already listed a few research directions that I believe to be promising. In the lines below, we take a few steps back from this heavy manuscript and discuss research directions more informally, and extend on a few shortcomings that the current literature suffers from.

We deepdived into the realm of Markov decision process under the average gain criterion, aiming at minimizing the regret and have presented many new results. In the minimax setting, the regret is scaling with the span of the bias function rather than the diameter, and the minimax lower bound can be reached by an algorithm running in polynomial time with no prior information. In the model dependent setting, the regret lower bound is a solution of an optimization problem combining information and navigation constraints. We have explained that minors play an important role, and that the navigational structure of Markov decision processes makes important to distinguish between exploration, co-exploration and exploitation to learn efficiently. The regret lower bound is shown tight, by providing an algorithm that approaches it with arbitrary precision. Beyond regret minimization which is only a matter of global performance, we pinpoint in the last part that the management of episodes is important as well, provided that one is concerned about the local behavior of algorithms. To that end, we design two new learning metrics (the regret of exploration and the sliding regret) to explain and quantify how efficient algorithms behave locally.

So, what's next?

Too much, actually.

To begin with, a fundamental theory is absent from the landscape described above: Bayesian approaches to reinforcement learning. In such approaches, the underlying model is drawn from a probability distribution instead of being fixed. The lowerbound-then-algorithm strategy could also be applied to this setting. There are already a few results that follow this principle, such as Lai (1987) for stochastic bandits. For Markov decision processes, I am very curious about what can be obtained. `PMEVI` and `ECoE`, that are respectively asymptotically optimal in the minimax and model dependent settings, are very different algorithms. They can hardly be compared actually. As a matter of fact, the model dependent and model independent regret lower bounds are extremely different in their structure. Would we find something different in the Bayesian setting again? Can these bounds be reconciliated? Or, is there a crucial difference of design between model dependent, minimax and Bayesian approaches? For stochastic bandits, model dependent and independent optimalities *are* reconciliatable — This is known as the best-of-both-worlds Bubeck and Slivkins (2012). There is also a parallel best-of-both-worlds for the model dependent & Bayesian approaches, see Lai (1987), and it would not be surprising if, for stochastic bandits at least, a best-of-**three**-worlds exists. For Markov decision processes, this is much less clear, for reasons that I discuss more deeply in the conclusion of Part III. If for communicating Markov decision processes, there is no best-of-three-worlds, then where is the barrier?

At the back of the room, there is always someone that bargains about the asymptotic nature of the analysis. Yes, the analysis provided in the model dependent and model independent

settings are very asymptotic. It is often opposed to asymptotic analysis that it is important to keep track of second order errors, and to report them. I would usually agree to this provided that the process comes for free, i.e., that it does not deteriorate the quality and clarity of the main focus of the proof. And yet, I would disagree with the idea. First, the size of second order terms may be, or may be not, artifacts of the analysis. Second, these second order terms mix asymptotic second order terms and transient terms. Hence, while they may tell that the analysis is too asymptotic, they tell very little about how efficient learners should behave during the transient stages of learning. Furthermore, in the conclusion of Part III, I claim that the model dependent lower bound may itself be very asymptotic in nature. Behind this claim is the idea that the information constraints of the lower bound are linked to Sanov's Theorem, and that the second order error in Sanov's Theorem is known to be linked to the geometry of the set from which we want to control the probability to fall into. This set is the set of confusing models. For stochastic bandits, this set is a half-space and its geometry is trivial, leading to small second order errors. For Markov decision processes, it is non-convex and its geometry may be awful. If the regret lower bound of Markov decision processes is too asymptotic in nature, then controlling the second order term is irrelevant. This encourages the development of learning theories that are specifically designed to understand how a learner should behave during the earlier stages of learning. Such theories would probably be closer to Bayesian approaches, and should not be mistaken for minimax approaches.

The last part of this document, which is about local regret guarantees, tries to escape the prison of standard regret minimization. Not everything is about regret minimization. To that extent, the various behaviors described in Part IV are arguably more important than the introduced learning metrics. The various questions related to how the trajectorial behaviors of classic algorithms may be studied, classified and quantified, seem very promising. The regret of exploration and the sliding regret are simple metrics that barely scratch the surface of what could be done. In the conclusion of Chapter 14, we explain that these two metrics fail to capture the very specific behavior of a few algorithms such as EXP3. The proper way to address that issue is still completely open.

And all this is only if we stick to the neighborhood of regret minimization concerns.

I defer the discussion of more *practical*, or *instantaneous*, research directions to the dedicated conclusion of every part of this manuscript, that are better places to discuss the direct follow-ups of well-confined collections of results.

List of Papers

This manuscript is a compilation a several works that I have completed during my PhD, yet not all of my works are covered. I provide a list of my main publications below. Works that are tagged as "**not covered**" are those of which the content does not appear in the present manuscript. Such works are put aside as they do not fit the story-line of the manuscript, which is exclusively about regret minimization in Markov decision processes.

Boone, V. and Gaujal, B. (2023a). Identification of Blackwell Optimal Policies for Deterministic MDPs. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 7392–7424. PMLR

Not covered.

Abstract. This work investigates the probably correct (PC) identification of Blackwell optimal policies in deterministic transition Markov decision processes, where the goal of the learner is to gather information and eventually stop to return a policy that has to be Blackwell optimal Blackwell (1962) with a fixed probability. We show that Blackwell optimal policies cannot be identified with arbitrarily high confidence unless the model is non-degenerate (Definition IV.3), in which case Blackwell optimality collapses to bias optimality (Definition I.9). We provide model-dependent lower and upper bounds on the number of samples required to identify such policies.

Boone, V. and Gaujal, B. (2023b). The Regret of Exploration and the Control of Bad Episodes in Reinforcement Learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2824–2856. PMLR

See Chapters 11 and 12.

Abstract. This paper is the first of the line of work on local regret considerations. It explains why the doubling trick is not satisfying, introduces the regret of exploration (Definition IV.2), the performance test (PT), provides minimax regret guarantees for UCRL-PT and regret of exploration guarantees for UCRL-PT when the underlying model has deterministic transitions.

Boone, V. (2023). The Sliding Regret in Stochastic Bandits: Discriminating Index and Randomized Policies. arXiv:2311.18437 [cs, eess, math, stat]

See Chapter 14.

Abstract. This paper is mostly similar to Chapter 14. We investigate the trajectorial behavior of standard algorithms for stochastic bandits. We introduce the sliding regret (Definition IV.7) that measures the worst local regret of the algorithm on a trajectory. We show that index algorithms such as UCB,

UCB-V, IMED and KL-UCB have linear sliding regret, resulting in infinitely many bursts of suboptimal play that may last arbitrarily long; While randomized algorithms such as TS and MED have optimal sliding regret, meaning that they play suboptimal actions sporadically.

This paper is under review at JMLR.

Boone, V. and Mertikopoulos, P. (2024). The equivalence of dynamic and strategic stability under regularized learning in games. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc

Not covered.

Abstract. This paper is not about reinforcement learning in Markov decision process hence is completely absent from the manuscript — This further requires to explain in more details what is done in this work. In this paper, we examine the long-run behavior of regularized, no-regret learning in finite games. A well-known result in the field states that the empirical frequencies of no-regret play converge to the game's set of coarse correlated equilibria; however, our understanding of how the players' actual strategies evolve over time is much more limited - and, in many cases, non-existent. This issue is exacerbated by a series of recent results showing that only strict Nash equilibria are stable and attracting under regularized learning, thus making the relation between learning and pointwise solution concepts particularly elusive. In lieu of this, we take a more general approach and instead seek to characterize the setwise rationality properties of the players' day-to-day play. To that end, we focus on one of the most stringent criteria of setwise strategic stability, namely that any unilateral deviation from the set in question incurs a cost to the deviator - a property known as closedness under better replies (club). In so doing, we obtain a far-reaching equivalence between strategic and dynamic stability: a product of pure strategies is closed under better replies if and only if its span is stable and attracting under regularized learning. In addition, we estimate the rate of convergence to such sets, and we show that methods based on entropic regularization (like the exponential weights algorithm) converge at a geometric rate, while projection-based methods converge within a finite number of iterations, even with bandit, payoff-based feedback.

Boone, V. and Zhang, Z. (2024). Achieving Tractable Minimax Optimal Regret in Average Reward MDPs. _eprint: 2406.01234

See Chapter 7.

Abstract. In this paper, we present the first tractable algorithm with minimax optimal regret of $\mathrm{O}(\sqrt{\mathrm{sp}(h)SAT\log(T)})$ without relying on prior knowledge. The algorithm relies on a novel subroutine, PMEVI (see Algorithm II.5) to compute bias-constrained optimal policies efficiently, improving on EVI from Auer et al. (2009). This subroutine can be applied to various previous algorithms to improve their regret bounds. The content of this paper is mostly similar to Chapter 7.

This paper has been accepted at NeurIPS 2024.

Boone, V. and Gaujal, B. (2024+). Local regret guarantees in average reward markov decision processes. To be submitted

See Chapter 13.

Abstract. This paper goes beyond Boone and Gaujal (2023b). The theory on the regret of exploration is more complete, with a concern about the

generality of the definition and the well-definition of the regret of exploration for general episodic algorithms. It suggests to replace the performance test by the vanishing multiplicative condition (VM) and provides a general analysis based on coherence (Definition IV.5) with a strong emphasis on the shrinking-shaking effect. The content of this paper is mostly similar to Chapter 13.

Boone, V. and Maillard, O.-A. (2024+). Lower bound of the regret for communicating markov decision processes. To be submitted

See Part III.

Abstract. The content of this paper is mostly similar to Part III. We provide a model dependent regret lower bound for Markov decision processes in the communicating setting together with a few approximations of it. The lower bound is shown to be optimal with the introduction of a new algorithm scheme, ECoE, of which the asymptotic expected regret is arbitrarily close to the lower bound. We additionally show that the computation of the lower bound, together with associated sub-problems, is computationally hard.

Bibliography

Abbasi-Yadkori, Y., Lazic, N., Szepesvari, C., and Weisz, G. (2019). Exploration-enhanced politex. *arXiv preprint arXiv:1908.10479*.

Agrawal, R. (1990). Adaptive control of Markov chains under the weak accessibility. *29th IEEE Conference on Decision and Control*, pages 1426–1431 vol.3.

Agrawal, R. (1991). Minimizing the learning loss in adaptive control of Markov chains under the weak accessibility condition. *Journal of Applied Probability*, 28:779 – 790.

Agrawal, R., Teneketzis, D., and Anantharam, V. (1988). Asymptotically efficient adaptive allocation schemes for controlled Markov chains: finite parameter space. In *Proceedings of the 27th IEEE Conference on Decision and Control*, pages 1198–1203 vol.2.

Agrawal, S. and Goyal, N. (2012). Analysis of Thompson Sampling for the multi-armed bandit problem. arXiv:1111.1797 [cs].

Agrawal, S. and Jia, R. (2023). Optimistic Posterior Sampling for Reinforcement Learning: Worst-Case Regret Bounds. *Mathematics of Operations Research*, 48(1):363–392. Publisher: INFORMS.

Arapostathis, A., Borkar, V. S., Fernández-Gaucherand, E., Ghosh, M. K., and Marcus, S. I. (1993). Discrete-time controlled Markov processes with average cost criterion: a survey. *Siam Journal on Control and Optimization*, 31:282–344.

Audibert, J.-Y. and Bubeck, S. (2009). Minimax policies for adversarial and stochastic bandits. In *COLT*, page 10.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902. Publisher: Elsevier.

Auer, P. (2002). Using Confidence Bounds for Exploitation-Exploration Trade-offs. *J. Mach. Learn. Res.*, 3:397–422.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331.

Auer, P., Jaksch, T., and Ortner, R. (2009). Near-optimal Regret Bounds for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Auer, P. and Ortner, R. (2006). Logarithmic Online Regret Bounds for Undiscounted Reinforcement Learning. *Proceedings of the 19th International Conference on Neural Information Processing Systems*.

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax Regret Bounds for Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 263–272. PMLR. ISSN: 2640-3498.

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19(3):357 – 367. Publisher: Tohoku University, Mathematical Institute.

Bartlett, P. L. and Tewari, A. (2009). REGAL: a regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 35–42, Arlington, Virginia, USA. AUAI Press.

Baudry, D., Kaufmann, E., and Maillard, O.-A. (2020). Sub-sampling for Efficient Non-Parametric Bandit Exploration. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5468–5478. Curran Associates, Inc.

Baudry, D., Suzuki, K., and Honda, J. (2023). A General Recipe for the Analysis of Randomized Multi-Armed Bandit Algorithms. *arXiv preprint arXiv:2303.06058*.

Bellman, R. (1957). A Markovian decision process. *Journal of mathematics and mechanics*, pages 679–684. Publisher: JSTOR.

Belomestny, D., Menard, P., Naumov, A., Tiapkin, D., and Valko, M. (2023). Sharp Deviations Bounds for Dirichlet Weighted Sums with Application to analysis of Bayesian algorithms. arXiv:2304.03056 [math, stat].

Berge, C. (1957). Topological spaces: Including a treatment of multi-valued functions, vector spaces and convexity.

Bernstein, S. (1924). On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49.

Bertsekas, D. (2012). *Dynamic programming and optimal control: Volume I*, volume 1. Athena scientific.

Bertsekas, D. P. and others (2011). Dynamic programming and optimal control 3rd edition, volume ii. *Belmont, MA: Athena Scientific*.

Blackwell, D. (1962). Discrete dynamic programming. *The Annals of Mathematical Statistics*, pages 719–726. Publisher: JSTOR.

Boone, V. (2023). The Sliding Regret in Stochastic Bandits: Discriminating Index and Randomized Policies. arXiv:2311.18437 [cs, eess, math, stat].

Boone, V. and Gaujal, B. (2023a). Identification of Blackwell Optimal Policies for Deterministic MDPs. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 7392–7424. PMLR.

Boone, V. and Gaujal, B. (2023b). The Regret of Exploration and the Control of Bad Episodes in Reinforcement Learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2824–2856. PMLR.

Boone, V. and Gaujal, B. (2024+). Local regret guarantees in average reward markov decision processes. To be submitted.

Boone, V. and Maillard, O.-A. (2024+). Lower bound of the regret for communicating markov decision processes. To be submitted.

Boone, V. and Mertikopoulos, P. (2024). The equivalence of dynamic and strategic stability under regularized learning in games. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Boone, V. and Zhang, Z. (2024). Achieving Tractable Minimax Optimal Regret in Average Reward MDPs. _eprint: 2406.01234.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities - A Nonasymptotic theory of independence*.

Bourel, H., Maillard, O., and Talebi, M. S. (2020). Tightening Exploration in Upper Confidence Reinforcement Learning. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1056–1066. PMLR.

Bubeck, S. and Slivkins, A. (2012). The best of both worlds: stochastic and adversarial bandits. arXiv:1202.4473 [cs].

Burnetas, A. and Katehakis, M. (1997). Optimal Adaptive Policies for Markov Decision Processes. *Mathematics of Operations Research - MOR*, 22:222–255.

Christ, M. and Yannakakis, M. (2023). The Smoothed Complexity of Policy Iteration for Markov Decision Processes. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1890–1903.

Cohen, M. B., Lee, Y. T., and Song, Z. (2020). Solving linear programs in the current matrix multiplication time.

d'Epenoux, F. (1960). Sur un probleme de production et de stockage dans l'aléatoire. *Revue Française de Recherche Opérationelle*, 14(3-16):4.

Efroni, Y., Merlis, N., Ghavamzadeh, M., and Mannor, S. (2019). Tight Regret Bounds for Model-Based Reinforcement Learning with Greedy Policies. *arXiv:1905.11527 [cs, stat]*. arXiv: 1905.11527.

Filippi, S., Cappé, O., and Garivier, A. (2010). Optimism in Reinforcement Learning and Kullback-Leibler Divergence. *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122. arXiv: 1004.5229.

Fox, B. L. and Rolph, J. E. (1973). Adaptive policies for Markov renewal programs. *The Annals of Statistics*, 1(2):334–341. Publisher: Institute of Mathematical Statistics.

Freedman, D. A. (1975). On tail probabilities for martingales. *the Annals of Probability*, pages 100–118. Publisher: JSTOR.

Fruit, R. (2019). *Exploration-exploitation dilemma in Reinforcement Learning under various form of prior knowledge*. PhD Thesis, Université de Lille 1, Sciences et Technologies; CRIStAL UMR 9189.

Fruit, R., Pirotta, M., and Lazaric, A. (2020). Improved Analysis of UCRL2 with Empirical Bernstein Inequality. *ArXiv*, abs/2007.05456.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. (2018). Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning. *Proceedings of the 35 th International Conference on Machine Learning*.

Garivier, A. and Cappé, O. (2011). The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings.

Garivier, A., Hadiji, H., Ménard, P., and Stoltz, G. (2022). KL-UCB-Switch: Optimal Regret Bounds for Stochastic Bandits from Both a Distribution-Dependent and a Distribution-Free Viewpoints. *Journal of Machine Learning Research*, 23(179):1–66.

Garivier, A., Ménard, P., and Stoltz, G. (2018). Explore First, Exploit Next: The True Shape of Regret in Bandit Problems. _eprint: 1602.07182.

Gast, N., Gaujal, B., and Khun, K. (2023). Testing indexability and computing Whittle and Gittins index in subcubic time. *Mathematical Methods of Operations Research*, 97(3):391–436. Publisher: Springer.

Goyal, V. and Grand-Clement, J. (2023). A first-order approach to accelerated value iteration. *Operations Research*, 71(2):517–535. Publisher: INFORMS.

Honda, J. and Takemura, A. (2010). An Asymptotically Optimal Policy for Finite Support Models in the Multiarmed Bandit Problem.

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

Howard, R. A. (1960). Dynamic programming and markov processes. Publisher: John Wiley.

Jonsson, A., Kaufmann, E., Ménard, P., Darwiche Domingues, O., Leurent, E., and Valko, M. (2020). Planning in markov decision processes with gap-dependent sample complexity. *Advances in Neural Information Processing Systems*, 33:1253–1263.

Kakade, S., Wang, M., and Yang, L. F. (2020). Variance Reduction Methods for Sublinear Reinforcement Learning.

Kallenberg, L. (2016). *Markov Decision Processes: Lecture Notes*.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models. *Journal of Machine Learning Research*, 17(1):1–42.

Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson Sampling: An Asymptotically Optimal Finite Time Analysis. *arXiv:1205.4217 [cs, stat]*. arXiv: 1205.4217.

Lai, T. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22.

Lai, T. L. (1987). Adaptive Treatment Allocation and the Multi-Armed Bandit Problem. *The Annals of Statistics*, 15(3):1091 – 1114. Publisher: Institute of Mathematical Statistics.

Lattimore, T. (2018). Refining the confidence level for optimistic bandit strategies. *The Journal of Machine Learning Research*, 19(1):765–796. Publisher: JMLR. org.

Lattimore, T. and Hutter, M. (2012). PAC Bounds for Discounted MDPs. In Bshouty, N. H., Stoltz, G., Vayatis, N., and Zeugmann, T., editors, *Algorithmic Learning Theory*, Lecture Notes in Computer Science, pages 320–334, Berlin, Heidelberg. Springer.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Leizarowitz, A. (2002). On Optimal Policies of Multichain Finite State Compact Action Markov Decision Processes. In *Decision & Control in Management Science: Essays in Honor of Alain Haurie*, pages 79–95. Springer.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Generative Model. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12861–12872. Curran Associates, Inc.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2021). Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. 34:17762–17776.

Maillard, O.-A. (2019). *Mathematics of Statistical Sequential Decision Making*. Habilitation à diriger des recherches, Université de Lille Nord de France.

Maillard, O.-A., Mann, T. A., and Mannor, S. (2014). How hard is my MDP?" The distribution-norm to the rescue". In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Maillard, O.-A., Munos, R., and Stoltz, G. (2011). A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 497–514. JMLR Workshop and Conference Proceedings.

Mallows, C. and Robbins, H. (1964). Some problems of optimal sampling strategy. *Journal of Mathematical Analysis and Applications*, 8(1):90–103. Publisher: Academic Press.

Marjani, A. A., Garivier, A., and Proutiere, A. (2021). Navigating to the Best Policy in Markov Decision Processes. *arXiv:2106.02847 [cs, stat]*. arXiv: 2106.02847.

Marjani, A. A. and Proutiere, A. (2021). Adaptive Sampling for Best Policy Identification in Markov Decision Processes. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7459–7468. PMLR. ISSN: 2640-3498.

Maxwell, M. and Woodroofe, M. (2000). Central limit theorems for additive functionals of Markov chains. *Annals of probability*, pages 713–724. Publisher: JSTOR.

Munos, R. and Moore, A. (1999). Influence and variance of a Markov chain: Application to adaptive discretization in optimal control. In *Proceedings of the 38th IEEE Conference on Decision and Control (Cat. No. 99CH36304)*, volume 2, pages 1464–1469. IEEE.

Nilim, A. and El Ghaoui, L. (2005). Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798. Publisher: INFORMS.

Ok, J., Proutiere, A., and Tranos, D. (2018). Exploration in Structured Reinforcement Learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Ortner, R. (2010). Online regret bounds for Markov decision processes with deterministic transitions. *Theoretical Computer Science*, 411(29):2684–2695.

Ortner, R. (2013). Adaptive aggregation for reinforcement learning in average reward markov decision processes. *Annals of Operations Research*, 208:321–336.

Ortner, R. (2020). Regret Bounds for Reinforcement Learning via Markov Chain Concentration. *Journal of Artificial Intelligence Research*, 67:115–128.

Osband, I. and Roy, B. V. (2017). Why is Posterior Sampling Better than Optimism for Reinforcement Learning? In *Proceedings of the 34th International Conference on Machine Learning*, pages 2701–2710. PMLR. ISSN: 2640-3498.

Osband, I., Russo, D., and Van Roy, B. (2013). (More) Efficient Reinforcement Learning via Posterior Sampling. *arXiv:1306.0940 [cs, stat]*. arXiv: 1306.0940.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. (2017). Learning Unknown Markov Decision Processes: A Thompson Sampling Approach. *arXiv:1709.04570 [cs]*. arXiv: 1709.04570.

Pardalos, P. M. and Schnitger, G. (1988). Checking local optimality in constrained quadratic programming is NP-hard. *Operations Research Letters*, 7:33–35.

Peligrad, M. (2020). A new CLT for additive functionals of Markov chains. *Stochastic Processes and their Applications*, 130(9):5695–5708. Publisher: Elsevier.

Pesquerel, F. and Maillard, O.-A. (2022). IMED-RL: Regret optimal learning of ergodic Markov decision processes. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 26363–26374. Curran Associates, Inc.

Pirutinsky, D. (2020). *On Asymptotically Optimal Reinforcement Learning*. PhD Thesis, Rutgers The State University of New Jersey, Graduate School-Newark.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1 edition.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527 – 535. Publisher: American Mathematical Society.

Saber, H., Pesquerel, F., Maillard, O.-A., and Talebi, M. S. (2024). Logarithmic regret in communicating MDPs: Leveraging known dynamics with bandits. In Yanıkoğlu, B. and Buntine, W., editors, *Proceedings of the 15th Asian Conference on Machine Learning*, volume 222 of *Proceedings of Machine Learning Research*, pages 1167–1182. PMLR.

Schweitzer, P. J. (1985). On undiscounted Markovian decision processes with compact action spaces. *RAIRO-Operations Research*, 19(1):71–86. Publisher: EDP Sciences.

Schweitzer, P. J. (1987). A Brouwer fixed-point mapping approach to communicating Markov decision processes. *Journal of mathematical analysis and applications*, 123(1):117–130. Publisher: Elsevier.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Talebi, M. S. and Maillard, O.-A. (2018). Variance-Aware Regret Bounds for Undiscounted Reinforcement Learning in MDPs. *Journal of Machine Learning Research*, pages 1–36. Publisher: Microtome Publishing.

Tewari, A. and Bartlett, P. (2007). Optimistic linear programming gives logarithmic regret for irreducible MDPs. *Advances in Neural Information Processing Systems*, 20.

Thompson, W. R. (1933). On the Likelihood that One Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3-4):285–294.

Tossou, A., Basu, D., and Dimitrakakis, C. (2019). Near-optimal Optimistic Reinforcement Learning using Empirical Bernstein Inequalities. *arXiv:1905.12425 [cs, stat]*. arXiv: 1905.12425.

Tranos, D. and Proutiere, A. (2021). Regret Analysis in Deterministic Reinforcement Learning. *arXiv:2106.14338 [cs, stat]*. arXiv: 2106.14338.

Tuynman, A., Degenne, R., and Kaufmann, E. (2024). Finding good policies in average-reward Markov Decision Processes without prior knowledge.

Vogel, W. (1960). An Asymptotic Minimax Theorem for the Two Armed Bandit Problem. *The Annals of Mathematical Statistics*, 31(2):444–451. Publisher: Institute of Mathematical Statistics.

Wang, J., Wang, M., and Yang, L. F. (2022). Near Sample-Optimal Reduction-based Policy Learning for Average Reward MDP. *arXiv preprint arXiv:2212.00603*.

Wang, S., Blanchet, J., and Glynn, P. (2024). Optimal Sample Complexity for Average Reward Markov Decision Processes.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8:279–292. Publisher: Springer.

Watkins, C. J. C. H. (1989). Learning from delayed rewards. Publisher: King's College, Cambridge United Kingdom.

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. (2020). Model-free Reinforcement Learning in Infinite-horizon Average-reward Markov Decision Processes. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10170–10180. PMLR.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. (2003). Inequalities for the L1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*.

Zanette, A. and Brunskill, E. (2019). Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7304–7312. PMLR. ISSN: 2640-3498.

Zhang, Z. and Ji, X. (2019). Regret Minimization for Reinforcement Learning by Evaluating the Optimal Bias Function. In Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhang, Z., Ji, X., and Du, S. S. (2021). Is Reinforcement Learning More Difficult Than Bandits? A Near-optimal Algorithm Escaping the Curse of Horizon. *arXiv:2009.13503 [cs, stat]*. arXiv: 2009.13503.

Zhang, Z. and Xie, Q. (2023). Sharper Model-free Reinforcement Learning for Average-reward Markov Decision Processes. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5476–5477. PMLR.

Zhang, Z., Zhou, Y., and Ji, X. (2020). Almost Optimal Model-Free Reinforcement Learning via Reference-Advantage Decomposition. *arXiv:2004.10019 [cs, stat]*. arXiv: 2004.10019.

Zurek, M. and Chen, Y. (2024). Span-Based Optimal Sample Complexity for Average Reward MDPs.

Abstract

In this manuscript, we investigate the problem of regret minimization in Markov decision processes under the average gain criterion. In both the model independent (*aka* minimax) and model dependent settings, we provide new lower bounds on the expected regret as well as algorithmic methods achieving them — hence being optimal. Beyond regret minimization, we further study the trajectorial behavior of classical algorithms from a novel local viewpoint, through the lens of new learning metrics that quantify how algorithms choose actions locally rather than globally.

Résumé

Cette thèse est dédiée à la minimisation du regret dans les processus de décisions Markoviens en gain moyen. On y développe des bornes inférieures sur le regret en espérance, ainsi que des algorithmes atteignant les-dites bornes et de ce fait optimaux, dans deux cas fréquentistes classiques: le cas où le modèle est fixé et où la borne dépend du modèle, et le cas minimax où la borne s'applique uniformément à l'intégralité de la classe considérée. Au delà de la minimisation du regret, on étudie également le comportement local d'algorithmes classiques que le regret échoue à capturer, puisque intrinsèquement global. De nouvelles mesures de performances sont introduites dans cette optique.