Exploration-exploitation dilemma in Reinforcement Learning under various form of prior knowledge

Ronan Fruit




École doctorale n°072: *Sciences Pour l'Ingénieur Université Lille Nord-de-France*

Doctorat Université de Lille

THÈSE

pour l'obtention du grade de docteur délivré par

Spécialité doctorale "Informatique"

présentée et soutenue publiquement par

Ronan FRUIT

le 6 novembre 2019

Exploration–exploitation dilemma in Reinforcement Learning under various form of prior knowledge

Impact des connaissances a priori sur le compromis exploration-exploitation en apprentissage par renforcement

Directeur de thèse : **Daniil RYABKO**
Co-encadrant de thèse : **Alessandro LAZARIC**

Jury

M. Peter Auer,	Professeur	Rapporteur
Mme Shipra Agrawal,	Professeur Adjoint	Rapporteur
M. Marc Tommasi,	Professeur	Examinateur
Mme Doina Precup,	Professeur Associé	Examinateur
M. Aurélien Garivier,	Professeur	Examinateur
M. Alessandro Lazaric,	Chargé de recherche	Co-encadrant

Inria Lille – Nord-Europe, Équipe SequeL

Contents

Abstract

In combination with Deep Neural Networks (DNNs), several Reinforcement Learning (RL) algorithms such as "Q-learning" of "Policy Gradient" are now able to achieve super-human performaces on most Atari Games as well as the game of Go. Despite these outstanding and promising achievements, such Deep Reinforcement Learning (DRL) algorithms require millions of samples to perform well, thus limiting their deployment to all applications where data acquisition is *costly*. The lack of sample efficiency of DRL can partly be attributed to the use of DNNs, which are known to be *data-intensive* in the training phase. But more importantly, it can be attributed to the type of Reinforcement Learning algorithm used, which only perform a very inefficient *undirected* exploration of the environment. For instance, Q-learning and Policy Gradient rely on *randomization* for exploration. In most cases, this strategy turns out to be very ineffective to properly balance the *exploration* needed to discover unknown and potentially highly rewarding regions of the environment, with the *exploitation* of rewarding regions already identified as such. Other RL approaches with theoretical guarantees on the *exploration-exploitation trade-off* have been investigated. It is sometimes possible to formally prove that the performances almost match the theoretical optimum. This line of research is inspired by the Multi-Armed Bandit literature, with many algorithms relying on the same underlying principle often referred as *"optimism in the face of uncertainty"*. Even if a significant effort has been made towards understanding the exploration-exploitation dilemma generally, many questions still remain open. In this thesis, we generalize existing work on exploration-exploitation to different contexts with different amounts of *prior knowledge* on the learning problem. We introduce several algorithmic improvements to current state-of-the-art approaches and derive a new theoretical analysis which allows us to answer several open questions of the literature. We then relax the (very common although not very realistic) assumption that a path between any two distinct regions of the environment should always exist. Relaxing this assumption highlights the impact of prior knowledge on the intrinsic limitations of the exploration-exploitation dilemma. Finally, we show how some prior knowledge such as the range of the value function or a set of macro-actions can be efficiently exploited to speed-up learning. In this thesis, we always strive to take the *algorithmic complexity* of the proposed algorithms into account. Although all these algorithms are somehow computationally "efficient", they all require a planning phase and therefore suffer from the well-known "curse of dimensionality" which limits their applicability to real-world problems. Neverthe-

less, the main focus of this work is to derive *general principles* that may be combined with more heuristic approaches to help overcome current DRL flaws.

Résumé

Combinés à des réseaux de neurones profonds (”Deep Neural Networks”), certains algorithmes d'apprentissage par renforcement tels que ”Q-learning” ou ”Policy Gradient” sont désormais capables de battre les meilleurs joueurs humains à la plupart des jeux de console Atari ainsi qu'au jeu de Go. Malgré des résultats spectaculaires et très prometteurs, ces méthodes d'apprentissage par renforcement dit ”profond” (”Deep Reinforcement Learning”) requièrent un nombre considérable d'observations pour apprendre, limitant ainsi leur déploiement partout où l'obtention de nouveaux échantillons s'avère *coûteuse*. Le manque d'efficacité de tels algorithmes dans l'exploitation des échantillons peut en partie s'expliquer par l'utilisation de réseaux de neurones profonds, connus pour être très *gourmands* en données. Mais il s'explique surtout par le recours à des algorithmes de renforcement explorant leur environnement de manière inefficace et *non ciblée*. Ainsi, des algorithmes tels que Q-learning ou encore Policy-Gradient exécutent des actions partiellement randomisées afin d'assurer une exploration suffisante. Cette stratégie est dans la plupart des cas inappropriée pour atteindre un bon compromis entre l'*exploration* indispensable à la découverte de nouvelles régions avantageuses (aux récompenses élevées), et l'*exploitation* de régions déjà identifiées comme telles. D'autres approches d'apprentissage par renforcement ont été développées, pour lesquelles il est possible de garantir un meilleur *compromis exploration-exploitation*, parfois proche de l'optimum théorique. Cet axe de recherche s'inspire notamment de la littérature sur le cas particulier du problème du bandit manchot, avec des algorithmes s'appuyant souvent sur le principe *”d'optimisme dans l'incertain”*. Malgré les nombreux travaux sur le compromis exploration-exploitation, beaucoup de questions restent encore ouvertes. Dans cette thèse, nous nous proposons de généraliser les travaux existants sur le compromis exploration-exploitation à des contextes différents, avec plus ou moins de *connaissances a priori*. Nous proposons plusieurs améliorations des algorithmes de l'état de l'art ainsi qu'une analyse théorique plus fine permettant de répondre à plusieurs questions ouvertes sur le compromis exploration-exploitation. Nous relâchons ensuite l'hypothèse peu réaliste (bien que fréquente) selon laquelle il existe toujours un chemin permettant de relier deux régions distinctes de l'environnement. Le simple fait de relâcher cette hypothèse permet de mettre en lumière l'impact des connaissances a priori sur les limites intrinsèques du compromis exploration-exploitation. Enfin, nous montrons comment certaines connaissances a priori comme l'amplitude de la fonction valeur ou encore des ensembles de macro-actions

peuvent être exploitées pour accélérer l'apprentissage. Tout au long de cette thèse, nous nous sommes attachés à toujours tenir compte de la *complexité algorithmique* des différentes méthodes proposées. Bien que relativement efficaces, tous les algorithmes présentés nécessitent une phase de planification et souffrent donc du problème bien connu du "fléau de la dimension", ce qui limite fortement leur potentiel applicatif (avec les méthodes actuelles). L'objectif phare des présents travaux est d'établir des *principes généraux* pouvant être combinés avec des approches plus heuristiques pour dépasser les limites des algorithmes actuels.

Acknowledgements

Prima di tutto, vorrei ringraziare Alessandro Lazaric, il mio PhD advisor, per le grandi competenze e aiuto che mi ha fornito durante i tre anni, oltre alla pazienza, disponibilità e gentilezza mostratemi in ogni occasione. Senza te, questo lavoro non avrebbe preso vita! Le tue osservazioni mi hanno spronato ad andare in profondità ed esigere coerenza da me stesso. Di questo non potrò mai esserti abbastanza riconoscente. Grazie per tutto quello che mi hai insegnato e tutte le opportunità che mi hai offerte durante questa esperienza! Grazie anche per i tuoi incoraggiamenti e consigli in questi anni!

Vorrei anche ringraziare (Egr. Dottore) Daniele Calandriello per aver sempre risposto in modo esaustivo a tutte le domande a cui non avevo mai pensato, per la vasta conoscenza che ha messo a mia disposizione su quasi tutti gli argomenti esistenti, e per avermi spiegato il senso della vita. Ho sempre provato ammirazione per la tua capacità di apparire nei momenti più inaspettati e sparire altrettanto improvvisamente. Più seriamente, sono stato fortunato di avere avuto un amico come te dal primo giorno della tesi. Ho veramente apprezzato le discussioni con te e anche il tuo umorismo. Grazie per avermi strappato un sorriso anche quando ero giù di morale (ad Inria o in conferenza). Mi mancheranno le discussioni con te in Italiano (a dire il vero, devo ancora perfezionare il mio dialetto romano).

Per finire, un pensiero speciale lo dedico a Matteo Pirotta (anche conosciuto sotto il nome "Mathéau" qui in Francia). Mi hai sempre incoraggiato e sostenuto nelle mie ricerche, anche nei momenti di dubbio e di interrogazioni. Grazie per avermi trasmesso entusiasmo e coraggio durante questi 2+ anni. Non potrò mai ringraziarti per tutto l'aiuto che mi hai dato (sia in LaTeX, linux, git, python, per citarne alcuni). Senza te, arrivare alla difesa del dottorato sarebbe stato molto più doloroso! Collaborare con te è stata un esperienza fantastica. Grazie per avermi dato continuamente nuovi spunti di approfondimento per le mie ricerche. È difficile per me esprimere quanto ti sia grato (non solo perché scrivo in Italiano, in Francese sarebbe lo stesso). Grazie di cuore.

Rád by som venoval osobitnú myšlienku Michalovi Valkovi za všetko, čo pre mňa počas môjho doktorandského štúdia urobil. Či už v práci, alebo v osobnom živote. Nikdy sa ti nepod'akujem dostatočne a úprimne dúfam, že v budúcnosti budeme v kontakte.

During my PhD, I have had the opportunity to collaborate with many outstanding researchers who greatly inspired the work presented in this thesis. They gave me many in-

sightful comments and suggestions that often steered me in the right direction. I would like to express my deepest gratitude to Ronald Ortner for our fruitful collaborations and also for offering me the opportunity to start a postdoc at Loeben university (even if it did not happen). I would also like to thank Emma Brunskill for inviting me to CMU and Stanford University to collaborate with her team and her. I am also extremely grateful to Joelle Pineau for offering me the opportunity to do an internship at FAIR Montreal under her supervision. During this internship, I had the privilege to work with remarkable people like Mohammad Ghavamzadeh and Ahmed Touati. Finally, many thanks to Jian Qian who gave me many new stimulating ideas while he was an intern at Inria.

Mes remerciements se tournent tout d'abord vers ceux qui, par leur soutien inconditionnel ainsi que leurs nombreux encouragements durant les quatre dernières années, m'ont porté, à savoir ma famille (mes parents Sylvie et Éric, mes frères et soeurs Inès et Côme ainsi que mes grands-parents Anne-Marie, Dolorès, Bernard et Daniel), mais également mes proches et amis (Pauline mon amie d'enfance, Hugo mon ami du primaire avec qui je rêvais de devenir scientifique, mes amis du lycée: Camille, Corentin, Damien, Elena, Marion, Nicolas, Rodolphe, Urbain, et du supérieur: Niels, Vianney, Hamza). J'aimerais particulièrement remercier Delphine Hove d'avoir eu la patience de m'écouter présenter mes travaux de recherches avant de partir en conférence. Tes précieux conseils m'ont beaucoup aidé à progresser. Merci aussi d'avoir relu et corrigé les paragraphes de ce manuscrit écrits en français.

Mes remerciements s'adressent aussi à chacun des membres de l'équipe SequeL –actuels ou passés– que j'ai eu la chance de rencontrer: Philippe, Emilie, Odalric, Romaric, Jérémy, Olivier, Bilal, Daniil, Pierre, Sadegh, Lilian, Mahsa, Nicolas, Yannis, Guillaume, Jean-Bastien, Edouard, Hassan, Mathieu, Xuedong, Florian, Marta, Frédéric, Alexandre, Merwan, Tomáš, Victor, Gergely, Marc. Tous, par leur excellence scientifique, mais aussi leur bienveillance et leur enthousiasme, ont contribué à mon épanouissement en tant que chercheur. Je pense aussi à tous ceux qui sur mon parcours de thèse sont devenus des amis en dehors de mon équipe de recherche: Géraud, Claire, Guillaume.

Je tiens à remercier Inria dont le rayonnement international m'a permis de développer connaissances et compétences au contact de chercheurs de tous horizons.

Merci également à la région Haut de France qui a assuré son soutien matériel à mes recherches.

1 Introduction

1.1 Topic of the thesis

In this thesis we study the problem of a *"rational agent"* evolving in an unknown *"environment"*. The goal of the agent is to *learn* a "good" *behavior* (according to some notion of preferences) from the experience directly collected while exploring the environment.

Reinforcement Learning (RL) formalizes this problem through an *"economic"* perspective: the agent aims at maximizing some notion of cumulative *reward* (or equivalently, at minimizing a cumulative *loss*). In order to account for the presence of random events in the environment, it is usually assumed that the agent satisfies Von Neumann–Morgenstern's axioms of *rationality* (Von Neumann and Morgenstern, 1947). Under these axioms, Von Neumann–Morgenstern's *"utility theorem"* implies that the *"preferences"* of the agent can be expressed as maximizing the *expectation* of a certain *utility function* (which corresponds to the cumulative reward in an RL context).

The environment of an RL problem, or RL *"task"*, is traditionally modeled by a Markov Decision Process (MDP). An MDP consists of a set of states (usually functions of some observables) and actions. When the agent decides to *"play"* a certain action in a given state, it receives some (possibly random) reward and moves to the next state according to a certain probability distribution over the state space. By definition, this type of process satisfies the *Markov property* i.e., future events depend only upon the present state and chosen action, and not the whole past history. This restrictive assumption enables to considerably simplify the problem. It is always possible to expand the state space so as to enforce the Markov property, at the expense of increasing the complexity of the problem. In practice, the size of the state space must be traded-off with the accuracy of the Markov property.

While evolving in an MDP, an agent aims at identifying which control *policy* to execute i.e., which action to perform depending on past observations. When the MDP is completely known, finding an *"optimal"* policy is a dynamic programming problem (Bellman, 1954). An even more challenging setting is when the MDP is unknown and has to be *learned* (RL problem). In this thesis, we restrict attention to *online* RL. In this setting, data about the environment becomes available in a *sequential order* as the agent *explores* the MDP. As the

MDP is being explored, the agent needs to update its behavior so as to be able to make *better decisions*. But unlike in other branches of Machine Learning like *supervised learning*, any present decision impacts future observations. As a consequence, the agent has to deal with two conflicting objectives, namely:

1. collecting information about the dynamics and reward of the environment which may allow to make better decisions in the future (*exploration*),

2. using the experience gathered so far to maximize the chances to gain as much reward as possible quickly (*exploitation*).

This problem is known as the *exploration-exploitation dilemma*. The work presented in this thesis focuses on the exploration-exploitation dilemma in an on-line RL setting, under various assumtions, and in different contexts. This problem was first studied in the simplified case of *Multi-armed bandit* (MAB) in the seminal works of Thompson (1933a); Lai and Robbins (1985). Since then, considerable progress has been made although many open questions still remain unanswered.

1.2 Motivations

One of the long-standing goal of *Artificial Intelligence* (AI) is to design robust, autonomous agents able to perform well in complex, real-world environments. Reinforcement Learning provides a promising framework to achieve some of these goals as evidenced by recent empirical achievements. In combination with *Deep Learning* techniques, RL algorithms are now able to achieve super-human performances on Atari games (Mnih et al., 2015b) or the challenging game of Go (Silver et al., 2016, 2017; Silver et al., 2017). Nevertheless, *Deep Reinforcement Learning* (DRL) algorithms require millions of *samples* to be trained, and can perform very poorly in environments with sparse reward like Atari 2600 game Montezuma's Revenge. In such environments, the agent only observes a reward signal after completing specific series of actions over extended periods of time, making the *exploration* of the environment very challenging. In other domains, samples can be expensive to collect (computationally or in terms of actual cost). Unfortunately, most of potential real-world applications of RL have these characteristics.

The lack of *sample efficiency* of DRL is a major obstacle to its deployment in real-world applications. This lack of sample efficiency mainly comes from the exploration strategy used, which often relies on randomization to discover unknown regions of the environment (e.g., ε-greedy strategies may require an exponential amount of time in the parameters of the MDP to converge). We say that the exploration is *undirected*. A major open question in RL is how to design efficient *directed* exploration strategies that make best use of all the *prior information* available about the problem being solved. The work of this thesis is motivated by a better understanding of the exploration-exploitation dilemma in RL, and the impact of *prior knowledge* on the intrinsic difficulty of this dilemma. We hope this work helps suggest promising research directions to improve the sample-efficiency of existing RL algorithms.

1.3 Scientific approach

Instead of restricting attention to very specific RL tasks/applications, we analyse the *theoretical properties* of some general RL problems. We study various settings, which mostly differ by the amount of prior knowledge available to the learning agent. For all these different settings, we analyse the learning limitations (e.g., impossibility results) and derive learning algorithms that attempt to achieve the best possible exploration-exploitation performance given these limitations.

While efficient exploration-exploitation strategies in RL are directly inspired by the MAB literature, RL poses specific challenges (e.g., how "local" uncertainty propagates through the Markov dynamics), which requires a more sophisticated theoretical analysis. Most of the algorithms that have been analysed theoretically belong to one of the following two categories:

1. *optimistic* algorithms,
2. *posterior sampling* (also known as Thomson sampling) algorithms.

Optimisitc algorithms implement the *"Optimism in the face of uncertainty"* principle which essentially prescribes to play the optimal policy of the most rewarding environment compatible with the current level of uncertainty (often quantified by confidence sets). Posterior sampling involves sampling a statistically plausible set of environments and selecting the best policy. The sampling distribution is then updated based on new observations. While both methods can be proved to achieve good exploration-exploitation performance in MAB (Kaufmann et al., 2012), so far optimistic approaches appear more promising in the general RL setting. For this reason, all the algorithms presented in this thesis are of optimistic nature.

For all the proposed algorithms, we apply a unified *statistical analysis* and systematically rely on the same mathematical tools/arguments. This allows to easily compare settings and to better understand the impact of assumptions on the learning capabilities.

The statistical analysis of RL algorithms help make a clear distinction between the intrinsic difficulty of an RL task (e.g., Montezuma) and the lack of efficiency of the algorithm used (e.g., DQN). Unfortunatley, none of the algorithms proposed in this thesis *scale* to large dimensional problems due to the notorious *"curse of dimensionality"* that also appear in dynamic programming (Bellman, 1954). Despite this lack of scalability, we hope to provide insightful *principles* that can inspire future research and algorithm design.

1.4 Open research questions of the literature

We list two very general research questions that were open at the beginning of this work in 2015 and will be only partly answered in the rest of the thesis.

- **What is the best exploration-exploitation trade-off an RL algorithm can achieve and how?** This question is the main leitmotiv of the thesis. In the next chapter we will see that the learning capabilities of any learning algorithm are intrinsically limited,

and these limitations can be statistically quantified. One natural objective is to design algorithms that can achieve the best trade-off given these inherent restrictions. Back in 2015, no exisitng algorithm had been proved "optimal" in this sense. This question is very general and the answer is of course problem-dependent and depends on many different aspects of the setting studied. For a more technical and detailed overview of some specific sub-questions, one may refer to the presentation given by Ortner (2016).

- **Under which conditions hierarchical approaches (such as options) help speed-up the learning process?** The option framework was developped to incorporate temporally extended actions and hierarchical reasoning to RL. The motivation is to mimic the ability of humans to identify and exploit the hierarchical structure of many RL tasks which naturally decompose into easier subtasks. It is believed that this partly explains how we (humans) manage to learn so well. Unfortunately, a formal understanding of how and when options are efficient was still missing.

1.5 Outline of the thesis

The thesis is organized as folows:

- **Chapter 2.** This chapter provides a brief introduction to the exploration-exploitation dilemma in RL and reviews the state-of-the-art literature relevant for the rest of the thesis. At first, we review the concept of Markov Decision Process and several optimality criteria. After the introduction of a dynamic programming algorithm known as value iteration, we briefly review the stochastic shortest path problem. In the second part, we focus on the exploration-exploitation literature in the specific case of infinite horizon undiscounted setting. We formally define a useful exploration-exploitation performance measure named "regret" and present several regret upper and lower-bounds. The reader who is already familiar with these topics may skip this chapter.

- **Chapter 3.** In this chapter, we present and analyse UCRLB, a variant of the learning algorithm UCRL2 (Jaksch et al., 2010). We prove that our version of the algorithm achieves better regret guarantees (i.e., exploration-exploitation trade-off), thus answering some of the open-questions on the gap between upper and lower regret bounds. All the other learning algorithms presented in this thesis will share many algorithmic bricks with UCRLB, and the structure of the regret proofs will be re-used across all chapters. In order to prepare for subsequent chapters, we prove intermediate results in their full generality. Several key passages of the regret proofs are presented from a slightly different perspective than is usually done in the existing literature (e.g., proof of optimism, bound on the optimistic bias). We recommand to carefully go through the entire chapter before reading the rest of the thesis.

- **Chapter 4.** In this chapter, we provide the first learning algorithm achieving near-optimal regret guarantees when the diameter of the MDP is infinite i.e., some states cannot be reached. This answers one of the open questions of the literature. We show that such setting poses specific challenges and we derive an impossibility result that we

believe is new to the exploration-exploitation literature. This is all the more surprising as it appears to apply to most RL tasks encountered in practice.

- **Chapter 5.** This chapter extends the work of Bartlett and Tewari (2009) by showing how to exploit prior knowledge on the range of the optimal bias span of the MDP to improve the learning performance (the regret). The methodology and mathematical tools used in this chapter provide a lot of insights on the minimal key properties needed to derive regret guarantees using an optimistic UCRL2-like approach. It also highlights the importance of focusing on operators (rather than MDPs) to derive and analyse RL algorithms. This follows the initial ideas of Bellman (1954) developped in the context of planning, and later extended to different RL settings.
- **Chapter 6.** In this last chapter, we analyze the exploration-exploitation trade-off in the presence of options. Our results shows when options provide a useful prior knowledge to address the exploration-exploitation dilemma.

2 Statistical analysis of the exploration-exploitation dilemma in RL

In this chapter we give a brief overview of the state-of-the-art literature on exploration–exploitation in RL. In Sec. 2.1, we formally define the notion of Markov Decision Process (MDP) used to mathematically describe the environment in which the learning agent evolves. An MDP describes a discrete-time decision problem where at each time step, the agent can choose between different available "actions" and is given some form of immediate motivation encoded into a "reward function". Because some decisions may have long-term consequences, it is not always easy to identify the best "policy" (mapping observations to actions) even when the MDP is completely known. We describe how to perform efficient planning in this case. Identifying the optimal policy becomes even more challenging when the MDP is unknown (learning setting). This problem is the focus of Sec. 2.2, where we survey the literature on exploration–exploitation in the infinite horizon undiscounted setting. We present several algorithms that can be proved to efficiently balance exploration and exploitation, and discuss their limitations.

2.1 Markov Decision Processes

In this section we briefly introduce the formalism of *Markov Decision Processes* and present several notions of *optimality*. We also recall all well-known results that will be useful for the next chapters (see e.g., Puterman, 1994; Bertsekas, 2007). We mainly follow the notations of Puterman (1994).

2.1.1 Definitions

States, actions, rewards and transitions

A *Markov Decision Process* M is formally defined as a 4-tuple $\langle \mathcal{S}, \mathcal{A}, r, p \rangle$. $\mathcal{S}$ and $\mathcal{A} = \bigcup_{s \in \mathcal{S}} \mathcal{A}_s$ respectively denote the *state* and *action* space. When in state s, an agent can choose to *play* any of the actions contained in $\mathcal{A}_s$. After playing action a in state s, the agent receives a

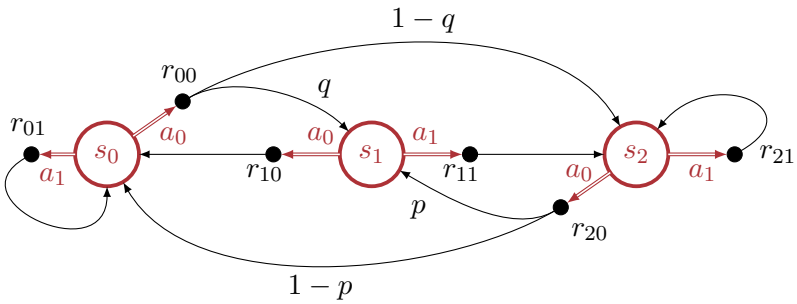


Figure 2.1: Graphical illustration of an MDP with 3 states (s_0, s_1 and s_2) and 2 actions per state (a_0 and a_1).

random *reward* with *expected value* $r(s,a)$, and then moves to a *new state* in $\mathcal{S}$ sampled according to a *stationary* distribution $p(\cdot|s,a)$. More precisely, the probability that the new state is s' is denoted $p(s'|s,a)$. By definition, $p(\cdot|s,a) \in \Delta_S$ where

Definition 2.1

$\Delta_S := \left\{q \in [0,1]^S : \ \sum_{s\in\mathcal{S}} q(s) = 1\right\}$ *is the S-dimensional probability simplex.*

The sampled reward and next state only depend on s and a and are *independent* of everything else. In this thesis, we restrict attention to MDPs with *finite* state space and denote by $S = |\mathcal{S}|$ the total *number of states*. We will consider MDPs with *finite* as well as *compact* action spaces[1]. When the action space is *finite*, we will denote by $A = \max_{s\in\mathcal{S}} |\mathcal{A}_s|$ the maximal *number of actions* available in every state. All *sampled* rewards are assumed to be *bounded* and without loss of generality, we assume that they lie in $[0, r_{\max}]$ where $r_{\max} > 0$. When the action space is *compact*, we further assume that for any two states $s, s' \in \mathcal{S}$, $a \mapsto r(s,a)$ and $a \mapsto p(s'|s,a)$ are *continuous* functions of a. Under these assumptions and –unless stated otherwise– all the results of this Chapter hold for *both* finite and compact action spaces. Although the action space is state-dependent, in the rest of the thesis, we will slightly abuse notation and denote by $\mathcal{S}\times\mathcal{A}$ the set of *"admissible"* state-action pairs i.e., the set $\{(s,a) : \ s\in\mathcal{S},\ a\in\mathcal{A}_s\}$. An example of graphical representation of an MDP is given in Fig. 2.1.

Sequential decision making

In this thesis, we assume that an *agent* can only make "decisions" at *discrete* time steps (often called "epochs") and so we exclusively focus on *discrete* sequences indexed by $t \in \mathbb{N}^+$, where $\mathbb{N}^+ := \mathbb{N}\setminus\{0\}$ is the set of (strictly) positive integers. At any time $t \geq 1$, the agent is in state s_t and plays action a_t. The (random) reward earned by the agent and the next state are respectiely denoted by r_t and s_{t+1}. This procedure is repeated thus generating a sequence of the form $(s_1, a_1, r_1, \ldots, s_t, a_t, r_t, \ldots)$ that we call a *"history"* (sometimes called a "sampled path"). The set of all possible histories up to time $t \geq 1$ is formally defined as

$$\mathcal{H}_t := \left\{(s_1, a_1, r_1, \ldots s_{t-1}, a_{t-1}, r_{t-1}, s_t) : \ \forall l \leq t,\ s_l \in \mathcal{S},\ a_l \in \mathcal{A}_{s_l},\ r_l \in [0, r_{\max}]\right\}. \quad (2.1)$$

[1] In this thesis, a compact set always refer to the compact subset of a metric space.

Policies and induced stochastic processes

The set of all probability distributions over the state space $\mathcal{S}$ (resp. action space $\mathcal{A}$) is denoted by $\mathcal{P}(\mathcal{S})$ (resp. $\mathcal{P}(\mathcal{A})$). For any $t \geq 1$, a *decision rule* $d_t : \mathcal{H}_t \to \mathcal{P}(\mathcal{A})$ maps *histories* of past *observations* (i.e., past states, actions and rewards) to *distributions* over actions. The set of decision rules is denoted D^{HR} where HR stands for *"history-dependent"*. This is the *most general* definition of decision rule we can think of. Decisions based on future events are *forbidden* to avoid causal inconsistency. We also introduce two specific types of decision rules. A *Markov randomized* decision rule $d : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ maps states to distributions over actions while a *Markov deterministic* decision rule $d : \mathcal{S} \to \mathcal{A}$ maps states to actions. Markov decision rules only take into account the *current state* and completely ignore previous observations. The subset of Markov randomized decision rules is denoted D^{MR}, while the subset of Markov deterministic decision rules is denoted D^{MD}. For any Markov decision rule $d \in D^{\text{MR}}$, $P_d \in \mathbb{R}^{S \times S}$ and $r_d \in \mathbb{R}^S$ denote the *transition matrix* and *reward vector* associated with d i.e.,

$$P_d(s'|s) := \sum_{a \in \mathcal{A}_s} d(a|s) p(s'|s,a) \quad \text{and} \quad r_d(s) := \sum_{a \in \mathcal{A}_s} d(a|s) r(s,a), \quad \text{for all } s, s' \in \mathcal{S}, \tag{2.2}$$

where $d(a|s)$ is the probability to sample a in state s when using d.

A *policy* $\pi = (d_1, d_2, d_3 \dots) \in \left(D^{\text{HR}}\right)^{\mathbb{N}^+}$ is a *sequence* of decision rules. At every time step $t \geq 1$, an agent executing policy π samples an action a_t from the distribution $d_t(h_t)$ that *only* depends on the past "observed" trajectory $h_t \in \mathcal{H}_t$. The set of all policies is denoted by Π. A *stationary* policy $\pi = (d, d, \dots) =: d^\infty$ repeatedly applies the same *Markov* decision rule $d \in D^{\text{MR}}$ over time. The set of stationary policies defined by Markov randomized (resp. deterministic) decision rules is denoted by Π^{SR} (resp. Π^{SD}). In the rest of the thesis, we will slightly abuse notations and use d and π *interchangeably* when $\pi = d^\infty \in \Pi^{\text{SR}}$ is stationary.

For a given MDP M, a policy $\pi \in \Pi$ and an initial distribution over states $\mu_1 \in \mathcal{P}(\mathcal{S})$, the induced sequence $(s_1, a_1, r_1, \dots, s_t, a_t, r_t, \dots)$ is a *stochastic process* with a well-defined *probability distribution* (Puterman, 1994, Section 2.1.6) (in particular, the items s_t, a_t and r_t are *random variables*). In the rest of the thesis, we will denote by $\mathbb{P}^\pi(\cdot|s_1 \sim \mu_1)$ the *probability measure* associated with this stochastic process and denote by $\mathbb{E}^\pi[\cdot|s_1 \sim \mu_1]$ the corresponding *expectation*. When there is ambiguity on which MDP we are considering, we use M as a subscript $\mathbb{P}^\pi_M(\cdot|s_1 \sim \mu_1)$ to denote the probability in MDP M.

In the special case where the policy $\pi \in \Pi^{\text{SR}}$ is stationary, the induced sequence of visited states $(s_1, s_2, \dots)$ is a specific stochastic process called a (discrete-time stationary) *Markov Chain* (MC). On the other hand, the stochastic process corresponding to the sequence of states and rewards $(s_1, r_1, s_2, r_2, \dots)$ is a (discrete-time stationary) *Markov Reward Process* (MRP). The interested reader may refer to Puterman (1994, Appendix A) for a brief overview of the theory on Markov Chains and Markov Reward Processes, and to Bremaud (1999); Grinstead and Snell (2003) for more details.

We classify MDPs depending on the *chain structure* of stationary policies (i.e., depending on how states are connected to each other through the dynamics). For the following definition,

we assume the reader to be familiar with the notions of *transient* and (positive) *recurrent* states and/or class of a Markov Chain (for more details, refer to Puterman (1994, Appendix A)).

Definition 2.2 (Classification of MDPs)

We say that an MDP is:

1. **ergodic** *if the Markov Chain induced by any deterministic stationary policy consists of a single* recurrent class *(i.e., all states are visisted infinitely often with probability* 1 *independently of the starting state)*
2. **unichain** *if the Markov Chain induced by any deterministic stationary policy consists of a single recurrent class plus a –possibly empty– set of* transient *states (i.e., there exists a subset of states that are visisted infinitely often with probability* 1 *independently of the starting state)*
3. **communicating** *if for every pair of states* $(s, s') \in \mathcal{S}$*, there exists a deterministic stationary policy under which* s' *is* accessible *from* s *in finite time with non-zero probability,*
4. **weakly communicating** *if the state space can be* partitioned *into two subsets* $\mathcal{S}^C$ *and* $\mathcal{S}^T$ *(with* $\mathcal{S}^T$ *possibly empty), such that for every pair of states* $(s, s') \in \mathcal{S}^C$*, there exists a deterministic stationary policy under which* s' *is accessible from* s *in finite time with non-zero probability, and all states in* $\mathcal{S}^T$ *are transient under all deterministic stationary policies.*

When we want to emphasize that we do not make any of the above assumptions but rather consider a general MDP, we will use the terminology "multi-chain" MDP.

In this thesis, we will see that the chain structure of the MDP can *limit* the performance of an (optimal) RL algorithm.

2.1.2 Finite horizon problems

Now that we have formally defined how an agent sequentially interacts with its environment in the MDP framework, we need to formulate the problem we want to solve i.e., the goal of the agent. Intuitively, the agent aims at executing a policy maximizing the *sum* of collected rewards $\sum_t r_t$. Unfortunately, this series will often diverge as $t \to +\infty$ and it is a priori not obvious how to compare infinite quantities. A first simple setting where this problem does not occur is when the agent maximizes the *cumulative reward* up to a *fixed* horizon H i.e., maximizes $\sum_{t=1}^{H} r_t$. Since $(r_t)_{t\geq 1}$ is a stochastic process, this sum cannot always be maximized and the agent will try to maximize the *expected value* instead (in line with Von Neumann–Morgenstern's axioms of rationality (Von Neumann and Morgenstern, 1947)). Formally, in the *finite horizon setting* –with horizon H– the goal is to solve the following optimization problem:

$$\sup_{\pi \in \Pi} \left\{ \mathbb{E}^{\pi} \left[\sum_{t=1}^{H} r_t \middle| s_1 \sim \mu_1 \right] \right\} \tag{2.3}$$

Algorithm 1 Backward Induction

Input: Operators $L : \mathbb{R}^S \mapsto \mathbb{R}^S$ and $G : \mathbb{R}^S \mapsto D^{\text{MR}}$, horizon H
Output: Optimal n-step expected cumulative sum of rewards v_n^* and n-step optimal policy π_n^* for $n \in \{1 \dots H\}$

Initialize $v_{H+1} := 0$
for $n = H \dots 1$ **do**
  $(v_n^*, d_n^*) := (Lv_{n+1}^*, Gv_{n+1}^*)$ ▷ Lv_{n+1}^* and Gv_{n+1}^* can be computed simultaneously
  $\pi_n^* \leftarrow (d_n^*, \dots, d_H^*)$
end for

where the initial state s_1 is sampled from distribution $\mu_1 \in \mathcal{P}(\mathcal{S})$. It is well known (see e.g., Puterman, 1994, Chapter 4) that there always exists an *optimal* policy $\pi^* = (d_1^*, d_2^*, \dots, d_H^*)$ solution to (2.3) for any $\mu_1 \in \mathcal{P}(\mathcal{S})$ and such that for all $H \geq t \geq 1$, $d_t^* \in D^{\text{MD}}$ i.e., d_t^* is *Markov deterministic* and *independent* of the initial state disribution.

For any Markov decision rule $d \in D^{\text{MR}}$, we define L_d the *Bellman evaluation operator* of d as

$$\forall v \in \mathbb{R}^S, \quad L_d v := r_d + P_d v. \tag{2.4}$$

We also define L the *optimal Bellman operator*

$$\forall v \in \mathbb{R}^S, \quad Lv := \max_{d \in D^{\text{MD}}} \{L_d v\}, \tag{2.5}$$

as well as the *greedy operator*[2]

$$\forall v \in \mathbb{R}^S, \quad Gv \in \arg\max_{d \in D^{\text{MR}}} \{L_d v\}. \tag{2.6}$$

It is always possible to compute an optimal policy π^* of (2.3) by *backward induction* as described in Alg. 1. The following proposition is a well-known result of the literature on *dynamic programming* (Puterman, 1994, Section 4.3).

Proposition 2.1

For all $n = 1 \dots H$ and all $s \in \mathcal{S}$, the value functions v_n^* and policies π_n^* returned by Alg. 1 satisfy

$$v_n^*(s) = \max_{\pi \in \Pi} \mathbb{E}^\pi \left[\sum_{t=n}^{H} r_t \middle| s_n = s \right] \quad \text{and} \quad \pi_n^* = (d_n^*, \dots, d_H^*) \in \arg\max_{\pi \in \Pi} \mathbb{E}^\pi \left[\sum_{t=n}^{H} r_t \middle| s_n = s \right]$$

A direct consequence of Prop. 2.1 is that $\pi^* = (d_1^*, \dots, d_H^*)$ is a maximizer of (2.3) for any $\mu_1 \sim \mathcal{P}(\mathcal{S})$ and $\mu_1^\intercal v_1^*$ is the corresponding maximum.

[2] We break ties arbitrarily when several greedy decision rules exist. It is always possible to choose $d \in D^{\text{MD}}$ but for the sake of generality, we allow any greedy *randomized* decision rule D^{MR}.

2.1.3 Infinite horizon problems

Maximizing the cumulative sum of rewards only up to a *pre-defined* horizon H is not adapted to all problems. In many scenarios, there is no "obvious" way to define what a "good" horizon is. Most of the time, we ideally want an horizon that is as big as possible i.e., such as $H \to +\infty$. In this section, we review several well-established *optimality criteria* in the *infinite horizon setting*.

Discounted optimality

One of the most commonly used optimality criterion in infinite horizon problems is *discounted optimality*. Instead of maximizing a finite sum of rewards, the idea is to maximize an infinite sum of rewards *discounted* by a fixed pre-defined *discount factor* $0 < \gamma < 1$ i.e.,

$$\sup_{\pi \in \Pi} \left\{ \mathbb{E}^{\pi} \left[\sum_{t=1}^{+\infty} \gamma^{t-1} r_t \middle| s_1 \sim \mu_1 \right] \right\} \tag{2.7}$$

Since $0 < \gamma < 1$ and $r_t \in [0, r_{\max}]$, the infinite sum of rewards is a geometric series and remains bounded between 0 and $r_{\max}/(1-\gamma)$. The series always converges and is called the value function of policy π. It will be denoted v_{γ}^{π}. The maximization of (2.7) is therefore well-defined. It has long been known (Puterman, 1994, Chapter 6) that there always exists an *optimal* policy π^* solution to (2.7) for *all* initial distributions $\mu_1 \in \mathcal{P}(\mathcal{S})$ such that $\pi^* \in \Pi^{\text{SD}}$ i.e., there exists a *stationary deterministic* optimal policy that *does not depend* on the initial distribution $\mu_1 \in \mathcal{P}(\mathcal{S})$. This makes the solution of (2.7) even "simpler" than the solution of (2.3) (the optimal policy associated to (2.3) is not stationary in general). Moreover, the following proposition holds.

Proposition 2.2

There exists a unique *solution* v_{γ}^* *to the fixed-point equation* $v_{\gamma}^* = L_{\gamma} v_{\gamma}^*$ *where* L_{γ} *is the* discounted *optimal Bellman operator i.e.,* $L_{\gamma} v := \max_{d \in D^{MD}} \{r_d + \gamma P_d v\}$ *for all* $v \in \mathbb{R}^S$ *(see Eq. 2.5). In addition, for all* $s \in \mathcal{S}$*,*

$$v_{\gamma}^*(s) = \max_{\pi \in \Pi} \left\{ \mathbb{E}^{\pi} \left[\sum_{t=1}^{+\infty} \gamma^{t-1} r_t \middle| s_1 = s \right] \right\}$$

Finally, a stationary policy $\pi^* = (d^*)^{\infty} \in \Pi^{SR}$ *is optimal (i.e., solution to (2.7)) if and only if* $d^* = G_{\gamma} v_{\gamma}^* \in \arg\max_{d \in D^{MR}} \left\{ r_d + \gamma P_d v_{\gamma}^* \right\}$ *i.e.,* π^* *is a* greedy *policy with respect to* v_{γ}^**.*

Prop. 2.2 holds both for finite and compact action spaces and is merely a direct consequence of Banach fixed-point theorem applied to the γ*-contractive* operator L_{γ} in ℓ_{∞}-norm ($0 < \gamma < 1$). Due to Prop. 2.2, it is always possible to compute an optimal policy π^* of (2.7) by first finding a solution v_{γ}^* to the discounted *Bellman optimality equation* $v_{\gamma}^* = L_{\gamma} v_{\gamma}^*$ and then

Algorithm 2 (Discounted) Value Iteration

Input: Operators $L_\gamma : \mathbb{R}^S \mapsto \mathbb{R}^S$ and $G_\gamma : \mathbb{R}^S \mapsto D^{\text{MD}}$, discount factor $\gamma \in]0,1[$, accuracy $\varepsilon \in]0, r_{\max}[$

Output: Value function $v \in \mathbb{R}^S$ and stationary deterministic policy $\pi \in \Pi^{\text{SD}}$

Initialize $n = 0$ and $v_0 := 0$
$v_1 := L_\gamma v_0$
while $\max\{v_{n+1} - v_n\} - \min\{v_{n+1} - v_n\} > \frac{(1-\gamma)\varepsilon}{\gamma}$ **do** ▷ Loop until termination
    Increment $n \leftarrow n + 1$
    $(v_{n+1}, d_n) := (L_\gamma v_n, G_\gamma v_n)$ ▷ $L_\gamma v_n$ and $G_\gamma v_n$ can be computed simultaneously
    $d_n \in \arg\max_{d \in D^{\text{MD}}} \{L_d^\gamma v_n\}$
end while
Set $v := v_n$ and $\pi := (d_n)^\infty$

considering a *greedy policy* w.r.t. v_γ^*. In order to find an ε-approximate solution (in ℓ_∞-norm) to (2.7), it is possible to apply the same *iterative scheme* as in the finite horizon case (Alg. 1) but with few modifications, as reported on Alg. 2. This algorithm is known as *value iteration*. Since L_γ is a γ-contraction, value iteration always *converges*: $\lim_{n \to +\infty} v_n = v_\gamma^*$ (this is also a consequence of the Banach fixed point theorem). Therefore, Alg. 2 always stops after a finite number of iterations and the policy π returned by Alg. 2 is such that $\|v_\gamma^\pi - v_\gamma^*\| \leq \varepsilon$. Finally, the maximum of (2.7) is equal to $\mu_1^\intercal v_\gamma^*$.

The discounted setting is particularly well-suited for problems with a pre-defined *random* horizon H that follows a *geometric distribution* with parameter $1-\gamma$ (note that in Sec. 2.1.2, H is deterministic). In this view, the agent is seen as "tossing a coin" at every time steps $t \geq 1$ and stopping collecting rewards in the MDP with probability $1-\gamma$ (and keeping going on with probability γ). Then, the expected discounted sum of rewards corresponds exactly to the expected total sum of rewards (accounting for the random horizon H) i.e.,

$$\mathbb{E}^\pi \left[\sum_{t=1}^{+\infty} \gamma^{t-1} r_t \middle| s_1 \sim \mu_1 \right] = \mathbb{E}^\pi \left[\sum_{t=1}^{H} r_t \middle| s_1 \sim \mu_1, H \sim \text{Geom}(1-\gamma) \right].$$

The expected value of H is $1/(1-\gamma)$ and so the discounted setting somehow resembles the finite horizon setting with $H = \Theta(1/(1-\gamma))$. As a result, it suffers the same problem as before: in many scenarios there is no obvious way to define γ and we want to set it as close to 1 as possible i.e., $\gamma \to 1$.

Gain optimality

We now present the *infinite horizon undiscounted setting* which uses the *gain* –or *long-term average reward*– as optimality criterion. Formally, in this setting the agent aims at solving the following optimization problem:

$$\sup_{\pi \in \Pi} \left\{ \liminf_{T \to +\infty} \mathbb{E}^\pi \left[\frac{1}{T} \sum_{t=1}^{T} r_t \middle| s_1 \sim \mu_1 \right] \right\}. \tag{2.8}$$

Since for all $t \geq 1$, r_t lies in $[0, r_{\max}]$ (by assumption), so does $1/T \cdot \sum_{t=1}^{T} r_t$. When the policy is *stationary* i.e., $\pi \in \Pi^{\text{SR}}$, the lim inf in Eq. 2.8 actually matches the lim sup. The *limit* is therefore well-defined and is called the *gain* (Puterman, 1994, Section 8.2.1). More precisely, the gain of policy $\pi \in \Pi^{\text{SR}}$ starting from initial state $s \in \mathcal{S}$ is defined as

$$g^{\pi}(s) := \lim_{T \to +\infty} \mathbb{E}^{\pi}\left[\frac{1}{T}\sum_{t=1}^{T} r_t \middle| s_1 = s\right]. \tag{2.9}$$

The gain $g^{\pi}(s)$ corresponds to the *asymptotic per-step* reward earned when executing policy π starting from $s \in \mathcal{S}$. This notion *generalizes* both the finite and the discounted setting when $H \to +\infty$ and $\gamma \to 1$ respectively since it can be shown (Puterman, 1994, Sections 8.2.1 and 8.2.2) that for all $s \in \mathcal{S}$

$$\mathbb{E}^{\pi}\left[\sum_{t=1}^{H} r_t \middle| s_1 = s\right] \underset{H \to +\infty}{\sim} g^{\pi}(s) \cdot H \quad \text{and} \quad \mathbb{E}^{\pi}\left[\sum_{t=1}^{+\infty} \gamma^{t-1} r_t \middle| s_1 = s\right] \underset{\gamma \to 1}{\sim} g^{\pi}(s)/(1-\gamma).$$

As a result, if $\pi, \pi' \in \Pi^{\text{SR}}$ are two stationary policies such that $\mu_1^{\intercal} g^{\pi} \geq \mu_1^{\intercal} g^{\pi'}$, then for H big enough and γ close enough to 1 we have that $\mathbb{E}^{\pi}\left[\sum_{t=1}^{H} r_t \middle| s_1 \sim \mu_1\right] \geq \mathbb{E}^{\pi'}\left[\sum_{t=1}^{H} r_t \middle| s_1 \sim \mu_1\right]$ and $\mathbb{E}^{\pi}\left[\sum_{t=1}^{+\infty} \gamma^{t-1} r_t \middle| s_1 \sim \mu_1\right] \geq \mathbb{E}^{\pi'}\left[\sum_{t=1}^{+\infty} \gamma^{t-1} r_t \middle| s_1 \sim \mu_1\right]$.

Any *stationary* policy $\pi \in \Pi^{\text{SR}}$ also has an associated *bias* function defined for all $s \in \mathcal{S}$ as

$$h^{\pi}(s) := \underset{T \to +\infty}{C\text{-}\lim}\, \mathbb{E}^{\pi}\left[\sum_{t=1}^{T} \left(r_t - g^{\pi}(s_t)\right) \middle| s_1 = s\right], \tag{2.10}$$

that measures the expected *cumulative difference* between the *immediate* reward r_t and the long term *asymptotic* reward $g^{\pi}(s)$ in *Cesaro-limit* (denoted C-lim). The Cesaro-limit is always well-defined unlike the "classical" limit as the series may *cycle* i.e., have several *accumulation points*[3]. Accordingly, the difference of bias values $h^{\pi}(s) - h^{\pi}(s')$ quantifies the (dis-)advantage of starting in state s rather than s'. We denote by $sp(h^{\pi}) := \max_s h^{\pi}(s) - \min_s h^{\pi}(s)$ the *span* (i.e., range) of the bias function. It is well-known (Puterman, 1994, Section 6.6) that the span defines a *semi-norm* on $\mathbb{R}^S$.

For any $d \in D^{\text{MR}}$, we also define the *limiting matrix* $P_d^* := \underset{n \to +\infty}{C\text{-}\lim}\, P_d^n$ (Puterman, 1994, Appendix A.4). The Cesaro limit always exists and so P_d^* is always well-defined. It is possible to express g^{π} (where $\pi = d^{\infty}$) in terms of P_d^* and r_d i.e., $g^{\pi} = P_d^* r_d$. The matrix $(I - P_d + P_d^*)$ is always invertible and $h^{\pi} = (I - P_d + P_d^*)^{-1}(I - P_d^*) r_d$ (Puterman, 1994, Appendix A). The matrix $H_{P_d} := (I - P_d + P_d^*)^{-1}(I - P_d^*)$ is called the *deviation matrix* and is the *Drazin inverse* of the matrix $I - P_d$.

Definition 2.3

In the rest of the thesis, we will define vector $e := (1, \ldots, 1)^{\intercal} \in \mathbb{R}^d$ *as the* d*-dimesional vector of all ones (*d *can vary depending on the context) and* $e_i := (0, \ldots, 1, \ldots, 0)^{\intercal}$ *as the* i*-th cartesian coordinate in* $\mathbb{R}^d$*.*

[3] Accumulation points are sometimed called "cluster points". Note that for policies with an *aperiodic chain*, the standard limit exists.

Proposition 2.3 (Theorem 8.2.6 of Puterman (1994))

For any policy $\pi = d^\infty \in \Pi^{SR}$, the gain g^π and bias h^π satisfy the following system of Bellman evaluation equations*:*

$$g = P_d g \quad \text{and} \quad h + g = L_d h. \tag{2.11}$$

Conversely, if $(g, h) \in \mathbb{R} \times \mathbb{R}^S$ is a solution to (2.11), then $g = g^\pi$ and $h = h^\pi + u$ where $u = P_d u$. Finally, if $P_d^* h = 0$ then $h = h^\pi$.

Similarly to the discounted case, there always exists an optimal policy $\pi^* \in \Pi^{\text{SD}}$ (stationary deterministic) solution to (2.8) for any $\mu_1 \in \mathcal{P}(\mathcal{S})$. Prop. 2.4 extends Prop. 2.2 to the undiscounted setting.

Proposition 2.4

Let M be a weakly communicating *MDP and denote by $\Pi^* \subseteq \Pi^{SD}$ the set of maximizers of (2.8) in Π^{SD}. If any of the following assumptions hold:*

1. *the action space A is* finite,
2. *$\Pi^* \neq \emptyset$ and $\sup_{\pi \in \Pi^*} sp\left(h^\pi\right) < +\infty$,*

then there exists a solution $(g^*, h^*) \in \mathbb{R} \times \mathbb{R}^S$ to the fixed point equation $h^* + g^* e = L h^*$. Moreover, for any such solution (g^*, h^*) and for all $s \in \mathcal{S}$,

$$g^* = \max_{\pi \in \Pi} \left\{ \liminf_{T \to +\infty} \mathbb{E}^\pi \left[\frac{1}{T} \sum_{t=1}^{T} r_t \middle| s_1 \sim s \right] \right\}.$$

Finally, any stationary policy $\pi^* = (d^*)^\infty$ satisfying $d^* \in \arg\max_{d \in D^{MR}} \{r_d + P_d h^*\}$ (i.e., greedy policy) is optimal i.e., $\pi^* \in \Pi^*$.

The proof of Prop. 2.4 is not as straightforward as the proof of Prop. 2.2 (discounted case). A complete proof of Prop. 2.4 can be found in (Puterman, 1994, Chapter 9) for finite action spaces, and (Schweitzer, 1985, Theorem 1) for compact action spaces[4]. Schweitzer (1985, Example 2) also presents a counter-example of weakly-communicating MDP for which the optimality equation does not admit any solution and $\sup_{\pi \in \Pi^*} sp\left(h^\pi\right) = +\infty$. In order to *relax* assumption 2 in Prop. 2.4, one needs to further assume that the MDP is *unichain*[5] (communicating is still not enough) as shown by Schweitzer (1985, Theorem 2). Note that the assumption that the MDP is *weakly communicating* is essential to show that the optimal gain is *state-independent* i.e., $sp\left(g^*\right) = 0$. In the general case where the MDP is *multi-chain*, the fixed point equation $h^* + g^* = L h^*$ no longer characterizes optimality i.e., other *equations* are needed (see (Puterman, 1994, Chapter 9) and (Schweitzer, 1985, Equation 1.1)). Note also

[4] Schweitzer (1985) actually proves a more general theorem from which Prop. 2.4 can be deduced.

[5] If the MDP is unichain, then assumption 2 is always satisfied and so Prop. 2.4 holds (Schweitzer, 1985, Theorem 2).

Algorithm 3 (Relative) Value Iteration

Input: Operators $L : \mathbb{R}^S \mapsto \mathbb{R}^S$ and $G : \mathbb{R}^S \mapsto D^{\text{MR}}$, accuracy $\varepsilon \in]0, r_{\max}[$, initial vector $v_0 \in \mathbb{R}^S$, arbitrary reference state $\overline{s} \in \mathcal{S}$

Output: Gain $g \in [0, r_{\max}]$, bias vector $h \in \mathbb{R}^S$ and stationary deterministic policy $\pi \in \Pi^{\text{SD}}$

1: Initialize $n = 0$
2: $v_1 := Lv_0$
3: **while** $sp\,(v_{n+1} - v_n) > \varepsilon$ **do** ▷ Loop until termination
4: Increment $n \leftarrow n + 1$
5: Shift $v_n \leftarrow v_n - v_n(\overline{s})e$ ▷ Avoids numerical instability ($v_n \not\to +\infty$)
6: $(v_{n+1}, d_n) := (Lv_n, Gv_n)$ ▷ Lv_n and Gv_n can be computed simultaneously
7: **end while**
8: Set $g := \frac{1}{2}\Big(\max\{v_{n+1} - v_n\} + \min\{v_{n+1} - v_n\}\Big)$, $h := v_n$ and $\pi := (d_n)^\infty$

that unlike Prop. 2.2, Prop. 2.4 only claims *uniqueness* of g^* but not of h^* in the optimality equation $h^* + g^* e = Lh^*$. For example, h^* can be shifted by any arbitrary constant without affecting the validity of the equation. But there may also exist other solutions that do not just differ by a constant shift (see Prop. 2.3). There is also no *strict equivalence* between *optimal* stationary policies and *greedy* policies $(d^*)^\infty$ with $d^* \in \arg\max_{d \in D^{\text{MR}}} \{r_d + P_d h^*\}$ (some optimal policies may rather satisfy an optimality equation with a different h^*, or may not even satisfy any optimal policy).

Topology of the optimal Bellman operator. In Prop. 2.5, we present few important properties of the optimal Bellman operator L that are central for the rest of the thesis. The proofs can be found in (Puterman, 1994).

Proposition 2.5

Let v and u be any two vectors in $\mathbb{R}^S$, then:

(a) L is monotone*: $v \geq u \implies Lv \geq Lu$.*

(b) L is non-expansive *both in span semi-norm and ℓ_∞-norm:*

$$sp\,(Lv - Lu) \leq sp\,(v - u) \quad \text{and} \quad \|Lv - Lu\|_\infty \leq \|v - u\|_\infty.$$

(c) L is linear[6]*: $\forall \lambda \in \mathbb{R}, \quad L(v + \lambda e) = Lv + \lambda e$.*

Computing a near optimal policy. To compute an ε-approximate solution to (2.8), we can use Alg. 3 –also known as *relative value iteration*, see Section 8.5.5 of Puterman (1994)– which is very similar to Alg. 2. Note that by definition, $sp\,(v_{n+1} - v_n) = \max\{v_{n+1} - v_n\} - \min\{v_{n+1} - v_n\}$ and so the stopping condition of Alg. 3 is comparable to the stopping condition of Alg. 2 (without involving γ). At line 5 of Alg. 3, just before computing v_{n+1}, the vector v_n is *"shifted"* by substracting the value $v_n(\overline{s})$ to $v_n(s)$ for every $s \in \mathcal{S}$ ($\overline{s}$ is an arbitrary

[6]Operator L is not a linear operator (like in linear algebra) but the property that $L(v + \lambda e) = Lv + ce$ for any $(\lambda, v) \in \mathbb{R} \times \mathbb{R}^S$ is often called the "linearity" property of L.

"reference" state). This is because the optimal Bellman operator L is not a contraction w.r.t. any-norm –unlike L_γ– and in general v_n asymptotically grows as ng^*e when $n \to \infty$ (see Section 8.2.1 of Puterman (1994)). Since in most MDPs $g^* > 0$, this means that $v_n \to +\infty$ as $n \to +\infty$, potentially causing *numerical instabilities*. However, under the conditions of Prop. 2.6 below, v_n will converge in span semi-norm i.e., will converge in the *quotient space* induced by the semi-norm $sp(\cdot)$ on $\mathbb{R}^S$. This is related to the remark we made earlier about h^* being defined up to a constant shift in the optimality equation (h^* is uniquely defined in the quotient space when there is exists a single optimal policy for example). Shifting v_n before any new update ensures convergence in the original space $\mathbb{R}^S$ (convergence in ℓ_∞-norm as opposed to span semi-norm). Note that neither the stopping condition (line 3 of Alg. 3) nor the other outputs g and π of Alg. 3 are affected by the shift of line 5. If line 5 was removed, Alg. 3 would stop after the same number of iterations and would return the same gain g and policy π. Only the final bias h as well as all the intermediate vectors v_n are shifted by a big constant (which grows linearly with n). Indeed, the difference $Lv_n - v_n$ remains unchanged after a constant shift in v_n: for all $c \in \mathbb{R}$, $L(v_n + ce) - (v_n + ce) = Lv_n - v_n$ due to the *linearity* property of the optimal Bellman operator (Prop. 2.5 (c)). Prop. 2.6 and Lem. 2.7 below also hold if line 5 of Alg. 3 (constant shift) is removed (except that in this case v_n diverges in $\mathbb{R}^S$ and converges only in the quotient space, as explained above). These results hold both for MDPs with finite and compact action spaces.

Proposition 2.6 (Theorems 9.4.5 of Puterman (1994) adapted by Jaksch et al. (2010))
Consider the sequences of vectors $(v_n)_{n\in\mathbb{N}}$ and Markov decision rules $(d_n)_{n\in\mathbb{N}}$ obtained while executing Alg. 3. If Prop. 2.4 holds and either:

1. *every average optimal stationary deterministic policy has an* aperiodic *transition matrix,*
2. *or the transition matrices P_{d_n} are* aperiodic *for all $n \geq 1$,*

then there exists $h^* \in \mathbb{R}^S$ such that $\lim_{n\to+\infty} v_n = h^*$ and $Lh^* = h^* + g^*e$.

Proof. In his Section 9.4.1, Puterman (1994) provides a complete proof in the general *multi-chain* case with *finite* action space and when every average optimal stationary deterministic policy has an aperiodic transition matrix (assumption 1). However, the proof only uses the existence of a solution of the Bellman optimality equation, which is always guaranteed under the assumptions of Prop. 2.4. Only his Lemma 9.4.3 uses the finiteness of D^{MD} and Π^{SD} but this lemma trivially holds when M is weakly communicating (instead of just multi-chain). Therefore, the result also holds for compact action spaces as long as all the assumptions of Prop. 2.4 are satisfied. While Puterman (1994) only provides a proof in the case where every average optimal stationary deterministic policy has an aperiodic transition matrix (assumption 1), Jaksch et al. (2010, Appendix B) showed how to extend it to the case where the transition matrices P_{d_n} are aperiodic for all $n \geq 1$ (assumption 2). ■

Since $sp(g^*e) = 0$ and $sp()$ is a continuous function (as a semi-norm), when the assumptions of Prop. 2.6 hold the stopping condition of Alg. 3 is necessarily met after a *finite* number of iterations. Moreover, it is possible to characterize by how much the gain g returned by Alg. 3

differs from g^*.

Proposition 2.7

Consider the gain g and bias h returned by Alg. 3. Under the same assumptions as Prop. 2.6, $|g - g^*| \leq \varepsilon/2$ and for all $s \in \mathcal{S}$, $|Lh(s) - h(s) - g| \leq \varepsilon$, where $\varepsilon \in]0, r_{\max}[$ is the accuracy given as input of Alg. 3.

Proof. The fact that $|g - g^*| \leq \varepsilon/2$ is just the application of Theorem 8.5.6 and Corollary 9.4.6 of Puterman (1994) (see also Section 9.5). For the other inequalities, we introduce the quantities $\mathfrak{M} := \max\{Lh - h\}$ and $\mathfrak{m} := \min\{Lh - h\}$. The condition $sp\,(Lh - h) \leq \varepsilon$ (line 3 of Alg. 3) is equivalent to $\mathfrak{M} - \mathfrak{m} \leq \varepsilon$. Using inequality $|g - g^*| \leq \varepsilon/2$ and the definition of g (line 8 of Alg. 3) we deduce

$$\frac{1}{2}(\mathfrak{M} + \mathfrak{m}) \geq g^* - \frac{\varepsilon}{2} \implies \mathfrak{m} \geq g^* - \frac{\varepsilon}{2} - \frac{1}{2}(\mathfrak{M} - \mathfrak{m}) \geq g^* - \varepsilon$$
$$\frac{1}{2}(\mathfrak{M} + \mathfrak{m}) - g^* \leq \frac{\varepsilon}{2} \implies \mathfrak{M} \leq g^* + \frac{\varepsilon}{2} + \frac{1}{2}(\mathfrak{M} - \mathfrak{m}) \leq g^* + \varepsilon.$$

In conclusion, for all $s \in \mathcal{S}$, $g^* - \varepsilon \leq \mathfrak{m} \leq Lh(s) - h(s) \leq \mathfrak{M} \leq g^* + \varepsilon$ which concludes the proof. ■

Prop. 2.7 states that not only g is an ε-approximation of g^* but $(g, h) \in \mathbb{R} \times \mathbb{R}^S$ approximately satisfies the Bellman optimality equation as $\|Lh - h - ge\|_\infty \leq \varepsilon$. The condition that P_{d_n} is aperiodic for all $n \geq 1$ is not always satisfied. Fortunately, there is a way to modify the transition probabilities of the MDP to enforce this property while impacting neither the optimal gain g^* nor the stationary optimal policy(ies) π^*. This modification is called the *aperiodicity transformation* (Puterman, 1994, Section 8.5.4).

Aperiodicity transformation. Instead of applying Alg. 3 to the original MDP M, we first construct a *transformed MDP* M_α where $\alpha \in]0, 1]$. M_α is similar to M with the only difference that for all Markov decision rules $d \in D^{\text{MR}}$, the transition matrix P_d is transformed into $P_d^\alpha := \alpha P_d + (1 - \alpha)I$ where I is the $S \times S$ identity matrix. We first note that if M is weakly-communicating, so is M_α as long as $1 \geq \alpha > 0$ (more generally, the aperiodicity transformation does not change the chain structure of the MDP). As shown by Puterman (1994, Proposition 8.5.8), this transformation does not affect the gain of any stationary policy meaning that for any $\pi \in \Pi^{\text{SR}}$, $g_\alpha^\pi = g^\pi$.[7] We denote by L_α the optimal Bellman operator of M_α. We note that:

$$\forall v \in \mathbb{R}^S,\ L_\alpha v := \max_{d \in D^{\text{MD}}} \{r_d + \alpha P_d^\alpha v\} + (1 - \alpha)v. \tag{2.12}$$

For $\alpha \in]0, 1[$, all the transition matrices of M_α are aperiodic and so Prop. 2.6 and Lem. 2.7 apply. If g_α and h_α denote the gain and bias returned by Alg. 3 applied to M_α, and if $(v_n^\alpha)_{n \in \mathbb{N}}$

[7] The transformation introduced by (Puterman, 1994, Section 8.5.4) is slightly different as the rewards are all multiplied by α. Therefore, Proposition 8.5.8 of Puterman (1994) states that the gain is also multiplied by α i.e., $g_\alpha^\pi = \alpha \cdot g^\pi$. However, it is straightforward to adapt the proof of Proposition 8.5.8 of Puterman (1994) to our case.

is the sequence of vectors obtained while executing the algorithm, then $\lim_{n\to+\infty}(v_{n+1}^{\alpha}-v_n^{\alpha}) = g^*$ (since $g_\alpha^* = g^*$), $|g_\alpha - g^*| \leq \varepsilon/2$ and $\|Lh_\alpha - h_\alpha - g_\alpha e\|_\infty \leq \varepsilon$. Note that this holds for any $\alpha \in]0,1[$ but in practice, the closer α is to 0, the *slower the convergence* of value iteration (more iterations are needed to meet the stopping condition of line 3).

Episodic problems. To conclude this section, we highlight the connection between the undiscounted inifinite horizon setting and the *episodic setting*. It is very common in practice that an RL task ends as soon as a certain *termination condition* is met, after which the problem is *reset* to an initial state (or initial distribution over states). Each reset defines a new *"episode"*. The restart condition is often assumed to be Markovian i.e., to depend only on the current state and action. The goal is then to maximize the cumulative reward over episodes. If the restart condition satisfies the Markov property, it can simply be interpreted as a transition probability of the MDP, in which case the gain is a good optimality criterion. Actually, in the episodic setting, it is a well-known result of *renewal theory* that the gain $g^\pi(s)$ of a policy π starting in state $s \in \mathcal{S}$ is equal to the ratio $\mathbb{E}^\pi[R|s_1 = s]/\mathbb{E}^\pi[\tau|s_1 = s]$, where R and τ denote respectively the total reward accumulated during an episode and the total duration of the episode. It seems reasonable that we should not just aim at maximizing $\mathbb{E}^\pi[R|s_1 = s]$, but we should also take into account $\mathbb{E}^\pi[\tau|s_1 = s]$. Indeed, it might sometimes be more rewarding *on the long-term* to run short episodes with relatively small cumulative reward rather than episodes with high reward but extremely long duration.

Refined optimality (Bias and Blackwell optimality)

In many MDPs, there is not a *single* gain-optimal policy although it is clear that among the gain-optimal policies, some are preferable in terms of reward. For example, while two different policies may have the same asymptotic per-step reward, one of them may accumulate more reward while converging to the asymptotic gain. It turns out that this is formally described by the notion of *bias optimality* (Lewis and Puterman, 2002) which refines gain optimality (bias optimal implies gain optimal but not conversely). Bias optimality can be further refined by the notions of *sensitive discount optimality* and *Blackwell optimality* which provides a comprehensive understanding of infinite horizon problems in the absence of a discount factor. All these refinements go beyond the scope of this thesis and from now on we will restrict attention to gain optimality.

2.1.4 Stochastic shortest path

In this section we review some important results on the *stochastic shortest path problem* (Bertsekas, 1995, Chapter 2). These results will be extremely useful to understand how difficult it is for an agent to *navigate* between the states of an MDP. Unlike in previous sections, we assume that the rewards of the MDP are all *non-positive* and lie in $[-r_{\max}, 0]$. When action a is played in state s, the absolute value of the reward $|r(s,a)|$ should be interpreted as the *expected time* before reaching the next state in the MDP. $|r(s,a)|$ can be seen as the *"length"* or expected *"duration"* of a transition (which only depends on the current

state s and action a, and not on the next state). In the stochastic shortest path problem, we consider an agent travelling from a state x to a state s. The *total length* of a *sampled path* $(s_1 = x, a_1, r_1, \dots, s_\tau = s)$ is defined as $|\sum_{t=1}^{\tau} r_t| = \sum_{t=1}^{\tau} |r_t| = -\sum_{t=1}^{\tau} r_t$. We introduce the following definition:

Definition 2.4

For any state $s \in \mathcal{S}$, we define $\tau(s) := \inf\{t \geq 1 : s_t = s\}$ the first hitting *time of s. Note that $\tau(s) \in \mathbb{N} \cup \{+\infty\}$.*

The goal of the stochastic shortest path problem is to find the *shortest expected distance* between states x and s in the MDP i.e., to solve

$$\inf_{\pi \in \Pi} \left\{ \mathbb{E}^\pi \left[\sum_{t=1}^{\tau(s)-1} |r_t| \middle| s_1 = x \right] \right\} \quad \Leftrightarrow \quad \sup_{\pi \in \Pi} \left\{ \mathbb{E}^\pi \left[\sum_{t=1}^{\tau(s)-1} r_t \middle| s_1 = x \right] \right\}. \tag{2.13}$$

Although the optimization problem in Eq. 2.13 seems very different from the optimization problem in Eq. 2.8, the two problems are related through the Bellman optimality equation. The stochastic shortest path problem can somehow be interpreted as a specific case of finding a *bias-optimal* policy when the optimal gain g^* is 0. The optimality equation can then be written as $Lh^* = h^*$. This statement is made more formal in Prop. 2.8 below. For all pairs of states $(x, s) \in \mathcal{S} \times \mathcal{S}$, the value of the supremum in (2.13) (right-hand side) is denoted $h^*_{\mapsto s}(x)$. By definition, $h^*_{\mapsto s}(x) \leq 0$ for all $x \in \mathcal{S}$ and $h^*_{\mapsto s}(s) = 0$.

Proposition 2.8

Let $M = \{\mathcal{S}, \mathcal{A}, r, p\}$ be a communicating *MDP (finite or compact $\mathcal{A}$) with* negative rewards *$r(s,a) \in [-r_{\max}, 0]$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. For any state $s \in \mathcal{S}$, consider the* Bellman shortest path operator *$L_{\mapsto s} : \mathbb{R}^S \longmapsto \mathbb{R}^S$ defined for all $v \in \mathbb{R}^S$ as:*

$$\forall x \in \mathcal{S}, \; L_{\mapsto s} v(x) := \begin{cases} \max_{a \in \mathcal{A}_x} \left\{ r(x,a) + \sum_{y \in \mathcal{S}} p(y|x,a) v(y) \right\} & \text{if } x \neq s \\ v(s) & \text{otherwise} \end{cases}. \tag{2.14}$$

$h^*_{\mapsto s}$ is the (componentwise) maximal non-positive solution *of the* Bellman shortest path optimality equation *$L_{\mapsto s} h^*_{\mapsto s} = h^*_{\mapsto s}$. Moreover, if $d^*_{\mapsto s}$ is a* greedy *decision rule w.r.t. $h^*_{\mapsto s}$ i.e., $d^*_{\mapsto s}(x) \in \arg\max_{a \in \mathcal{A}_x} \left\{ r(x,a) + \sum_{y \in \mathcal{S}} p(y|x,a) h^*_{\mapsto s}(y) \right\}$ for all $x \neq s$, then $\pi^*_{\mapsto s} := (d^*_{\mapsto s})^\infty$ is an optimal solution to Eq. 2.13.*

Proof. $L_{\mapsto s}$ corresponds to the optimal Bellman operator of a modified MDP $M_{\mapsto s}$ where all actions are unchanged except the actions in state s. These actions are assigned a reward 0 i.e., $r(s,a) = 0$ for all $a \in \mathcal{A}_s$, and loop on s with probability 1 i.e., $p(s|s,a) = 1$ for all $a \in \mathcal{A}_s$. In $M_{\mapsto s}$, problem (2.13) can be equivalently formulated with $\tau(s)$ replaced by $+\infty$ (the reward is always zero once state s is reached). Therefore, (2.13) is an instance of an *expected total-reward problem with negative model* (Puterman, 1994, Section 7.3). Since M is communicating, there exists a policy π such that $\mathbb{E}^\pi \left[\sum_{t=1}^{+\infty} r_t \middle| s_1 = x \right] > -\infty$ in $M_{\mapsto s}$ (e.g., any policy reaching s in finite time almost surely) and so Assumption 7.3.1. of Puterman (1994) holds. The fact that $h^*_{\mapsto s}$ is the maximal non-positive solution of $L_{\mapsto s} h^*_{\mapsto s} = h^*_{\mapsto s}$ is a

consequence of Theorem 7.3.3. (a) of Puterman (1994) (proved for both finite and compact action spaces). The fact that $\pi^*_{\mapsto s}$ is optimal is a consequence of Theorem 7.3.5 of Puterman (1994). ■

Value iteration (Alg. 3) converges to $h^*_{\mapsto s}$ (for both finite and compact $\mathcal{A}$) but no aperiodicity condition is needed in this case.

Proposition 2.9

Let MDP M satisfy the assumptions of Prop. 2.8. If Alg. 3 is run with operator $L_{\mapsto s}$, $v_0 := 0$ and reference state $\overline{s} := s$, then v_n converges monotonically *to $h^*_{\mapsto s}$ and so Alg. 3 stops after a finite number of iterations. Moreover, the vector h output by Alg. 3 satisfies $-\varepsilon e \leq L_{\mapsto s}h - h \leq 0$.*

Proof. Since the reference state is s and $v_0 = 0$, by induction $v_n(s) = 0$ for all $n \geq 0$ so that line 5 of Alg. 3 (constant shift) can be ignored i.e., $v_n = L^n 0$. Then, the monotone convergence of $(v_n)_{n\in\mathbb{N}}$ is a direct consequence of Theorem 7.3.10. (a) of Puterman (1994). Therefore, $v_0 = 0 \geq v_1 \geq ... \geq h \geq L_{\mapsto s}h \geq h^*_{\mapsto s}$ (first inequality). When Alg. 3 terminates, we have $sp\,(L_{\mapsto s}h - h) \leq \varepsilon$. We introduce the quantities $\mathfrak{M} := \max\{L_{\mapsto s}h - h\}$ and $\mathfrak{m} := \min\{L_{\mapsto s}h - h\}$ so that $sp\,(L_{\mapsto s}h - h) = \mathfrak{M} - \mathfrak{m} \leq \varepsilon$. Since $v_0 = 0$ and $L_{\mapsto s}v(s) = v(s)$ for all $v \in \mathbb{R}^S$ by definition, $v_n(s) = 0$ for all $n \geq 0$ and so $L_{\mapsto s}h(s) = h(s) = 0$ and $\mathfrak{M} = 0$. The condition $sp\,(L_{\mapsto s}h - h) \leq \varepsilon$ implies $L_{\mapsto s}h - h \geq \mathfrak{m}e \geq -\varepsilon e$. ■

Bias and aperiodicity transformation. We already showed that the aperiodicity transformation (Sec. 2.1.3) does not affect the gain, we will now investigate the impact on the shortest path. Although such a transformation is not needed to enforce convergence of value iteration in a stochastic shortest path setting, Thm. 2.1 (below) will later be useful in this thesis.

Theorem 2.1

Let MDP M satisfy the assumptions of Prop. 2.8. Let $\alpha \in]0, 1]$ and M_α be the MDP obtained after applying the aperiodicity transformation of parameter α to M. M_α also satisfies the assumptions of Prop. 2.8 and so $h^{\alpha *}_{\mapsto s}$ is well-defined for all $s \in \mathcal{S}$. Moreover, $\alpha \cdot h^{\alpha *}_{\mapsto s} = h^*_{\mapsto s}$.

Proof. One way to interpret the aperiodicity transformation is that at every time step, an agent evolving in M_α "loops" on the current state with probability $1 - \alpha$, and follows the dynamics of M with probability α. Therefore, all the paths that exist in M also exist in M_α but they are "longer". So if M is communicating, M_α is communicating as well. The rewards are not affected by the transformation so if the rewards of M are non-positive, so are the rewards of M_α. Furthermore, by definition, $h^{\alpha *}_{\mapsto s}$ is a fixed point of $L^\alpha_{\mapsto s}$ and $h^*_{\mapsto s}$ is a fixed

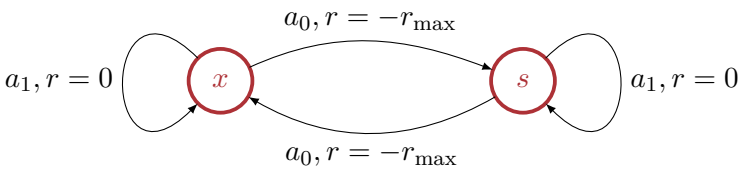


Figure 2.2: Example of communicating MDP where the "shortest path" from x to s (2.13) is such that $\pi^*(x) = a_1$ and $h^*_{\mapsto s}(x) = 0$. Under π^*, $\tau(s) = +\infty$ almost surely.

point of $L_{\mapsto s}$. Let's denote by $p_{\mapsto s}$ the transition probability of $M_{\mapsto s}$ (see proof of Prop. 2.8).

$$
\begin{aligned}
L^{\alpha}_{\mapsto s} h^{\alpha *}_{\mapsto s} = h^{\alpha *}_{\mapsto s} \Leftrightarrow & \max_{a \in \mathcal{A}_x} \left\{ r(x,a) + \alpha \sum_y p_{\mapsto s}(y|x,a) h^{\alpha *}_{\mapsto s}(y) \right\} + (1-\alpha) h^{\alpha *}_{\mapsto s}(x) = \cancel{h^{\alpha *}_{\mapsto s}}(x) \\
\Leftrightarrow & \max_{a \in \mathcal{A}_x} \left\{ r(x,a) + \sum_y p_{\mapsto s}(y|x,a) \left(\alpha h^{\alpha *}_{\mapsto s}(y) \right) \right\} = \alpha h^{\alpha *}_{\mapsto s}(x) \\
\Leftrightarrow & L_{\mapsto s} \left(\alpha h^{\alpha *}_{\mapsto s} \right) = \alpha h^{\alpha *}_{\mapsto s}.
\end{aligned}
$$

So $\alpha h^{\alpha *}_{\mapsto s}$ is a fixed point of $L_{\mapsto s}$ and conversely $h^*_{\mapsto s}/\alpha$ is a fixed point of $L^{\alpha}_{\mapsto s}$. Moreover, $\alpha > 0$, $h^{\alpha *}_{\mapsto s} \leq 0$ and $h^*_{\mapsto s} \leq 0$ implying that $\alpha h^{\alpha *}_{\mapsto s} \leq 0$ and $h^*_{\mapsto s}/\alpha \leq 0$. Since $h^*_{\mapsto s}$ is the maximum non-positive fixed point of $L_{\mapsto s}$ and $\alpha h^{\alpha *}_{\mapsto s} \leq 0$, necessarily $h^*_{\mapsto s} \geq \alpha h^{\alpha *}_{\mapsto s}$. Symmetrically, $h^{\alpha *}_{\mapsto s}$ is the maximum non-positive fixed point of $L^{\alpha}_{\mapsto s}$ and $h^*_{\mapsto s}/\alpha \leq 0$ so necessarily $h^{\alpha *}_{\mapsto s} \geq h^*_{\mapsto s}/\alpha$. In conclusion, $\alpha h^{\alpha *}_{\mapsto s} = h^*_{\mapsto s}$. ■

Infinite hitting time. In this section we considered a slightly more general formulation of the shortest path problem than Bertsekas (1995, Chapter 2). In our formulation, it is possible that the policy π^* achieving the maximum in (2.13) satisfies $\mathbb{E}^{\pi^*}[\tau(s)|s_1 = x] = +\infty$, while the maximum in (2.13) is always bounded (under the assumption that M is communicating). In this case, the solution of (2.13) does not exactly match the intuitive notion that we have of a "shortest path" to a target state s. We give an example of such a scenario in Fig. 2.2. Nevertheless, all the results presented in this section hold whether $\tau(s)$ is almost surely *finite* or not. This is because the problem can be expressed as as a specific instance of *expected total-reward problem with negative model* (Puterman, 1994, Section 7.3) (see proof of Prop. 2.8). Note that if all the rewards are *strictly negative* (as opposed to just non-positive), then necessarily $\mathbb{E}^{\pi^*}[\tau(s)|s_1 = x] < +\infty$ and the solution of the problem is a *"proper"* shortest path (this is the specific case analysed in Bertsekas (1995, Chapter 2)).

2.1.5 Uncertain MDPs: between discrete and continuous MDPs

In this thesis, we will have to deal with MDPs with unknown r and p but for which we know some confidence sets. A convenient way to describe an *uncertain MDP* is through the notions of *"bounded-parameter MDPs"* and *"extended MDPs"*.

Bounded-parameter MDP. A bounded-parameter MDP is a collection of MDPs –with identical state-action spaces– specified by *confidence bounds* on the parameters (rewards and transition probabilities) representing the uncertainty about the true values. Formally, a bounded parameter MDP $\mathcal{M}$ is usually characterized by some compact sets $B_r(s,a) \subseteq [0, r_{\max}]$ and $B_p(s,a) \subseteq \Delta_S$ (see Def. 2.1):

$$\mathcal{M} = \Big\{ M = \langle \mathcal{S}, \mathcal{A}, r, p \rangle : \; r(s,a) \in B_r(s,a), \; p(\cdot|s,a) \in B_p(s,a), \; \forall (s,a) \in \mathcal{S} \times \mathcal{A} \Big\}. \quad (2.15)$$

Bounded-parameter MDPs were first introduced by Givan et al. (2000) in the infinite horizon discounted setting, and later used by Tewari and Bartlett (2007a) in the undiscounted setting. The bounded parameter MDP will typically be constructed so as to *include* the true MDP with high probability (w.h.p.).

Extended MDP. As pointed out by Jaksch et al. (2010, Section 3.1.1), any bounded parameter MDP can be equivalently represented by an *"extended MDP"*. The idea is to combine all MDPs into a single MDP with identical state space $\mathcal{S}$ but with an *extended compact action space* $\mathcal{A}^+$. The extended MDP corresponding to the bounded parameter MDP $\mathcal{M}$ defined in Eq. 2.15 is formally defined as $\mathcal{M}^+ = \langle \mathcal{S}, \mathcal{A}^+, r^+, p^+ \rangle$ where for all $s \in \mathcal{S}$:

$$\begin{aligned} \mathcal{A}^+_s &:= \bigcup_{a \in \mathcal{A}_s} \{a\} \times B_r(s,a) \times B_p(s,a) \\ \forall a^+ = (a, r, p) \in \mathcal{A}^+_s, \; &\begin{cases} r^+(s, a^+) := r \\ p^+(\cdot|s, a^+) := p \end{cases} . \end{aligned} \quad (2.16)$$

Every possible value in the compact sets $B_r(s,a)$ and $B_p(s,a)$ is considered as an "extended" action in $\mathcal{M}^+$. For any MDP $M = \langle \mathcal{S}, \mathcal{A}, r, p \rangle \in \mathcal{M}$ and any stationary deterministic policy $\pi \in \Pi^{\text{SD}}_M$ defined on M, let's define the stationary deterministic policy $\pi^+ \in \Pi^{\text{SD}}_{\mathcal{M}^+}$ on $\mathcal{M}^+$ by $\pi^+(s) := (\pi(s), r(s, \pi(s)), p(\cdot|s, \pi(s)))$. It is immediate to see that the *Markov Reward Processes* (MRP) induced by π on M is exactly the same as the MRP induced by π^+ on $\mathcal{M}^+$. Conversely, for any policy $\pi^+ \in \Pi^{\text{SD}}_{\mathcal{M}^+}$, the MDP $M = \langle \mathcal{S}, \mathcal{A}, r, p \rangle \in \mathcal{M}$ and policy $\pi \in \Pi^{\text{SD}}_M$ defined as follows induce the same MRP as π^+:

$$\forall s \in \mathcal{S}, \begin{cases} \pi(s) := a \\ r(s,a) := r \\ p(\cdot|s,a) := p \end{cases} \quad \text{where } (a, r, p) := \pi^+(s), \text{ and } \forall b \neq a \begin{cases} r(s,b) \in B_r(s,b) \text{ (any value)} \\ p(\cdot|s,b) \in B_p(s,b) \text{ (any value)} \end{cases}$$

There is a *one-to-one correspondence* between the pairs $(M, \pi) \in \mathcal{M} \times \Pi^{\text{SD}}_M$ and the policies $\pi^+ \in \Pi^{\text{SD}}_{\mathcal{M}^+}$. In the rest of the thesis, we will use the *same notation* $\mathcal{M}$ for an extended MDP (2.16) and the corresponding bounded-parameter MDP (2.15) (they are essentially the "same" mathematical object). We will also slightly abuse terminology and say that an MDP *"belongs to"* an extended MDP when it is actually contained in the corresponding bounded-parameter MDP.

Extended optimal Bellman operator. The optimal Bellman operator $\mathcal{L}$ of an extended MDP is called an *"extended optimal Bellman operator"* and is defined as:

$$\forall v \in \mathbb{R}^S,\ \forall s \in \mathcal{S},\ \mathcal{L}v(s) := \max_{a \in \mathcal{A}_s} \left\{ \max_{r \in B_r(s,a)} r + \max_{p \in B_p(s,a)} p^\intercal v \right\} \tag{2.17}$$

In the specific case where the confidence sets $B_p(s,a)$ are *polytopes*, the inner maximum $\max_{p \in B_p(s,a)} \{p^\intercal v\}$ is reached on at least one *vertex*[8], meaning that we can *restrict* $B_p(s,a)$ to its vertices without impacting the result (there are only finitely many vertices on a polytope). Moreover, $\max_{r \in B_r(s,a)} \{r\}$ is always reached on the maximal value of $B_r(s,a)$ and so it can be replaced by a singleton without changing anything. In conclusion, $\mathcal{L}$ can be expressed as an optimal Bellman operator with *finite action space*. In this thesis, all the extended optimal Bellman operators that we will deal with will satisfy satisfy this property. This simplifies a lot the thoretical analysis (see Prop. 2.4 and 2.6).

2.2 On-line Reinforcement Learning in the infinite horizon undiscounted setting

In the previous section, we used the formalism of MDPs to describe an agent interacting with its environment. Depending on the chosen optimality criterion, we showed how to compute a (near-)optimal policy when the parameters of the MDP are fully *known*. In this section, we will address the case when all or part of the MDP is *unknown* and needs to be *learned* by the agent. We restrict attention to the *infinite horizon undiscounted setting* which will be the main focus of this thesis. Although it is not always the most appropriate setting (e.g., when there is a pre-defined horizon or discount factor), it is perhaps the most general (in the limit, see Sec. 2.2) among all the settings presented in Sec. 2.1. It is also the most challenging to analyse.

2.2.1 The learning problem

We consider the learning problem where $\mathcal{S}$, $\mathcal{A}$ and $r_{\max}$ are *known*, while rewards r and transition probabilities p are *unknown* and need to be estimated *on-line* i.e., in a *sequential* fashion. The *planning* algorithms presented in Sec. 2.1 cannot be used directly to compute an optimal policy and *samples* of r and p need to be collected first.

Rather than focusing on learning a (near-)optimal policy (e.g., with the best possible accuracy given an horizon T), we will be interested in maximizing the cumulative reward $\sum_{t=1}^{T} r_t$ collected up to time T. As T grows to infinity, maximizing $\sum_{t=1}^{T} r_t$ amounts to learning a gain-optimal policy since *in the limit* the series eventually grows as Tg^* (Puterman, 1994, Chapter 8), which is the best asymptotic growth rate achievable. But in the meanwhile, the learning agent needs to efficiently trade-off the *exploration* needed to collect information about the dynamics and reward, and the *exploitation* of the experience gathered so far to

[8] This is a well-known property of *linear programs*.

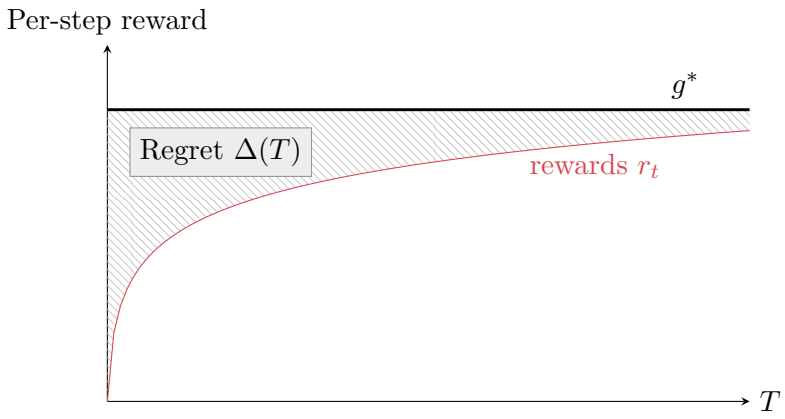


Figure 2.3: Graphical illustration of Def. 2.5.

gain as much reward as possible. In order to quantitatively assess the *exploration-exploitation* performance we use the concept of *regret* which compares the rewards accumulated by the agent and an optimal policy i.e., $\mu_1^\intercal v_T^* - \sum_{t=1}^T r_t$. To simplify this definition, we observe that

$$v_T^* = L^T 0 = L^T h^* + L^T 0 - L^T h^* = T g^* e + h^* + L^T 0 - L^T h^*.$$

Using the fact that L is non-expansive in ℓ_∞-norm (property (b) of Prop. 2.5) we obtain

$$\|v_T^* - T g^* e\|_\infty \leq \|h^*\|_\infty + \|L^T 0 - L^T h^*\|_\infty \leq 2\|h^*\|_\infty.$$

h^* is independent of T and measures the expected cumulative difference between the optimal asymptotic stationary regime g^* and the actual reward at time step t. It somehow quantifies the unavoidable expected regret incurred when the optimal policy is executed starting from a distribution different than the optimal asymptotic regime. We therefore introduce the following definition.

Definition 2.5

Let $(r_t)_{t\geq 1}$ denote the sequence of rewards collected while executing learning algorithm $\mathfrak{A}$ in MDP M, with initial state distribution μ_1. The regret after T time steps is defined as

$$\Delta(M, \mathfrak{A}, \mu_1, T) := \sum_{t=1}^{T} (g^* - r_t) = T g^* - \sum_{t=1}^{T} r_t.$$

Graphically, the regret corresponds to the hatched area between the black and red curves on Fig. 2.3. Given that the term Tg^* is *algorithm-independent*, maximizing $\sum_{t=1}^T r_t$ is equivalent to minimizing the regret.

Since the regret is a *random variable*, we cannot minimize it directly. One possibility is to analyse the *expected regret* $\mathbb{E}^{\mathfrak{A}}\left[\Delta(M, \mathfrak{A}, \mu_1, T) | s_1 \sim \mu_1\right]$, whre $\mathfrak{A} \in \Pi$ is interpreted as an (priori non-stationary) policy. Another possibility is to bound the regret in *high probability* i.e., with probability $1 - \delta$ where δ is a level of *confidence* given as input to $\mathfrak{A}$. A high probability bound is usually considered a *stronger* result: it is always possible to convert a high probability bound into a bound on expectation by carefully tuning the confidence δ.

The analysis of the regret in expectation and in high probability both belong to the *frequentist* approach: the result gives an indication of what happens if the learning process is *repeated* several times in the same conditions with different random samplings (different "seeds"). Another line of research consists in analysing the *expected Bayesian regret*. With this approach, the true unkown MDP is assumed to be sampled from a known *prior distribution* and the goal is to minimize $\mathbb{E}_M \left[\mathbb{E}^{\mathfrak{A}} \left[\Delta(M, \mathfrak{A}, \mu_1, T) | s_1 \sim \mu_1 \right] \right]$ where $\mathbb{E}_M$ is the expectation over the prior. Bayesian regret bounds provide weaker guarantees since they only hold on expectation over a set of plausible MDPs, and not always for a specific instance. In this thesis, we will exclusively focus on frequentist approaches, mainly high probability regret bounds.

2.2.2 Theoretical benchmarks

We say that an algorithm *learns* if and only if $\Delta(M, \mathfrak{A}, \mu_1, T) = o(T)$ when $T \to +\infty$ (either in expectation or with high probability). But we also care about how *fast* the algorithm can learn. Before describing learning algorithms and analysing their regret, we first discuss fundamental *limitations* of the learning abilities of *any* algorithm. We summarize several existing *regret lower-bounds* which provide insightful benchmarks when designing algorithms.

Asymptotic lower-bounds

The first regret lower-bound in an RL setting was proved by Burnetas and Katehakis (1997). The lower-bound is proved for the restricted family of *ergodic MDPs*. In such MDPs, the optimal bias h^* is *unique* up to a constant shift (Lewis and Puterman, 2002, Proposition 2.3) and all gain-optimal stationary deterministic policies are greedy w.r.t. Lh^* i.e., the Bellman optimality equation fully characterizes gain-optimal policies (Lewis et al., 1999). We define $\Pi^* \subseteq \Pi^{\mathrm{SD}}$ the set of such greedy policies i.e., the set of all stationary deterministic gain-optimal policies in M. Because h^* is unique up to constant shift, for any ergodic MDP $M = \langle \mathcal{S}, \mathcal{A}, r, p \rangle$ we can define the *state-action gaps* for all state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$ without any ambiguity

$$\delta(s,a) := \underbrace{\max_{b \in \mathcal{A}_s} \{r(s,b) + p(\cdot|s,b)^\intercal h^*\}}_{=Lh^*(s)=h^*(s)+g^*} - r(s,a) - p(\cdot|s,a)^\intercal h^*, \tag{2.18}$$

where h^* is any optimal bias of M. Burnetas and Katehakis (1997) assume that the reward function is known and only the transition probabilities need to be learned. They also define the *state-action KL divergences* between two MDPs M and M' that differ only by their dynamics p and p'

$$\mathrm{KL}_{M \| M'}(s,a) := \mathrm{KL}\left(p(\cdot|s,a) \| p'(\cdot|s,a)\right). \tag{2.19}$$

Finally, the sets of *confusing models* w.r.t. M are defined as

$$\Phi(s,a) := \Big\{ M' = \langle \mathcal{S}, \mathcal{A}, r, p' \rangle : \; p'(\cdot|x,b) = p(\cdot|x,b) \text{ for all } (x,b) \neq (s,a), \\ \delta(s,a) > 0 \text{ and } \delta'(s,a) = 0 \Big\}. \tag{2.20}$$

We report the lower-bound of Burnetas and Katehakis (1997) in Prop. 2.10 below.

Proposition 2.10 (Theorem 1 of Burnetas and Katehakis (1997))

Let M be an ergodic *MDP with finite state and action spaces $\mathcal{S}$ and $\mathcal{A}$, and $r_{\max} = 1$. Let $\mathfrak{A}$ be a learning algorithm s.t. $\mathbb{E}^{\mathfrak{A}}\left[\Delta(M', \mathfrak{A}, \mu_1, T) | s_1 \sim \mu_1\right] = o(T^\alpha)$ for all $\alpha > 0$, all ergodic MDPs M' and initial state distribution μ_1. The expected regret of $\mathfrak{A}$ is lower bounded as*

$$\liminf_{T\to\infty} \frac{\mathbb{E}^{\mathfrak{A}}\left[\Delta(M, \mathfrak{A}, \mu_1, T) | s_1 \sim \mu_1\right]}{\log T} \geq \sum_{\substack{s,a \\ \Phi(s,a)\neq\emptyset}} \frac{\delta(s,a)}{\inf_{M'\in\Phi(s,a)} KL_{M\|M'}(s,a)}.$$

In Prop. 2.10, the learning algorithm $\mathfrak{A}$ is assumed to be *uniformly good* i.e., to achieve *sub-polynomial* regret on all ergodic MDPs. Since $\mathfrak{A}$ is constrained to perform well on *all* instances, it cannot perform arbitrarily well on any specific instance, hence the lower-bound. This is reminiscent of the *"No Free Lunch"* theorem in supervised learning. Prop. 2.10 shows that the expected regret will eventually grow at least logarithmically with time.

The more sub-optimal sate-action (s,a) (i.e., the higher $\delta(s,a)$), the bigger the lower-bound: the regret incurred when playing this action is higher by definition. When the transition probability vector $p(\cdot|s,a)$ associated with a sub-optimal action a can easily be *confused* with another probability vector q that makes a optimal, the lower bound is also bigger (term $\mathrm{KL}_{M\|M'}(s,a)$). This is because a small error in the estimation of $p(\cdot|s,a)$ can lead to a potientially very sub-optimal behaviour. As shown by Ok et al. (2018, Section 4.1), the lower-bound of Prop. 2.10 can be upper-bounded by $\frac{2SA(sp(h^*)+1)^2}{\delta_{\min}}$ with the minimum gap

$$\delta_{\min} := \min_{\substack{s,a: \\ \delta(s,a)>0}} \delta(s,a). \tag{2.21}$$

$\delta_{\min} > 0$ except if $\Pi^{\mathrm{SD}} = \Pi^*$.

Ok et al. (2018) also extended Prop. 2.10 to any class of ergodic MDPs with arbitrary *structure* where the reward function is also unknown (see Prop. 2.11). We denote by $\mathcal{M}$ such a class of ergodic MDPs with $r_{\max} = 1$ (and potentially continuous state and action spaces). We generalize the definition of the set of confusing MDPs:

$$\Phi := \Big\{ M' = \langle \mathcal{S}, \mathcal{A}, r, p' \rangle : \; p'(s'|s,a) = 0 \implies p(s'|s,a) = 0, \; \forall s, s' \in \mathcal{S}, \forall a \in \mathcal{A} \\ p'(\cdot|s,\pi^*(s)) = p(\cdot|s,\pi^*(s)), \; \forall s \in \mathcal{S}, \forall \pi^* \in \Pi^*, \\ \text{and } \Pi^* \cap \Pi'^* = \emptyset \Big\}.$$

Proposition 2.11 (Theorem 1 of Ok et al. (2018))

Let $M \in \mathcal{M}$ and $\mathfrak{A}$ be a learning algorithm s.t. $\mathbb{E}^{\mathfrak{A}}\left[\Delta(M', \mathfrak{A}, \mu_1, T) | s_1 \sim \mu_1\right] = o(T^\alpha)$ for all $\alpha > 0$, all $M' \in \mathcal{M}$ and initial state distribution μ_1. The expected regret of $\mathfrak{A}$ is lower bounded as

$$\liminf_{T\to\infty} \frac{\mathbb{E}^{\mathfrak{A}}\left[\Delta(M, \mathfrak{A}, \mu_1, T) | s_1 \sim \mu_1\right]}{\log T} \geq K$$

where

$$\begin{aligned} K = \inf_{\eta \geq 0} & \sum_{s,a} \eta(s,a)\delta(s,a) \\ s.t. & \sum_{s,a} \eta(s,a) KL_{M\|M'}(s,a) \geq 1 \quad \forall M' \in \Phi. \end{aligned}$$

When $\mathcal{M}$ is *unstructured* (class of all ergodic MDPs), one can show (Ok et al., 2018, Section 4.1) that Prop. 2.11 allows to recover Prop. 2.10. Ok et al. (2018, Section 4.2) also analysed the case where $\mathcal{S}$ and $\mathcal{A}$ are subset of metric spaces and r and p are Lipschitz-continuous.

There are several major limitations to Prop. 2.10 and Prop. 2.11. The lower-bounds are derived only for ergodic MDPs and it is an open question whether the lower-bound increases when extended to more general chain structures (like communicating or weakly-communicating). But perhaps the main limitation is the *asymptotic* nature of the lower-bounds. These bounds provide no indication on the regret performance in *finite time*.

Minimax lower-bounds

We will now present a different type of lower-bound proved by Jaksch et al. (2010). Before that, we need to introduce the notion of *diameter* of an MDP.

Definition 2.6

The diameter of an MDP is defined as

$$D := \max_{s,s'} \min_{\pi \in \Pi^{SD}} \mathbb{E}^{\pi}\left[\tau(s') | s_1 = s\right] - 1 \tag{2.22}$$

where $\tau(s') := \inf\{t \geq 1 : s_t = s'\}$ is the first hitting time *in s'.*

From Def. 2.2 and Proposition 8.3.1 of Puterman (1994), it is clear that $D < +\infty$ if and only if M is communicating. The *diameter* of an MDP is the length of the *longest shortest path* in the MDP. In other words, it is the length of the shortest path between the two states that are the *most distant* from each other. It quantifies the difficulty to *navigate* in the MDP. We provide a graphical illustration on Fig 2.4.

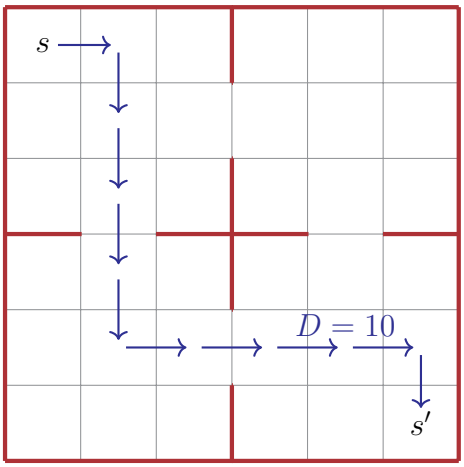


Figure 2.4: Graphical illustration of Def. 2.6. The MDP is a grid-world where every square represents a state. The four cardinal actions can be played in any state with success probability 1, except when there is a wall (red).

Proposition 2.12 (Theorem 5 of Jaksch et al. (2010))

For any algorithm $\mathfrak{A}$, any integers $S, A \geq 10$, $D \geq 20 \log_A(S)$, and $T \geq DSA$, there is an MDP M with at most S states, A actions, and diameter D, such that for any initial distribution $\mu_1 \in \Delta_S$, the expected regret of $\mathfrak{A}$ after T time steps is lower-bounded as

$$\mathbb{E}^{\mathfrak{A}}\left[\Delta(M, \mathfrak{A}, \mu_1, T) \middle| s_1 \sim \mu_1\right] \geq 0.015 \cdot r_{\max}\sqrt{DSAT}.$$

Prop. 2.12 significantly differ from Prop. 2.10 and 2.11. It shows that for any number of states S, number of actions A and diameter D, it is always possible to construct a *worst-case* MDP with these features that achieves a regret of order at least $\Omega\left(r_{\max}\sqrt{DSAT}\right)$. Unlike the bounds of Burnetas and Katehakis (1997) and Ok et al. (2018), Prop. 2.12 is not *problem-dependent* but it is also *not asymptotic*. Problem-dependent non-asymptotic bounds would combine the best of both worlds but to the best of our knowledge, no such bounds are currently available in the RL literature. Bounds on the worst-case regret are often referred as *"minimax" bounds*. Minimax bounds usually scale as $\sqrt{T}$ while problem dependent bounds scale logarithmically with T.[9]

The term D (diameter) appearing in the bound of Prop. 2.12 can be deceiving. The specific worst-case MDP constructed by Jaksch et al. (2010) to prove the lower-bound satisfies $D = 2sp\left(h^*\right)$ and so it is not clear whether to interpret the lower-bound in terms of diameter, range of the bias or yet another term. This ambiguity is one of the major issues with minimax lower-bounds.

Bartlett and Tewari (2009, Theorem 6) tried to improve the bound of Jaksch et al. (2010) but Osband and Van Roy (2016) later showed that their proof contains a mistake. The work presented in this thesis together with other recent work (Ortner, 2018; Tossou et al., 2019)

[9] In the bandit litterature,"problem-dependent" bounds are said to be *distribtion-dependent*, as opposed to minimax bounds which are said to be *distribution-free* (Garivier et al., 2018).

suggest that the lower-bound of Prop. 2.12 cannot be improved (without restricting the family of possible MDPs).

2.2.3 (Near) Optimal algorithms

A common strategy to efficiently balance exploration and exploitation in RL is to apply the *optimism in face of uncertainty* (OFU) principle: the agent maintains *optimistic estimates* of the MDP parameters and, at each step, executes the policy with highest optimistic "value" (e.g., gain, discounted value function, etc.). In this section, we will review some of the existing RL algorithms relying on OFU.

An alternative approach is posterior sampling (Thompson, 1933b), which maintains a Bayesian distribution over MDPs (i.e., dynamics and expected reward) and, at each step, samples an MDP and executes the corresponding optimal policy (e.g., Osband et al., 2013; Abbasi-Yadkori and Szepesvári, 2015; Osband and Roy, 2017; Ouyang et al., 2017a). Unfortunately, so far all existing posterior sampling algorithms only provide garantees on the Bayesian regret. A notable exception is the work of (Agrawal and Jia, 2017) which successfully combines posterior sampling with OFU to obtain guarantees on the frequentist regret. However, their algorithm requires to sample multiple times the posterior distribution over MDPs so as to obtain empirical high-probability confidence bounds, which somehow resembles what OFU methods do in a computationally more efficient way.

Asymptotically optimal algorithms

Burnetas and Katehakis (1997) proposed Optimal Adaptive Policies (OAP) that achieve the lower-bound of Prop. 2.10 i.e.,

$$\limsup_{T\to\infty} \frac{\mathbb{E}^{\mathfrak{A}}\left[\Delta(M,\mathrm{OAP},\mu_1,T)|s_1\sim\mu_1\right]}{\log T} \leq \sum_{\substack{s,a\\ \Phi(s,a)\neq\emptyset}} \frac{\delta(s,a)}{\inf_{M'\in\Phi(s,a)} \mathrm{KL}_{M\|M'}(s,a)}. \tag{2.23}$$

At each time step t, OAP computes an estimate $\widehat{M}_t = \langle \mathcal{S}, \mathcal{A}, r, \widehat{p}_t\rangle$ of the unknown MDP M based on past-observations ($\widehat{p}_t$ is the maximum-likelihood estimator of p). An optimal solution $(\widehat{g}_t, \widehat{h}_t) \in \mathbb{R}\times\mathbb{R}^S$ of the optimality equation of $\widehat{M}_t$ is then computed i.e., a solution to $\widehat{L}_t\widehat{h}_t = \widehat{h}_t + \widehat{g}_t e$ where $\widehat{L}_t$ is the optimal Bellman operator of $\widehat{M}_t$. Let $N_t(s,a)$ denote the number of past visits in state-action pair (s,a) and $\mathcal{L}_t : \mathbb{R}^S \mapsto \mathbb{R}^S$ be the operator defined for all $v\in\mathbb{R}^S$ and all $s\in\mathcal{S}$ by

$$\mathcal{L}_t v(s) := \max_{a\in\mathcal{A}_s}\left\{ r(s,a) + \max_{q\in B_p^t(s,a)} q^\intercal v \right\}, \tag{2.24}$$

where $B_p^t(s,a) := \{q\in\Delta_S :\ \mathrm{KL}(\widehat{p}_t(\cdot|s,a)\|q) \leq \ln(t)/N_t(s,a)\}$ is a high probability confidence set for $p(\cdot|s,a)$. $\mathcal{L}_t$ is an extended optimal Bellman operator (Sec. 2.1.5). At every time step t, the current state is denoted s_t and OAP plays any greedy action w.r.t. $\mathcal{L}_t\widehat{h}_t(s_t)$. To avoid *under-exploration* of some state-action pairs, OAP sometimes needs to play an action

that have not been visited *"sufficiently often"* instead i.e., among

$$\left\{ a \in \mathcal{A} : \; N_t(s,a) < \ln^2 \left(\sum_{b \in \mathcal{A}_s} N_t(s,b) + 1 \right) \right\} \subseteq \mathcal{A}_s.$$

Tewari and Bartlett (2007b) derived a similar algorithm called OLP (Optimistic Linear Programming) that defines confidence sets $B_p^t(s,a)$ using the ℓ_1-norm instead of the Kullback-Leibler divergence: $B_p^t(s,a) := \left\{ q \in \Delta_S : \; \|\widehat{p}_t(\cdot|s,a) - q\|_1 \leq \sqrt{2 \ln(t) / N_t(s,a)} \right\}$. The regret guarantees are slightly worse: the term $\mathrm{KL}_{M\|M'}(s,a)$ in (2.23) is replaced by a similar term depending on the distance in ℓ_1-norm rather than Kullback-Leibler divergence. However, computing the maximum over q in (2.24) becomes computationally easier: it can be expressed as a linear programming problem (hence the name OLP).

Both OAP and OLP implement the OFU maxim through the extended Bellman operator $\mathcal{L}_t$ which is an *"optimistic"* version of L (at least in high probability). Ok et al. (2018) derived Directed Exploration Learning (DEL) which is able to achieve the lower-bound of Prop. 2.11 with an *explicit* explore versus exploit strategy instead. Depending on past observation, DEL decides to *exploit* i.e., to take the greedy policy w.r.t. $\widehat{L}_t \widehat{h}_t(s_t)$ (rather than $\mathcal{L}_t \widehat{h}_t(s_t)$), or to *explore* by explicitly using the expression of the estimated lower-bound $\widehat{K}_t$ (solution to optimization problem in Prop. 2.11 with M replaced by $\widehat{M}_t$). Unlike OAP and OLP, DEL does not rely on OFU.

Optimal algorithms with finite time guarantees

UCRL. The first algorithm with provable *finite time* regret guarantees is UCRL (Upper Confidence Bounds Reinforcement Learning) introduced by Auer and Ortner (2007). For any ergodic MDP M, let $\Pi^* \subseteq \Pi^{\mathrm{SD}}$ be the set of stationary deterministic gain-optimal policies in M and

$$\tau_{\max} := \max_{s,s'} \max_{\pi \in \Pi^{\mathrm{SD}}} \mathbb{E}^{\pi} \left[\tau(s') | s_1 = s \right] - 1$$

the worst case *mixing time*. Unlike the diameter (Def. 2.6), $\tau_{\max}$ is a double maximum and so $\tau_M < +\infty$ only when M is *ergodic*. We also define

$$\kappa_{\max} := \frac{1}{2} \max_{\pi \in \Pi^{\mathrm{SD}}} \max_{s'} \frac{\max_s \mathbb{E}^{\pi} \left[\tau(s') | s_1 = s \right] - 1}{\mathbb{E}^{\pi} \left[\tau_2(s') | s_1 = s' \right] - 1}$$

the worst-case *condition number* of M (Kirkland et al., 2008, condition number κ_8), where $\tau_2(s') := \inf\{t \geq 2 : s_t = s'\}$ is the first *return time* in s'. Finally, the gap in gains is

$$\delta_g := g^* - \max_{\pi \in \Pi^{\mathrm{SD}} \setminus \Pi^*} \left\{ \max_{s: \; g^{\pi}(s) < g^*} g^{\pi}(s) \right\}.$$

Auer and Ortner (2007, Theorem 2) proved that there exists a numerical constant $\beta > 0$ such that for any *ergodic* MDP M, for all initial state distribution $\mu_1 \in \mathcal{P}(\mathcal{S})$ and for all $T > 1$:

$$\mathbb{E}\left[\Delta(M, \mathrm{UCRL}, \mu_1, T)\right] \leq \beta \cdot \frac{S^5 A \tau_{\max} \kappa_{\max}^2}{\delta_g} \ln(T) + 3 S^2 A^2 \tau_{\max} \log_2 \left(\frac{T}{SA} \right). \tag{2.33}$$

Algorithm 4 UCRL2

Input: Confidence $\delta \in]0,1[$, maximal reward $r_{\max}$, set of states $\mathcal{S}$, set of actions $\mathcal{A}$

Set initial time $t := 1$, observe initial state s_1 and initialize for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$:
- counters $N_1(s,a) := 0$,
- empirical averages $\widehat{p}_1(s'|s,a) := 0$ and $\widehat{r}_1(s,a) := 0$.

for episodes $k = 1, 2, ...$ **do**

Set the starting time of the episode $t_k := t$ and initialize for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$: episode counters $\nu_k(s,a,s') := 0$ and $\nu_k(s,a) := 0$, and cumulative rewards $R_k(s,a) := 0$.

For all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, compute upper confidence bounds:

$$\beta_{p,k}^{sas'} := \sqrt{\frac{14S \ln\left(\frac{2At_k}{\delta}\right)}{N_k^+(s,a)}} \tag{2.25}$$

$$\beta_{r,k}^{sa} := \sqrt{\frac{7 \ln\left(\frac{2SAt_k}{\delta}\right)}{2N_k^+(s,a)}} \tag{2.26}$$

Set $\mathcal{M}_k := \{\mathcal{S}, \mathcal{A}, r_k, p_k\}$ to be the extended MDP defined by the confidence intervals

$$p_k(s'|s,a) \in B_p^k(s,a) := \left\{ q \in \Delta_S : \ \left\| q - \widehat{p}_k(s'|s,a) \right\|_1 \leq \beta_{p,k}^{sa} \right\} \tag{2.27}$$

$$r_k(s,a) \in B_r^k(s,a) := \left[\widehat{r}_k(s,a) - \beta_{r,k}^{sa}, \widehat{r}_k(s,a) + \beta_{r,k}^{sa} \right] \cap [0, r_{\max}] \tag{2.28}$$

Compute policy π_k using ("extended") value iteration (Alg. 3):

$$(g_k, h_k, \pi_k) := \mathrm{EVI}\left(\mathcal{L}_k, \mathcal{G}_k, \frac{r_{\max}}{\sqrt{t_k}}, 0, s_1 \right) \tag{2.29}$$

while $\nu_k(s_t, \pi_k(s_t)) \leq N_k^+(s_t, \pi_k(s_t))$ **do**

Execute action $a_t := \pi_k(s_t)$, obtain reward r_t, and observe next state s_{t+1}.

Increment episode counters:

$$\nu_k(s_t, a_t, s_{t+1}) \leftarrow \nu_k(s_t, a_t, s_{t+1}) + 1 \quad \text{and} \quad \nu_k(s_t, a_t) \leftarrow \nu_k(s_t, a_t) + 1$$

Increment cumulative reward: $R_k(s_t, a_t) \leftarrow R_k(s_t, a_t) + r_t$

Increment time $t \leftarrow t + 1$

end while

Update counters, empirical averages and sample variances for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$:

$$N_{k+1}(s,a) := N_k(s,a) + \nu_k(s,a) \tag{2.30}$$

$$\widehat{p}_{k+1}(s'|s,a) := \frac{N_k(s,a)}{N_{k+1}^+(s,a)} \cdot \widehat{p}_k(s'|s,a) + \frac{\nu_k(s,a,s')}{N_{k+1}^+(s,a)} \tag{2.31}$$

$$\widehat{r}_{k+1}(s,a) := \frac{N_k(s,a)}{N_{k+1}^+(s,a)} \cdot \widehat{r}_k(s,a) + \frac{R_k(s,a)}{N_{k+1}^+(s,a)} \tag{2.32}$$

end for

Although it is difficult to compare (2.33) with the lower-bound of Prop. 2.10, the regret bound of UCRL is likely to be much worse given the dependency in S (among other things).

Similarly to OAP and OLP, UCRL maintains maximum-likelihood estimates of r and p as well as confidence sets $B_r(s,a)$ and $B_p(s,a)$ based on high probability confidence bounds. But unlike OAP and OLP, UCRL updates the policy only once the confidence bounds of at least one state-action pair have been *halved* since the last policy update. The time interval between two policy updates is called an *"episode"*. At each episode k, UCRL2 computes a policy

$$\pi_k \in \arg\max_{\pi \in \Pi^{\mathrm{SD}}} \sup_{M' \in \mathcal{M}_k} g^{\pi}_{M'}$$

where

$$\mathcal{M}_k := \left\{ M' = \langle \mathcal{S}, \mathcal{A}, r, p \rangle : \ M \text{ is ergodic}, \ r(s,a) \in B^k_r(s,a), \ B^k_p(s,a) \right\}$$

is the set of plausible ergodic MDPs compatible with the confidence sets. It is a bounded-parameter MDP (see Sec. 2.1.5) with the additional constraint that the MDPs it contains should all be ergodic. The confidence sets are constructed so that $M \in \mathcal{M}_k$ with high probability implying $\sup_{M' \in \mathcal{M}_k} g^*_{M'} \geq g^*$ i.e., π_k is *gain-optimistic*. UCRL is therefore another instance of RL algorithm relying on OFU.

UCRL2. Jaksch et al. (2010) later improved UCRL with UCRL2. Since all the RL algorithms presented in this thesis are variants of UCRL2, we report the detailed pseudo-code in Alg. 4. To improve the readibility of the algorithm, we use the notation $n^+ := \max\{1, n\}$ for any positive integer $n \in \mathbb{N}$.

UCRL2 and UCRL share a *similar structure*. Both algorithms proceed through *episodes*. At the beginning of each episode, a *stationary policy* is computed by taking into consideration the *past observations*. The policy computed also takes into account the *uncertainty* of observed data by constructing a bounded-parameter MDP $\mathcal{M}_k$ (similar to the bounded-parameter of UCRL without the constraint on ergodicity). This policy is executed until the end of the episode. A new episode then starts and the policy is updated based on the new observations gathered during the last episode. This procedure is *repeated* until the desired time horizon is reached.

When the MDP is communicating but not ergodic, switching stationary policies *too often* can cause a large –even linear– regret as shown by Ortner (2010, Example 1). To avoid too many non-stationarities in the policy executed by the algorithm, the episodes are designed to have a length that grows *exponentially* with time. This way, the number of episodes (i.e., the number of policy switches) is at most *logarithmic* in time causing only a minor increase in the regret. More precisely, an episode ends when the number of visit in a state-action pair has doubled since the end of the previous episode.

Given the bounded parameter MDP $\mathcal{M}_k$, UCRL2 executes a policy π_k which is an approximate solution to the following optimization problem:

$$\max_{\pi \in \Pi^{\mathrm{SD}}} \left\{ \sup_{M' \in \mathcal{M}_k} g^{\pi}_{M'} \right\} = \sup_{M' \in \mathcal{M}_k} \left\{ \max_{\pi \in \Pi^{\mathrm{SD}}} g^{\pi}_{M'} \right\} = \sup_{M' \in \mathcal{M}_k} g^*_{M'}. \tag{2.34}$$

If $M \in \mathcal{M}_k$ (with high probability), the solution of (2.34) is an upper-bound to g^* and so π_k is (nearly) gain-optimistic (like in UCRL). Since we do not restrict $\mathcal{M}_k$ to ergodic MDPs (like in UCRL), the associated optimal gain might not be state-independent and so the gains of two different MDPs might not always be *comparable*[10]. One might wonder whether (2.34) is well-posed and admits maximizer $(M_k, \pi_k^*) \in \mathcal{M}_k \times \Pi^{\mathrm{SD}}$. Using the mapping between bounded-parameters MDPs and extended MDPs (Sec. 2.1.5), it is possible to interpret $\mathcal{M}_k$ as an MDP. Eq. 2.16 can then be rewrtitten[11] as finding the optimal policy of $\mathcal{M}_k$:

$$\max_{\pi \in \Pi^{\mathrm{SD}}_{\mathcal{M}_k}} g^{\pi}_{\mathcal{M}_k} = g^*_{\mathcal{M}_k} \tag{2.35}$$

Since $M \in \mathcal{M}_k$ (with high probability) and M is communicating, so is $\mathcal{M}_k$. Moreover, the confidence sets $B_p^k(s,a)$ (2.27) are polytopes and we already explained in Sec. 2.1.5 that in this case, the action space can be restricted to a *finite* set. We can thus apply the tools of Sec. 2.1.3: we know that a maximizer of (2.35) always exists (Prop. 2.4) and we can compute an approximate solution using value iteration[12] (Alg. 3). Since value iteration is run with the extended optimal Bellman operator $\mathcal{L}_k$ of $\mathcal{M}_k$, we call the algorithm *"extended" value iteration* (EVI). The accuracy ε_k and extended greedy operator $\mathcal{G}_k$ given as input to EVI are respectively $r_{\max}/\sqrt{t_k}$ and

$$\forall s \in \mathcal{S}, \forall v \in \mathbb{R}^S, \quad \mathcal{G}_k v(s) \in \arg\max_{a \in \mathcal{A}_s} \left\{ \max_{r \in B_r^k(s,a)} r + \max_{p \in B_p^k(s,a)} p^\intercal v \right\}. \tag{2.36}$$

Jaksch et al. (2010, Section 3.1.3) showed that assumption 2 of Prop. 2.6 hold so that EVI converges and g_k approximates $g^*_{\mathcal{M}_k}$ with an $r_{\max}/\sqrt{t_k}$-accuracy. Enumerating the vertices of the sets $B_p^k(s,a)$ is not the most computationally efficient method to implement EVI. The maximization of $p^\intercal v$ under the constraint $p \in B_p^k(s,a)$ can be expressed as a *linear programming* (LP) problem (which can be solved efficiently using a generic solver). Strehl and Littman (2008a) provide a better algorithm that exploits the specific structure of this LP (see also Jaksch et al., 2010, Figure 2). It runs in $\mathcal{O}(S)$ once the vector v has been sorted in descending order. The sorting operation requires $\mathcal{O}(S \ln(S))$ operations but needs only be done once for all (s,a).

UCRL2 enjoys the following regret guarantees.

Proposition 2.13 (Theorem 4 of Jaksch et al. (2010))

For any communicating *MDP, there exists a constant $C(M)$ such that with probability at least $1-\delta$, it holds that for all initial state distributions $\mu_1 \in \Delta_S$ and for all time horizons $T > 1$:*

$$\mathbb{E}\left[\Delta(M, \mathrm{UCRL2}, \mu_1, T)\right] \leq 34^2 \cdot \frac{r_{\max} D^2 S^2 A}{\delta_g} \ln(T) + C(M). \tag{2.37}$$

[10] If the MDPs M_1 and M_2 both belong to $\mathcal{M}_k$ but have non constant optimal gains $g^*_{M_1}$ and $g^*_{M_2}$, it is possible that $g^*_{M_1}(s) > g^*_{M_2}(s)$ while $g^*_{M_1}(s') < g^*_{M_2}(s')$ for some $s' \neq s$.

[11] Eq. 2.34 and Eq. 2.35 are equivalent.

[12] In Alg. 4, we refer to value iteration applied to an extended Bellman operaor as "extended" value iteration (EVI) even though this is just a specific instance of value iteration.

The exact expression of the constant $C(M)$ can be found in (Jaksch et al., 2010). It depends on some form of worst-case *mixing time* of M (different than τ_M). The logarithmic term in 2.37 is tighter than (2.33) and holds for the broader class of *communicating* MDPs (rather than just ergodic MDPs). The bound is still difficult to compare with Prop. 2.10 but we can easily compare it with the worst-case upper-bound $\frac{2r_{\max}(sp(h^*)+1)^2SA}{\delta_{\min}}$ of Ok et al. (2018, Section 4.1) mentioned earlier. As shown by Bartlett and Tewari (2009, Theorem 4) (more details will be given in Sec. 3.3 of Chap. 3), the range of the bias function is at most $r_{\max}D$ i.e., $sp\left(h^*\right) \leq r_{\max}D$, and the equality holds in some MDPs. Moreover, the gap in gain δ_g is always smaller than $\delta_{\min}$ as shown in the following lemma.

Lemma 2.1

For any ergodic MDP, $\delta_g \leq \delta_{\min}$.

Proof. If $\Pi^* = \Pi^{\text{SD}}$ then $\delta_g = \delta_{\min} = 0$. Otherwise, we denote by $(s^-, a^-) \in \mathcal{S} \times \mathcal{A}$ the state-action pair achieving the minimum in (2.21) i.e., such that $\delta_{\min} = \delta(s^-, a^-)$. We define the action space $\mathcal{A}^-$ such that $\mathcal{A}^-_s = \mathcal{A}_s$ for all $s \neq s^-$ and $\mathcal{A}^-_{s^-} := \{a \in \mathcal{A}_{s^-} : \ \delta(s^-, a) > 0\}$. $\mathcal{A}^-$ contains all actions except optimal actions in state s^-. Let $M^- := \langle \mathcal{S}, \mathcal{A}^-, r, p \rangle$ be the MDP defined on the action space $\mathcal{A}^-$, with L^- and g^- the corresponding optimal Bellman operator and optimal gain (the gain is state-independent since both M and M^- are ergodic). By construction, $g^* > g^-$ and based on (2.18) we can write

$$\delta_{\min} = g^* - L^- h^*(s^-) + h^*(s^-) = g^* - \min_s \left\{L^- h^*(s) - h^*(s)\right\}$$

Theorem 8.5.5. of Puterman (1994) implies that

$$g^- \geq \min_s \left\{L^- h^*(s) - h^*(s)\right\}$$

and so necessarily $\delta_{\min} \geq g^* - g^- = g^* - g^{\pi^-}$ where $\pi^- \in \Pi^{\text{SD}}$ is any gain-optimal stationary deterministic policy of M^-. π^- is also a valid policy in the original MDP M with $g^{\pi^-} < g^*$. As a result, $\max_{\pi \in \Pi^{\text{SD}} \setminus \Pi^*} \{g^\pi\} \geq g^{\pi^-}$ which implies that $\delta_{\min} \geq \delta_g$. ■

Finally, *Multi-Armed Bandit* problems are specific instances of ergodic MDPs (with a single state) satisfying $\delta_g = \delta_{\min}$. In conclusion, the bound of Prop. 2.13 is always worse than $\frac{2r_{\max}(sp(h^*)+1)^2SA}{\delta_{\min}}$ but in the *worst case* the two expressions are *comparable up to a factor S*. This suggests that asymptotically, UCRL2 is at least S-loose in terms of regret, which is not so bad. The regret analysis of OAP is very different from the proof of Prop. 2.13. We conjecture that the proofs techniques of Burnetas and Katehakis (1997) can probably be applied to the analysis of UCRL2 and lead to an asymptotic regret bound almost matching the lower bound of Prop. 2.10 (probably up to a factor S).

In addition to the logarithmic regret bound of Prop. 2.13, Jaksch et al. (2010) also proved a *minimax bound* for UCRL2.

Proposition 2.14 (Theorem 2 of Jaksch et al. (2010))

For any communicating *MDP, with probability at least* $1-\delta$*, it holds that for all initial state distributions* $\mu_1 \in \Delta_S$ *and for all time horizons* $T > 1$*:*

$$\Delta(M, \text{UCRL2}, \mu_1, T) \leq 34 \cdot DS\sqrt{AT \ln\left(\frac{T}{\delta}\right)}. \tag{2.38}$$

Compared to the minimax lower-bound of Prop. 2.12, the bound of Prop. 2.14 is looser by a factor $\sqrt{DS}$ (ignoring logarithmic terms).

Extensions. Bartlett and Tewari (2009) tried to extend UCRL2 to the case where an upper-bound $c \geq sp\,(h^*)$ on the optimal bias span is known. The regret bound then scales with c instead of D. This will be the focus of Chap. 5. Filippi et al. (2010) derived a variant of UCRL2 (called KL-UCRL) that uses concentration inequalities on the Kullback-Leibler divergence (instead of Hoeffding/Weissman inequality) to construct confidence bounds. The regret upper-bound they prove is the same as in Prop. 2.14. Despite proving the same bound, the authors empirically observe the superiority of KL-UCRL over UCRL2. They provide some intuition to explain their results and Talebi and Maillard (2018b) later showed that the regret analysis can be refined. Talebi and Maillard (2018b) indeed showed that the regret of KL-UCRL scales as $\widetilde{\mathcal{O}}\left(\sqrt{S\sum_{s,a} V^*_{s,a} T} + D\sqrt{T}\right)$ (ignoring logarithmic terms) where $V^*_{s,a} := \mathbb{V}_{X\sim p(\cdot|s,a)}\left(h^*(X)\right)$ is the variance of the optimal bias w.r.t. the next state. Since $V^*_{s,a} \leq sp\,(h^*) \leq r_{\max} D$, the bound is smaller than in Prop. 2.14. Nevertheless, the bound only holds for ergodic MDPs and the logarithmic terms hidden in the $\widetilde{\mathcal{O}}$-notation can be very big. More recently, Ortner (2018) derived an algorithm called OSP (Optimistic Sample Path) which leverages Markov Chain concentration inequalities. When run on an unknown ergodic MDP with mixing time t_{mix} (the definition differ from $\tau_{\max}$), the regret can be bounded (w.h.p.) as $\mathcal{O}(\sqrt{t_{\text{mix}} SAT} \ln(T/\delta))$. In some specific MDPs, t_{mix} is comparable to D so that OSP achieves the minimax lower-bound (up to logarithmic factors).[13] However, OSP requires explicitly enumerating all A^S policies which makes it *intractable*. Finally, the work of Tossou et al. (2019) (still unpublished) suggests that it is possible to design a tractable algorithm (variant of UCRL2) called UCRL-V with optimal minimax regret guarantees.

[13] One example where t_{mix} is of the same order as D is actually the family of MDPs used by Jaksch et al. (2010) to prove Prop. 2.12.

3 Improved exploration-exploitation with Bernstein bounds

In the previous chapter, we gave a high-level overview of UCRL2 (Jaksch et al., 2010) and compared its regret performance (Prop. 2.13 and 2.14) to existing lower-bounds (Prop. 2.10 and 2.12). In this chapter, we introduce several modifications to the algorithm and improve the minimax regret guarantees of Prop. 2.14. Our proposed algorithm, UCRL2-Bernstein (UCRLB for short), leverages empirical Bernstein inequality as well as recent contributions of the literature (including from other settings e.g., infinite horizon discounted and finite horizon) to make a significant step towards closing the gap between minimax regret upper and lower-bounds. For any communicating MDP with S states, A actions, $\Gamma \leq S$ possible next states and diameter D, we show that UCRLB suffers at most $\widetilde{\mathcal{O}}\left(\sqrt{D\Gamma SAT}\right)$ regret (ignoring logarithmic terms). This saves a factor $\sqrt{DS/\Gamma}$ compared to the regret bound of UCRL2. Since in many MDPs $\Gamma = \mathcal{O}(1) \ll S$, this bound is also almost matching the minimax lower-bound of Prop. 2.12. Although many ideas presented in this chapter are not new, we make several important contributions to the regret analysis of this type of algorithms, and provide new insights on existing proofs techniques. For example, we provide a more generic and insightful proof of gain-optimism relying on the properties of the extended Bellman operator rather than the extended MDP. We also refine the regret analysis by introducing a new quantity called "travel-budget" of an MDP, that replaces the diameter in the bound.

Another objective of this chapter is to present a unified framework for the analysis of UCRL2-like algoritms. All the algorithms presented in the next chapters of this thesis will be variants of UCRLB and most of the analysis will be unchanged. To minimize redundancies and improve clarity, this is the only chapter where we will provide a fully detailed analysis. In subsequent chapters, we will refer to this chapter for the parts of the analysis that are similar, and only focus on what significantly differs. In order to keep the structure of the regret proofs identical across chapters, we consider a very general version of the algorithm, more than is actually needed for the setting of this chapter. For example, we allow the optimal optimisitc policy to be stochastic although a deterministic policy always exists. This will be useful in Chap. 5. We will also apply the aperiodicity transformation in EVI even if this is not strictly necessary with the extended MDP considered here.

Most of the work presented in this chapter has not been published in any venue so far.

3.1 Upper Confidence Reinforcement Learning with Bernstein bounds

UCRL2 and UCRLB are very similar from an algorithmic point of view. The main difference lies in the definition of the extended MDP. In the regret analysis (Sec. 3.4 and 3.5), we will show that the modifications that we propose result in a much more *sample efficient* algorithm. In this section, we start by giving an overview of the main *features* of UCRLB. We also highlight and explain the main *differences* with UCRL2.

3.1.1 Detailed algorithm and notations

The detailed *pseudo-code* of UCRLB is reported in Alg. 5. In what follows, we give additional explanations and we introduce several notations.

Time steps and visit counts. The time steps (occurences of a new state) are indexed by $T \geq t \geq 1$. The *state visited at time* t is denoted s_t while the *action played at time* t is denoted a_t. The episodes (corresponding to policy switches like in UCRL2, see Sec. 2.2.3) are indexed by k, and π_k is the policy executed during episode k. In some specific applications (see for example Chap. 5), it is too restrictive to constrain π_k to belong to the set of *deterministic* policies Π^{SD}. For this reason, we allow π_k to be a *stationary randomized* policy (even though in most cases this level of generality is not needed). At every time step t of episode k, a_t is sampled from the distribution $\pi_k(\cdot|s_t)$. After action a_t has been played, a reward r_t is earned and the next state s_{t+1} is observed. For all $k \geq 1$, we denote by t_k the starting time of episode k. The *fisrt episode* starts at $t_1 := 1$. A *new episode* starts whenever the stopping condition of the current episode is met i.e., whenever the number of visits in the state-action pair (s_t, a_t) has *doubled* during the episode. Formally, for all $k \geq 1$,

$$\begin{aligned} t_{k+1} &:= \inf \left\{ T \geq t > t_k : \sum_{\tau=1}^{t-1} \mathbb{1}\left\{s_\tau, a_\tau = s_t, a_t\right\} \geq \max \left\{ 1, 2 \sum_{\tau=1}^{t_k-1} \mathbb{1}\left\{s_\tau, a_\tau = s_t, a_t\right\} \right\} \right\} \\ &= \inf \left\{ T \geq t > t_k : \sum_{\tau=t_k}^{t-1} \mathbb{1}\left\{s_\tau, a_\tau = s_t, a_t\right\} \geq \max \left\{ 1, \sum_{\tau=1}^{t_k-1} \mathbb{1}\left\{s_\tau, a_\tau = s_t, a_t\right\} \right\} \right\} \end{aligned} \tag{3.11}$$

where $\inf\{\emptyset\} \leftarrow T + 1$ by convention. Note that by construction, the stopping condition of episode k is always met after at most t_k steps. For all $T > t \geq 1$, we define the *episode at time* t by

$$k_t := \sup\{k \geq 1 : \ t \geq t_k\}. \tag{3.12}$$

Algorithm 5 UCRL-Bernstein (UCRLB)

Input: Confidence $\delta \in]0,1[$, maximal reward $r_{\max}$, set of states $\mathcal{S}$, set of actions $\mathcal{A}$

Set initial time $t := 1$, observe initial state s_1 and initialize for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$:
- counters $N_1(s,a,s') := 0$ and $N_1(s,a) := 0$,
- empirical averages $\widehat{p}_1(s'|s,a) := 0$ and $\widehat{r}_1(s,a) := 0$,
- sample variances $\widehat{\sigma}^2_{p,1}(s'|s,a) := 0$ and $\widehat{\sigma}^2_{r,1}(s,a) := 0$.

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

Set the starting time of the episode $t_k := t$ and initialize for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$: episode counters $\nu_k(s,a,s') := 0$ and $\nu_k(s,a) := 0$, and cumulative (squared) rewards $R_k(s,a) := 0$ and $S_k(s,a) := 0$. ▷ Initialization of episode k

For all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, compute upper confidence bounds:

$$\beta^{sas'}_{p,k} := 2\sqrt{\frac{\widehat{\sigma}^2_{p,k}(s'|s,a)}{N^+_k(s,a)} \ln\left(\frac{6SAN^+_k(s,a)}{\delta}\right)} + \frac{6\ln\left(\frac{6SAN^+_k(s,a)}{\delta}\right)}{N^+_k(s,a)} \tag{3.1}$$

$$\beta^{sa}_{r,k} := 2\sqrt{\frac{\widehat{\sigma}^2_{r,k}(s,a)}{N^+_k(s,a)} \ln\left(\frac{6SAN^+_k(s,a)}{\delta}\right)} + \frac{6r_{\max}\ln\left(\frac{6SAN^+_k(s,a)}{\delta}\right)}{N^+_k(s,a)} \tag{3.2}$$

Set $\mathcal{M}_k := \{\mathcal{S}, \mathcal{A}, r_k, p_k\}$ to be the extended MDP defined by the confidence intervals

$$p_k(s'|s,a) \in B^k_p(s,a,s') := \left[\widehat{p}_k(s'|s,a) - \beta^{sas'}_{p,k}, \widehat{p}_k(s'|s,a) + \beta^{sas'}_{p,k}\right] \cap [0,1] \tag{3.3}$$

$$r_k(s,a) \in B^k_r(s,a) := \left[\widehat{r}_k(s,a) - \beta^{sa}_{r,k}, \widehat{r}_k(s,a) + \beta^{sa}_{r,k}\right] \cap [0, r_{\max}] \tag{3.4}$$

Compute policy π_k using extended value iteration (see Eq. 3.20 and Alg. 6):

$$(g_k, h_k, \pi_k) := \mathrm{EVI}\left(\mathcal{L}^k_\alpha, \mathcal{G}^k_\alpha, \frac{r_{\max}}{t_k}, 0, s_1\right) \tag{3.5}$$

Sample action $a_t \sim \pi_k(\cdot|s_t)$. ▷ Stochastic policies are allowed

while True do ▷ Execute policy π_k until the end of episode k

Execute action a_t, obtain reward r_t, and observe next state s_{t+1}.

Increment episode counters:

$$\nu_k(s_t,a_t,s_{t+1}) \leftarrow \nu_k(s_t,a_t,s_{t+1}) + 1 \quad \text{and} \quad \nu_k(s_t,a_t) \leftarrow \nu_k(s_t,a_t) + 1$$

Increment cumulative (squared) reward

$$R_k(s_t,a_t) \leftarrow R_k(s_t,a_t) + r_t \quad \text{and} \quad S_k(s_t,a_t) \leftarrow S_k(s_t,a_t) + r_t^2$$

if $\nu_k(s_t,a_t) \geq N^+_k(s_t,a_t)$ **then** ▷ Stopping condition of episode k

Increment time $t \leftarrow t+1$ and **Break**

else

Increment time $t \leftarrow t+1$ and sample action $a_t \sim \pi_k(\cdot|s_t)$.

end if

end while

Update counters, empirical averages and sample variances for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$:

$$N_{k+1}(s,a,s') := N_k(s,a,s') + \nu_k(s,a,s') \text{ and } N_{k+1}(s,a) := N_k(s,a) + \nu_k(s,a) \tag{3.6}$$

$$\widehat{p}_{k+1}(s'|s,a) := \frac{N_k(s,a)}{N^+_{k+1}(s,a)} \cdot \widehat{p}_k(s'|s,a) + \frac{\nu_k(s,a,s')}{N^+_{k+1}(s,a)} \tag{3.7}$$

$$\widehat{r}_{k+1}(s,a) := \frac{N_k(s,a)}{N^+_{k+1}(s,a)} \cdot \widehat{r}_k(s,a) + \frac{R_k(s,a)}{N^+_{k+1}(s,a)} \tag{3.8}$$

$$\widehat{\sigma}^2_{p,k+1}(s'|s,a) := \widehat{p}_{k+1}(s'|s,a)\left(1 - \widehat{p}_{k+1}(s'|s,a)\right) \tag{3.9}$$

$$\widehat{\sigma}^2_{r,k+1}(s,a) := \frac{S_k(s,a)}{N^+_{k+1}(s,a)} + \frac{N_k(s,a)}{N^+_{k+1}(s,a)} \cdot \left(\widehat{\sigma}^2_{r,k}(s,a) + \widehat{r}_k(s,a)\right) - \left(\widehat{r}_{k+1}(s,a)\right)^2 \tag{3.10}$$

end for

UCRLB keeps track of the number of observations of the sequence $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ *strictly before* and *during* episode k (repectively $N_k(s,a,s')$ and $\nu_k(s,a,s')$):

$$\nu_k(s,a,s') := \sum_{t=t_k}^{t_{k+1}-1} \mathbb{1}\left\{s_t = s, a_t = a, s_{t+1} = s'\right\} \quad \text{and} \quad N_k(s,a,s') := \sum_{l=1}^{k-1} \nu_l(s,a,s'). \tag{3.13}$$

UCRLB also keeps track of the number of visits in every state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ before and during episode k (repectively $N_k(s,a)$ and $\nu_k(s,a)$):

$$\nu_k(s,a) := \sum_{s' \in \mathcal{S}} \nu_k(s,a,s') \quad \text{and} \quad N_k(s,a) := \sum_{s' \in \mathcal{S}} N_k(s,a,s'). \tag{3.14}$$

In Alg. 5, $\nu_k(s,a,s')$ (resp. $\nu_k(s,a)$) is incremented after *every new visit* in (s,a,s') (resp. (s,a)) while $N_k(s,a,s')$ (resp. $N_k(s,a)$) is updated at the *end of every episode* using the recurrence relation $N_{k+1}(s,a,s') := N_k(s,a,s') + \nu_k(s,a,s')$ (resp. $N_{k+1}(s,a) := N_k(s,a) + \nu_k(s,a)$), with $N_1(s,a,s') := 0$ by definition (resp. $N_1(s,a) := 0$). Finally, $R_k(s,a)$ (resp. $S_k(s,a)$) denote the cumulative sum of rewards (resp. squared rewards):

$$R_k(s,a) := \sum_{t=t_k}^{t_{k+1}-1} \mathbb{1}\left\{s_t, a_t = s, a\right\} \cdot r_t \quad \text{and} \quad S_k(s,a) := \sum_{t=1}^{t_k-1} \mathbb{1}\left\{s_t, a_t = s, a\right\} \cdot r_t^2. \tag{3.15}$$

Episodes. The stopping condition of episodes implemented in UCRLB *slightly differ* from the stopping condition used in UCRL2. In UCRL2, an episode k stops whenever the algorithm *is about* to play an action a in a state s that already satisfies $\nu_k(s,a) = N_k^+(s,a)$. Action a is therefore *never* played and a *new policy* is computed instead. In UCRL2, for all state-action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\nu_k(s,a) \leq N_k^+(s,a)$ and $\nu_k(s,a) = N_k^+(s,a)$ holds true for *at least one* (s,a). However, it is possible that the equality holds for *several* state-action pairs. In UCRLB, an episode k stops as soon as the action a that has *just* been played (i.e., most recently) e.g., in state s, satisfies $\nu_k(s,a) = N_k^+(s,a)$. Action a is therefore *played* and a *new policy* is computed just *after* that. The reason we modified the doubling scheme of UCRL2 is only to *simplify* the theoretical analysis of the algorithm in the general case where the policy π_k played at episode k may be stochastic. Our stopping condition avoids introducing two actions at time t: the action that "could have been played" (if the episode had not been ended) and the one which is actually played.

Confidence bounds and extended MDP. At the beginning of every episode k, UCRLB uses the *sample means* $\widehat{p}_k$ and $\widehat{r}_k$ as (unbiased) estimators of p and r respectively. These estimators can be *efficiently* updated at the end of every episode using the usual *update rule* of the sample mean (see Eq. 3.7 and 3.8). While UCRL2 relies on *Hoeffding's concentration inequality* (HI) (Boucheron et al., 2013, Chapter 2.6) and *Weissman's concentration inequality* (Weissman et al., 2003, Theorem 2.1) to derive the confidence intervals needed to define the extended MDP $\mathcal{M}_k$ (see Eq. 2.25 and 2.26), UCRLB leverages on *empirical Bernstein's concentration inequality* (EBI) (Audibert et al., 2007; Maurer and Pontil, 2009) to derive the confidence bounds of Eq. 3.1 and 3.2 used in the definition of $\mathcal{M}_k$. EBI is *tighter* than HI (at least for

a sufficiently high number of observations). We recall both inequalities below.

Proposition 3.1 (Hoeffding inequality, Theorem 2.8 of Boucheron et al. (2013))
Let $(X_i)_{1\leq i\leq n}$ *be a collection of independent random variables s.t.* $\forall i \in \{1,...,n\}$, $\mathbb{P}(X_i \in [a_i, b_i]) = 1$ *and* $\mathbb{E}[X_i] = \mu_i$. *Then with probability at least* $1-\delta$ *it holds that*

$$\left|\sum_{i=1}^{n}(X_i - \mu_i)\right| \leq \sqrt{\frac{1}{2}\sum_{i=1}^{n}(b_i - a_i)^2 \ln\left(\frac{2}{\delta}\right)}. \tag{3.16}$$

Proposition 3.2 (Empirical Bernstein inequality, Theorem 1 of Audibert et al. (2009))
Let $(X_i)_{1\leq i\leq n}$ *be a collection of i.i.d. r.v. s.t.* $\forall i \in \{1,...,n\}$, $\mathbb{P}(X_i \in [a, b]) = 1$ *and* $\mathbb{E}[X_i] = \mu$. *Then with probability at least* $1-\delta$ *it holds that*

$$\left|\sum_{i=1}^{n}(X_i - \mu)\right| \leq \sqrt{\frac{2V_n(X)\ln(3/\delta)}{n}} + \frac{3(b-a)\ln(3/\delta)}{n},$$

wher $V_n(X)$ *is the* population variance[1]*:* $V'_n(X) := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{i=1}^{n}X_i\right)^2$.

For any state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, UCRLB uses EBI to bound $|\widehat{p}_k(s'|s,a) - p(s'|s,a)|$ for all s' w.h.p.. The ℓ_1-deviation $\|\widehat{p}_k(\cdot|s,a) - p(\cdot|s,a)\|_1 = \sum_{s'\in\mathcal{S}}|\widehat{p}_k(s'|s,a) - p(s'|s,a)|$ between the empirical and true transition probability is bounded (w.h.p.) by *taking a union bound* over all $s' \in \mathcal{S}$ and summing. Instead, UCRL2 uses a variant of Hoeffding's bound derived by Weissman et al. (2003) that directly bounds the ℓ_1-deviation. The use of EBI significantly improves the learning performances (see Sec. 3.4). Notice that Lattimore and Hutter (2012); Dann and Brunskill (2015); Lattimore and Hutter (2014) already proposed variants of UCRL2 that leverages on EBI. However, Lattimore and Hutter (2012, 2014) introduced and analysed their algorithm in the discounted setting (when a discount factor γ is given as input to the algorithm, see Sec. 2.1.3) while Dann and Brunskill (2015) focused on the finite horizon setting (when an horizon H is given as input to the algorithm, see Sec. 2.1.3). They both proved a bound on the *sample complexity* while we will analyse the *regret* of UCRLB.

Extra multiplicative factors appear in the logarithmic terms of (3.1) and (3.2) compared to the bound of Prop. 3.2. This is due to the use of *union bounds* (see Sec. 3.5 for more details). In Alg. 5, the *population variances* of $\widehat{p}_k(s'|s,a)$ and $\widehat{r}_k(s,a)$ are denoted by $\widehat{\sigma}^2_{p,k}(s'|s,a)$ and $\widehat{\sigma}^2_{r,k}(s,a)$ respectively. The estimated transition probability $\widehat{p}_k(s'|s,a)$ correspond to the sample mean of i.i.d. Bernoulli r.v. with mean $p(s'|s,a)$ and therefore the population variance

[1] Unlike the *sample variance* $V'_n(X) := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{i=1}^{n}X_i\right)^2$, the *population variance* $V_n(X)$ is a biased estimator of the true variance. The two estimators are equal up to a multiplicative factor $n/(n-1)$ called *"Bessel's correction"*: $V'_n(X) := \frac{n}{n-1}V_n(X)$.

can be easily computed as $\widehat{\sigma}^2_{p,k}(s'|s,a) := \widehat{p}_k(s'|s,a)\,(1-\widehat{p}_k(s'|s,a))$ (3.9). The population variance of the reward can be computed recursively at the end of every episode (3.10):

$$\begin{aligned}\widehat{\sigma}^2_{r,k+1}(s,a) &:= \frac{1}{N^+_{k+1}(s,a)}\left(\sum_{l=1}^{k} S_l(s,a)\right) - (\widehat{r}_{k+1}(s,a))^2 \\ &= \frac{S_k(s,a)}{N^+_{k+1}(s,a)} + \frac{N_k(s,a)}{N^+_{k+1}(s,a)}\left(\widehat{\sigma}^2_{r,k}(s,a) + (\widehat{r}_k(s,a))^2\right) - (\widehat{r}_{k+1}(s,a))^2\,.\end{aligned}$$

The extended MDP $\mathcal{M}_k$ is defined by the compact sets $B^k_r(s,a)$ (3.4) and

$$B^k_p(s,a) := \left\{p \in \Delta_S :\ p(s') \in B^k_p(s,a,s'),\ \forall s' \in \mathcal{S}\right\} \tag{3.17}$$

where $B^k_p(s,a,s')$ is defined in Eq. 3.3. UCRLB uses the known bound $r_{\max}$ on the reward in order to construct the confidence intervals $B^k_r(s,a)$[2] (Eq. 3.2 and 3.4).

Like UCRL2, UCRLB relies on *extended value iteration* (EVI) to find such an approximate solution (3.5). More details are given in the next section.

3.1.2 Extended value iteration

Like in UCRL2 (Sec. 2.2.3), the purpose of EVI is to find an approximate optimal policy of the extended MDP $\mathcal{M}_k$.[3] EVI is not the only algorithm able to solve this problem. For example, Lattimore and Szepesvári (2018, Section 37.3.1) describe how to solve this problem using the *ellipsoid method*. In this section (as well as in the whole thesis), we will focus exclusively on EVI. We recall that EVI is an intance of value iteration (Alg. 3) with an extended optimal Bellman operator $\mathcal{L}_k$ given as input, namely

$$\forall v \in \mathbb{R}^S,\ \forall s \in \mathcal{S},\quad \mathcal{L}_k v(s) := \max_{a\in\mathcal{A}_s}\left\{\max_{r\in B^k_r(s,a)}\{r\} + \max_{p\in B^k_p(s,a)}\{p^\intercal v\}\right\}. \tag{3.18}$$

The inner optimization problem $\max_{p\in B^k_p(s,a)}\{p^\intercal v\}$ is a *linear programming* (LP) problem since $p \longmapsto p^\intercal v$ is linear and $B^k_p(s,a)$ is only defined by linear constraints on p. It is possible to use a generic solver to find the solution of this problem. However, given that we need to solve SA different LP (one for every state-action pair) with the *same objective function* and with very simple constraints (the sets $B^k_p(s,a,s')$ are real intervals), it is computationally more *efficient* to first *sort* the vector v and then use the LPROBA algorithm described in Sec. 3.1.3 below. If $u := \textsc{Sort}(v)$ is the vector v sorted in descending order, then (3.18) can be re-written:

$$\mathcal{L}_k v(s) := \max_{a\in\mathcal{A}_s}\left\{\max_{r\in B_r(s,a)}\{r\} + \textsc{LProba}\Big(u, \left(B^k_p(s,a,s')\right)_{s'\in\mathcal{S}}\Big)\right\}. \tag{3.19}$$

[2] If UCRLB is given as input an (s,a)-dependent range $[r_{\min}(s,a), r_{\max}(s,a)]$, it is straighforward to adapt Eq. 3.2 and 3.4 in order to take advantage of this additional knowledge.

[3] It is sufficient to find a $r_{\max}/t_k$-approximation of optimization problem (2.35) in order to derive regret guarantees, see Sec.chap:ucrlb:sec:regret.proof.

Extended optimality equation. Since $B_p^k(s,a)$ is a polytope, $\mathcal{L}_k$ can be interpreted as an optimal Bellman operator with finitely many actions (see Sec. 2.1.5). Then, a sufficient condition to apply Prop. 2.4 (guaranteeing existence of a solution to the Bellman optimality equation) is to show that $\mathcal{M}_k$ is *weakly-communicating*. Since the true MDP M is *communicating* by assumption, if $M \in \mathcal{M}_k$ (which holds with high probability as will be shown later, see Prop. 3.1), then $\mathcal{M}_k$ is communicating as well (and therfeore weakly-communicating).

Even when $M \notin \mathcal{M}_k$, $\mathcal{M}_k$ is still communicating because for all 3-tuple (s,a,s'), there exists $q(\cdot|s,a) \in B_p^k(s,a)$ such that $q(s'|s,a) > 0$. Indeed, as will be clear in the next section, such $q(\cdot|s,a)$ can always be obtained by running LPROBA on $e_{s'}$ (the s'-th cartesian basis vector). This only works because $B_p^k(s,a,s')$ are real intervals. However, in some problems it is possible that some transitions $p(s'|s,a)$ of the true MDP are perfectly known beforehand. To remove the burden of learning these specific transitions (s,a,s'), we would like to restrict the corresponding intervals $B_p(s,a,s')$ to a *singleton*, potentially making $\mathcal{M}_k$ not communicating. In this case, it is preferable to expand the singletons by $\pm 1/t_k$, thus ensuring that $\mathcal{M}_k$ is communicating (no matter whether $M \in \mathcal{M}_k$ or not), while forbidding values too distant from $p(s'|s,a)$.

Convergence of EVI. Jaksch et al. (2010, Section 3.1.3) showed that assumption 2 of Prop. 2.6 (guaranteeing convergence of value iteration) always holds for UCRL2 (aperiodicity of the transition matrices encountered in EVI). This assumption also holds with the confidence sets $B_p^k(s,a)$ defined in Eq. 3.17. As we just mentioned, we might be tempted to reduce $B_p^k(s,a)$ to singletons, potentially violating assumption 2 of Prop. 2.6. To overcome this issue, we apply the aperiodicity transformation presented in Sec. 2.2, with aperiodicity coefficient α arbitrarily set to 0.9 The corresponding *aperiodic* optimal Bellman operator $\mathcal{L}_\alpha^k$ can be computed using the expression below.

$$\mathcal{L}_\alpha^k v(s) := \max_{a \in \mathcal{A}_s} \left\{ \max_{r \in B_r(s,a)} \{r\} + \alpha \cdot \text{LPROBA}\Big(u, \big(B_p^k(s,a,s')\big)_{s' \in \mathcal{S}}\Big) \right\} + (1-\alpha) \cdot v(s). \quad (3.20)$$

Similarly, we denote the aperiodic extended MDP $\mathcal{M}_\alpha^k$. Assumption 1 of Prop. 2.6 holds and so EVI converges. Prop. 2.7 also holds i.e.,[4]

$$|g_k - g_k^*| \leq \varepsilon_k/2 := \frac{r_{\max}}{2t_k} \text{ (where } g_k^* \text{ is the optimal gain of } \mathcal{M}_k) \quad (3.21)$$

$$\text{and } \|\mathcal{L}_\alpha^k h_k - h_k - g_k e\|_\infty \leq \varepsilon_k := \frac{r_{\max}}{t_k}. \quad (3.22)$$

Greedy policy. It is very likely that *several* greedy policy exist, especially at the beginning of the learning process when the uncertainty on p and r is high (so that many actions are equally optimistically optimal). When there is *ambiguity* on which action to play (there can be several optimal policies), UCRL2 break ties *arbitrarily* by playing only one of the actions (see Eq. 2.36). Thus the policy is always deterministic. The choice of the greedy policy that will be executed during the episode does not seem to impact the regret bound. Nevertheless, since all policies are in some sense *equivalent*, it is reasonable to play them

[4] We recall that the optimal gains of $\mathcal{M}_\alpha^k$ and $\mathcal{M}_k$ are equal (denoted by g_k^*), see Sec. 2.2.

Algorithm 6 Greedy operator used in UCRLB ($\mathcal{G}_\alpha^k$)

Input: Vector $v \in \mathbb{R}^S$, confidence sets $B_r(s,a)$ and $B_p(s,a,s')$ for all (s,a,s'), aperiodicity coefficient $\alpha \in]0,1]$
Output: Policy $\pi \in \Pi^{\mathrm{SR}}$

for $s \in \mathcal{S}$ **do** ▷ This loop can be parallelized to speed up running time
  $\mathcal{A}^+(s) := \operatorname{Arg\,max}_{a \in \mathcal{A}_s} \left\{ \max_{r \in B_r(s,a)} \{r\} + \alpha \cdot \textsc{LProba}\Big(u_n, (B_p(s,a,s'))_{s' \in \mathcal{S}}\Big) \right\}$
  for $a \in \mathcal{A}$ **do**
    if $a \in \mathcal{A}^+(s)$ **then**
      Set $\pi(a|s) := \frac{1}{|\mathcal{A}^+(s)|}$ ▷ All greedy actions are played with equal probability
    else
      Set $\pi(a|s) := 0$
    end if
  end for
end for

with equal probability in order to have a more *balanced exploration*. The implementation of the greedy operator $\mathcal{G}_k^\alpha$ given as input to EVI (Eq. 3.5) is reported in Alg. 6. Our goal in considering randomized policies is not just to artificially complexify the analysis but also to generalize UCRL2's analysis. This will be needed in Chap. 5 for example. It could also be useful for future work where no deterministic policy is optimal.

3.1.3 Linear Programming for extended value iteration

The detailed pseudo-code of LProba is reported in Alg. 7. Given an input vector $v \in \mathbb{R}^S$ and S intervals $([a_i, b_i])_{1 \le i \le S}$ satisfying $1 \ge b_i \ge a_i \ge 0$ and $\sum_{i=1}^S a_i \le 1 \le \sum_{i=1}^S b_i$, LProba solves the following LP:

$$\text{s.t.} \quad \begin{cases} \max \{p^\intercal v\} \\ \sum_{i=1}^S p_i = 1 \\ a_i \ge p_i \ge b_i, \forall i \in \mathcal{S} \end{cases} \tag{3.23}$$

The vector v is assumed to be *sorted* in decreasing order i.e., $v_1 \ge v_2 \ge \cdots \ge v_S$, which simplifies the resolution. The assumptions that $\sum_{i=1}^S a_i \le 1 \le \sum_{i=1}^S b_i$ and $1 \ge b_i \ge a_i \ge 0$ ensure that the *feasible* region defined by the constraints is non-empty. These assumptions are always met in UCRLB because $\widehat{p}(\cdot|s,a) \in B_p^k(s,a)$ by construction, $0 \le \widehat{p}(s'|s,a) \le 1$ for all $s' \in \mathcal{S}$, and $\sum_{s' \in \mathcal{S}} \widehat{p}(s'|s,a) = 1$. Alg. 7 was first introduced by Dann and Brunskill (2015) (the validity of the algorithm is proved in their Appendix A). The idea is to initialize p_i to its minimum value a_i for all $i \in \{1, \ldots, S\}$ and then allocate the *remaining probability mass* $1 - \sum_{i=1}^S a_i$ to p_1 which corresponds to the maximal value v_1. If there is still some probability mass left, it is assigned to p_2 (which corresponds to the second maximal value v_2) and so on in decreasing order until $\sum_{i=1}^S p_i = 1$ (LProba is therefore an instance of "greedy" procedure).

Algorithm 7 Linear Programming for probability maximization (LPROBA)

Input: A vector $v \in \mathbb{R}^S$ sorted in decreasing order $v(1) \geq v(2) \geq \cdots \geq v(S)$, S closed intervals $([a_i, b_i])_{1 \leq i \leq S}$ s.t. $1 \geq b_i \geq a_i \geq 0$ and $\sum_{i=1}^S a_i \leq 1 \leq \sum_{i=1}^S b_i$
Output: A scalar w

1: Set $w_0 := \sum_{i=1}^S a_i \times v(i)$, $\Delta_0 := 1 - \sum_{i=1}^S a_i$ and $i := 1$ ▷ Initialization
2: **while** $\Delta_{i-1} > 0$ **do** ▷ Main loop
3: Set $\delta_i := \min\{\Delta_{i-1},\ b_i - a_i\}$
4: Update $w_i \leftarrow w_{i-1} + \delta_i \times v(i)$ ▷ Assign allowed weights to highest values of v first
5: Update $\Delta_i \leftarrow \Delta_{i-1} - \delta_i$
6: Increment $i \leftarrow i + 1$
7: **end while**
8: Set $w := w_{i-1}$

Computational complexity. LPROBA terminates after at most S iterations. Therefore, the worst-case complexity of a single iteration of EVI is $O(S^2A + S\ln(S))$ where the $S\ln(S)$ term appears because of the sorting of v_n (the input vector of Alg. 7 should be sorted). Fotunately, the loop over states (line 9 of Alg. 3) can be parallelized, reducing the time complexity to $O(SA + S\ln(S))$. This is of the same *order of magnitude* as for value iteration (with discrete instead of compact action spaces) which has a computational complexity of order $O(S^2A)$ per iteration and time complexity $O(SA)$ when parallelized. Value iteration usually converges *exponentially fast* (Schweitzer and Federgruen, 1979) and so EVI is *computationally efficient*.

3.2 Gain-optimism in UCRLB

UCRLB implements the OFU principle. More precisely, it is *gain-optimistic* meaning that the optimal gain g_k^* of the extended MDP $\mathcal{M}_k$ is (w.h.p.) *bigger than or equal to* the optimal gain g^* of the true MDP (at every episode k). As briefly hinted in Sec. 2.2.3, this property is *essential* to guarantee a good *exploration-exploitation trade-off*, and more precisely to derive *near-optimal minimax regret bounds* (see Sec. 3.5). In this section we *formally* prove that UCRLB is gain-optimistic.

3.2.1 A new argument: optimistic Bellman operator

The way that optimism is proved in UCRL2 (Jaksch et al., 2010) is by showing that the true MDP M *belongs to* $\mathcal{M}_k$ w.h.p., which automatically implies that $g_k^* \geq g^*$ w.h.p. (see Sec. 2.1.5 and the equivalence between bounded parameter MDP and extended MDP). This *all-or-none* argument seems very *restrictive*. Indeed, to bound the regret it is *sufficient* to show that $g_k^* \geq g - \eta$ provided η is sufficiently small (the impact on the regret is not bigger than $\eta \cdot T$). Yet, a small perturbation in the definition of the extended MDP may cause the true MDP to be excluded and the argument of Jaksch et al. (2010) no longer applies. This would suggest that the regret can no longer be bounded which is rather unexpected. The difference $g_k^* - g^*$ should intuitively vary continuously as the extended MDP changes. In this section, we present a *new proof* of optimism that only relies on the properties of the *optimal*

Bellman operator of the extended MDP. We no longer require that the true MDP belongs to the extended MDP although it is a *sufficient condition* to apply our proof (our proof is therefore more general). We operate a *paradigm shift* in the way to prove (near) gain-optimism and to *interpret* the extended MDP: we show that what matters is not the inclusion of the true MDP in the corresponding bounded parameter MDP, but only the relationship between the Bellman operator of the extended MDP and the one of the true MDP. One might argue that this *change of perspective* does not result in a much different *implementation* since in the end, the policy executed is always the optimal policy of an extended MDP that will most likely contain the true MDP. But in some situations (see for example Chap. 5), our new argument allows to *restrict* the extended MDP (smaller confidence intervals that do not necessarily include the true parameters of the MDP). The optimism is therefore *tighter* which results in an improvement of the performance of the algorithm.

Our proof relies on the following very simple theorem proved by Puterman (1994):

Proposition 3.3 (Theorem 8.4.1 of Puterman (1994)[5])

Let L be the optimal Bellman operator of an MDP with S states and assume that the optimal gain g^* of this MDP is state-independent. If there exists a scalar g and a vector $h \in \mathbb{R}^S$ such that $Lh \geq h + ge$ (where $e = (1, \dots, 1)^\intercal$ is the S-dimensional vector of all ones), then $g^* \geq g$.

Let (g^*, h^*) be a solution of the Bellman optimality equation of the true MDP i.e., $Lh^* = h^* + g^* e$ where L is the optimal Bellman operator of the true MDP. Using Prop. 3.3, if we can show that $\mathcal{L}_k h^* \geq h^* + g^* e$ then $g_k^* \geq g^*$. Since $h^* + g^* e = Lh^*$, this is equivalent to showing that $\mathcal{L}_k h^* \geq Lh^*$. In other words, in order to prove gain-optimism we only need to show that the optimal Bellman operator of the extended MDP $\mathcal{M}_k$ is *optimistic* w.r.t. to the optimal Bellman operator of the true MDP, when applied to one *optimal bias vector*. Trivially, if the true MDP belongs to the extended MDP then this condition is satisfied. More generally, if there exists $\eta \geq 0$ such that $\mathcal{L}_k h^* \geq Lh^* - \eta e = h^* + (g^* - \eta)e$, then by applying Prop. 3.3 we have that $g_k^* \geq g^* - \eta$.

We call the statement of Lem. 3.3 the *"dominance property"* of operator L. As we just showed, it plays a key role in ensuring gain-optimism. It is also a much more "refined" argument than the one usually used (*"inclusion"* argument: $M \in \mathcal{M}_k$). In this thesis we will make an extensive use of this property and prove similar results for other operators than L.

3.2.2 Proof of optimism with concentration inequalities

We now prove that $M \in \mathcal{M}_k$ w.h.p. (Thm. 3.1) which implies that $\mathcal{L}_k h^* \geq Lh^*$ w.h.p. Thm. 3.1 is similar to Lemma 17 proved by Jaksch et al. (2010) except that we bound the

[5]The theorem proved by Puterman (1994) is more general but we only need this simplified version for our purpose.

probability of event $\bigcup_{k\geq 1}\{M \notin \mathcal{M}_k\}$ while they only bound the probability of $\{M \notin \mathcal{M}_k\}$ by a term that decreases with t_k. They then take a union bound in the regret proof. Thm. 3.1 will simplify the regret analysis and our proof allows to use confidence bounds that only grows logarithmically with N_k instead of t_k (Eq. 3.1 and 3.2). As a consequence, the confidence bounds associated to (s,a) *do not increase* over time when (s,a) is not visited (they remain constant) and UCRLB will not visit all (s,a) *infinitely often*. This is not surprising since we want to show a *uniform high probability regret bound* as opposed to a *uniform expected regret bound*. A uniform expected regret bound requires to visit all state-action pairs infinitely often and so to have a term t_k in the logarithm of the confidence bounds (3.1) and (3.2). For a more thorough discussion on this, see for example Dann et al. (2017, Section 4.1).

Theorem 3.1

The probability that there exists $k \geq 1$ s.t. the true MDP M does not belong to the extended MDP $\mathcal{M}_k$ defined by Eq. 3.3 and 3.4 is at most $\frac{\delta}{3}$, that is

$$\mathbb{P}\left(\exists k \geq 1,\ s.t.\ M \notin \mathcal{M}_k\right) \leq \frac{\delta}{3}.$$

Proof. We want to bound the probability of event $E := \bigcup_{k=1}^{+\infty}\{M \notin \mathcal{M}_k\}$. As explained by Lattimore and Szepesvári (2018, Section 4.4), when (s,a) is visited for the n-th times, the reward that we observe is the n-th element of an infinite sequence of i.i.d. r.v. lying in $[0, r_{\max}]$ with expected value $r(s,a)$. Similarly, the next state that we observe is the n-th element of an infinite sequence of i.i.d. r.v. lying in $\mathcal{S}$ with probability density function (pdf) $p(\cdot|s,a)$. In Alg. 5, we defined the sample means $\widehat{p}_k$ and $\widehat{r}_k$ (Eq. 3.7 and 3.8), and the confidence intervals B_p^k and B_r^k (Eq. 3.3 and 3.4) as depending on k. Actually, this quantities depends only on the first $N_k(s,a)$ elements of the infinite i.i.d. sequences that we just mentioned. For the rest of the proof, we will therefore slightly change our notations and denote by $\widehat{p}_n(s'|s,a)$, $\widehat{r}_n(s,a)$, $B_p^n(s'|s,a)$ and $B_r^n(s,a)$ the sample means and confidence intervals after the first n visits in (s,a). Thus, the r.v. that we denoted by $\widehat{p}_k$ in Alg. 5 actually corresponds to $\widehat{p}_{N_k(s,a)}$ with our new notation (and similarly for $\widehat{r}_k$, B_p^k and B_r^k). This change of notation will make the proof easier.

$M \notin \mathcal{M}_k$ means that there exists $k \geq 1$ s.t. either $p(s'|s,a) \notin B_p^{N_k(s,a)}(s,a,s')$ or $r(s,a) \notin B_r^{N_k(s,a)}(s,a)$ for at least one $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$. This means that there exists at least one value $n \geq 0$ s.t. either $p(s'|s,a) \notin B_p^n(s,a,s')$ or $r(s,a) \notin B_r^n(s,a)$. As a consequence we have the following inclusion

$$E \subseteq \bigcup_{s,a}\bigcup_{n=0}^{+\infty}\{r(s,a) \notin B_r^n(s,a)\} \cup \bigcup_{s'}\left\{p(s'|s,a) \notin B_p^n(s,a,s')\right\} \tag{3.24}$$

Using Boole's inequality we thus have:

$$\mathbb{P}(E) \leq \sum_{s,a}\sum_{n=0}^{+\infty}\left(\mathbb{P}\left(r(s,a) \notin B_r^n(s,a)\right) + \sum_{s'}\mathbb{P}\left(p(s'|s,a) \notin B_p^n(s,a,s')\right)\right) \tag{3.25}$$

Let's fix a 3-tuple $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and define for all $n \geq 0$

$$\epsilon_{p,n}^{sas'} := \widehat{\sigma}_{p,n}(s'|s,a)\sqrt{\frac{2\ln\left(30S^2A(n^+)^2/\delta\right)}{n^+}} + \frac{3\ln\left(30S^2A(n^+)^2/\delta\right)}{n^+} \tag{3.26}$$

$$\epsilon_{r,n}^{sa} := \widehat{\sigma}_{r,n}(s,a)\sqrt{\frac{2\ln\left(30SA(n^+)^2/\delta\right)}{n^+}} + \frac{3r_{\max}\ln\left(30SA(n^+)^2/\delta\right)}{n^+} \tag{3.27}$$

where $\widehat{\sigma}_{p,n}(s'|s,a)$ and $\widehat{\sigma}_{r,n}(s,a)$ denote the population variances obtained with the first n samples. It is immediate to verify that $\epsilon_{p,n}^{sas'} \leq \beta_{p,n}^{sas'}$ and $\epsilon_{r,n}^{sa} \leq \beta_{r,n}^{sa}$ a.s. (see Eq. 3.1 and 3.2 with $N_k(s,a)$ replaced by n). Using Prop. 3.2 we have that for all $n \geq 1$:

$$\mathbb{P}\left(|p(s'|s,a) - \widehat{p}_n(s'|s,a)| \geq \beta_{p,n}^{sas'}\right) \leq \mathbb{P}\left(|p(s'|s,a) - \widehat{p}_n(s'|s,a)| \geq \epsilon_{p,n}^{sas'}\right) \leq \frac{\delta}{10n^2S^2A} \tag{3.28}$$

$$\mathbb{P}\left(|r(s,a) - \widehat{r}_n(s,a)| \geq \beta_{r,n}^{sa}\right) \leq \mathbb{P}\left(|r(s,a) - \widehat{r}_n(s,a)| \geq \epsilon_{r,n}^{sa}\right) \leq \frac{\delta}{10n^2SA} \tag{3.29}$$

Note that when $n = 0$ (i.e., when there hasn't been any observation of (s,a)), $\epsilon_{p,0}^{sas'} \geq 1$ and $\epsilon_{r,0}^{sa} \geq r_{\max}$ so $\mathbb{P}\left(|p(s'|s,a) - \widehat{p}_0(s'|s,a)| \geq \epsilon_{p,0}^{sas'}\right) = \mathbb{P}\left(|r(s,a) - \widehat{r}_0(s,a)| \geq \epsilon_{r,0}^{sa}\right) = 0$ by definition. Since in addition (also by definition)

$$B_p^n(s,a,s') \subseteq \left[\widehat{p}_n(s'|s,a) - \beta_{p,n}^{sas'}, \widehat{p}_n(s'|s,a) + \beta_{p,n}^{sas'}\right] \text{ (see Eq. 3.3)}$$

and

$$B_r^n(s,a) \subseteq \left[\widehat{r}_n(s,a) - \beta_{r,n}^{sa}, \widehat{r}_k(s,a) + \beta_{r,n}^{sa}\right] \text{ (see Eq. 3.4)}$$

we conclude that for all $n \geq 1$

$$\mathbb{P}\left(p(s'|s,a) \notin B_p^n(s,a,s')\right) \leq \frac{\delta}{10n^2S^2A} \quad \text{and} \quad \mathbb{P}\left(r(s,a) \notin B_r^n(s,a)\right) \leq \frac{\delta}{10n^2SA}$$

and these probabilities are equal to 0 if $n = 0$. Plugging these inequalities into Eq. (3.25) we obtain:

$$\mathbb{P}\left(\exists T \geq 1, \exists k \geq 1 \text{ s.t.} M \notin \mathcal{M}_k\right) \leq \sum_{s,a}\left(0 + \sum_{n=1}^{+\infty}\left(\frac{\delta}{10n^2SA} + \sum_{s'}\frac{\delta}{10n^2S^2A}\right)\right) = \frac{2\pi^2\delta}{60} \leq \frac{\delta}{3}$$

which concludes the proof. ■

3.3 Bounding the optimistic bias of UCRLB: diameter and refinements

At every episode $k \geq 1$, EVI returns both a policy π_k, a gain g_k and a bias vector h_k. We refer to g_k as the (near) *optimistic gain* and h_k as the (near) *optimistic bias vector*. Note that the optimistic gain g_k is indeed (near) optimistic i.e., satisfies $g_k \geq g_k^* - \varepsilon_k/2 \geq g^* - \varepsilon_k/2$ (by combining Eq. 3.21 with the results of Sec. 3.2), while the optimistic bias vector h_k does not necessarily satisfy $h_k \gtrsim h^*$.[6] Actually, h_k is defined up to a *constant shift*. Nevertheless,

[6] (g^*, h^*) is a solution to the Bellman optimality equation of the true MDP M.

we will use the terminology "optimistic bias" to refer to h_k.

We will see in Sec. 3.5 that the *shape* of the optimistic bias h_k has a substantial impact on the regret analysis. In this section we focus on bounding the *range* of h_k i.e., bounding $sp(h_k)$.

3.3.1 Diameter

In this section we bound $sp(h_k)$ using the concept of *diameter* of an MDP (Def. 2.6). We start by recalling an important result proved by Bartlett and Tewari (2009):

Proposition 3.4 (Theorem 4 of Bartlett and Tewari (2009))

Let M be a communicating MDP with non-negative rewards and (g^*, h^*) a solution of the Bellman optimality equation i.e., $Lh^* = h^* + g^*e$. For any states s and s' and any stationary policy $\pi \in \Pi^{SR}$, we have:

$$h^*(s') - h^*(s) \leq g^* \cdot \mathbb{E}^\pi[\tau(s') - 1 | s_1 = s]$$

where $\tau(s') := \inf\{t \geq 1 : s_t = s'\}$ is the first hitting time *of s'.*

As a direct consequence of Prop. 3.4, we have the following corollary:

Corollary 3.1

Under the same assumptions as Prop. 3.4, the range of h^* can be bounded as $sp(h^*) \leq g^* D$ where D is the diameter of M.

Proof. By definition

$$sp(h^*) := \max_{s \in \mathcal{S}}\{h^*(s)\} - \min_{s \in \mathcal{S}}\{h^*(s)\} = \max_{s,s'}\{h^*(s') - h^*(s)\} \leq g^* \cdot \max_{s,s'} \mathbb{E}^\pi[\tau(s') - 1 | s]$$

where the last inequality is a direct consequence of Prop. 3.4 and the fact that $g^* \geq 0$. ∎

Let's first assume that EVI computes an exact solution (g_k^*, h_k^*) of the Bellman optimality equation $\mathcal{L}_\alpha^k h_k^* = h_k^* + g_k^*$. According to Cor. 3.1 we have $sp(h_k^*) \leq g_k^* \cdot D_\alpha^k$ (where D_α^k is the diameter of $\mathcal{M}_\alpha^k$). We now need to relate the parameters of the extended MDP g_k^* and D_α^k with the parameters of the true MDP.

Bounding g_k^*. The optimal gain g_k^* is always *smaller* than $r_{\max}$ by definition but can be *as big as* $r_{\max}$. For example at the beginning of the learning process, the *uncertainty is maximal* in all state-action pairs and so all optimistic rewards are set to $r_{\max}$ implying $g_k^* = r_{\max}$. But even after a rather long exploration phase, it is sufficient that one state-action pair is *poorly visited* to have $g_k^* = r_{\max}$. This is because the gain is a *global* quantity of the MDP (as opposed to the *local* rewards). As long as at least one state-action pair (s, a) is poorly visited, UCRLB

will *optimistically* set the reward $r_k(s,a) \leftarrow r_{\max}$ and transition $p_k(s|s,a) \leftarrow 1$ causing g_k^* to be maximal (if one policy of a communicating MDP loops on a single state with reward $r_{\max}$, then the optimal gain is $r_{\max}$ *independently* of the rest of the MDP as shown in Theorem 8.3.2 of Puterman (1994)). Since in general we *cannot control* how UCRLB *explores* the MDP, the tightest upper-bound that we can derive for g_k^* is $r_{\max}$.

Bounding D_k. Jaksch et al. (2010, Section 4.3.1) showed that the diameter D_k of the extended MDP constructed by UCRL2 at every episode $k \geq 1$ is *smaller than or equal to* the diameter of the true MDP D. Their proof relies on the same argument used to prove optimism (inclusion argument): since $M \in \mathcal{M}_k$ w.h.p., the shortest path to go from any state to any other state is always shorter in the extended MDP and so $D_k \leq D$. We can use the same argument in our case to show that under the same event as in Thm. 3.1, $D_k \leq D$ (where D_k is the diameter of $\mathcal{M}_k$). However, as already argued in Sec. 3.2.1, this *"inclusion"* argument is rather *restrictive* and *"non-smooth"*. Proving a *more general result* helps provide better *intuitions* and opens the way for *extensions*. Similarly to what we did in Sec. 3.2, we generalize the argument of Jaksch et al. (2010) by showing that it is sufficient to analyze the relationship between the Bellman operator of the extended MDP $\mathcal{M}_k$ and the true MDP M to connect D_k with D. We no longer require that $M \in \mathcal{M}_k$. The following proposition is another declination of the *dominance property* (analogue to Prop. 3.3) in the context of (generalized) *stochastic shortest path problems* (see Sec. 2.1.4).

Proposition 3.5 (Theorem 7.3.2. of Puterman (1994))

Let $M = \langle \mathcal{S}, \mathcal{A}, r, p \rangle$ be a communicating *MDP (finite or compact $\mathcal{A}$) with* negative rewards. *For any state $s \in \mathcal{S}$, consider the* Bellman shortest path operator $L_{\mapsto s}$ *with maximal non-positive fixed point $h^*_{\mapsto s}$ (see Prop. 2.8). If there exists $h \in \mathbb{R}^S$ such that $h \leq 0$ and $L_{\mapsto s} h \geq h$ then $h^*_{\mapsto s} \geq h$.*

Let's consider $M' = \{\mathcal{S}, \mathcal{A}, r', p\}$ the MDP with identical transition probabilities p than the true MDP M and rewards r' equal to -1 everywhere (for all state and actions). For all $s \in \mathcal{S}$, denote by $L_{\mapsto s}$ the *Bellman shortest path operator* of M' and $h^*_{\mapsto s'}$ its fixed point (Prop. 3.5). By Prop. 2.8, $-h^*_{\mapsto s'}(s) := \min_{\pi \in \Pi^{\text{SR}}} \mathbb{E}^\pi[\tau(s')|s_1 = s] - 1$ for all $s, s' \in \mathcal{S}$ (see Eq. 2.13), and so by definition $D := \max_{s \in \mathcal{S}} \{\|h^*_{\mapsto s}\|_\infty\}$. Let's denote by $\mathcal{L}^k_{\mapsto s}$ the analogue of $L_{\mapsto s}$ for the extended MDP $\mathcal{M}'_k$ (identical to $\mathcal{M}_k$ with all rewards replaced by -1), and by $h^k_{\mapsto s}$ its maximal non-positive fixed-point. Under the high probability event of Thm. 3.1, $\mathcal{L}^k_{\mapsto s} h^*_{\mapsto s} \geq L_{\mapsto s} h^*_{\mapsto s} = h^*_{\mapsto s}$ and so by Prop. 3.5, $h^k_{\mapsto s} \geq h^*_{\mapsto s}$ ($h^*_{\mapsto s} \leq 0$ by definiton). It follows that $D_k \leq D$.

As in the case of gain-optimism (Sec. 3.2), we see that in order to show that $D_k \leq D$, it is sufficient to prove that $\mathcal{L}^k_{\mapsto s} h^*_{\mapsto s} \geq L_{\mapsto s} h^*_{\mapsto s}$ for all $s \in \mathcal{S}$ (optimism of the Bellman operator on a specific vector). More generally, let's assume that there exists $1 > \eta \geq 0$ such that $\mathcal{L}^k_{\mapsto s} h^*_{\mapsto s} \geq L_{\mapsto s} h^*_{\mapsto s} - \eta e_{|s}$ (where $e_{|s}$ is the S-dimensional vector of all ones, except the s-th coordinate which is zero). Let's define $\mathcal{L}^{k,\eta}_{\mapsto s}$ the analogue of $\mathcal{L}^k_{\mapsto s}$ with all rewards equal to $-1 + \eta$ instead of -1, and $h^{k,\eta}_{\mapsto s}$ the corresponding maximal non-positive fixed point. It is

immediate from the definition of the operators that $\mathcal{L}^k_{\mapsto s} h^*_{\mapsto s} \geq L_{\mapsto s} h^*_{\mapsto s} - \eta e_{|s}$ is equivalent to $\mathcal{L}^{k,\eta}_{\mapsto s} h^*_{\mapsto s} \geq L_{\mapsto s} h^*_{\mapsto s}$. According to Prop. 3.5, we therefore have $h^{k,\eta}_{\mapsto s} \geq h^*_{\mapsto s}$. Since the rewards associated to $\mathcal{L}^{k,\eta}_{\mapsto s}$ are the same for all policies and they only differ from the rewards of $\mathcal{L}^k_{\mapsto s}$ by a multiplicative factor $(1-\eta)$, it is immediate to see that $h^{k,\eta}_{\mapsto s} = (1-\eta) h^k_{\mapsto s}$. In conclusion, $(1-\eta) h^k_{\mapsto s} \geq h^*_{\mapsto s}$ and so $D_k \leq D/(1-\eta)$. The impact of a small perturbation $1 > \eta > 0$ on the diameter is non-linear in η while the impact on the gain is linear (see Sec. 3.2).

Diameter and aperiodicity transformation: D_k vs D^k_α. So far, we have bounded the diameter of the extended MDP $\mathcal{M}_k$, ignoring the aperiodicity transformation. Thm. 2.1 shows how to relate D_k with D^k_α: $D^k_\alpha = D_k/\alpha$. After combining all the inequalities derived in Sec. 3.3.1, we obtain $sp\,(h^*_k) \leq r_{\max} D/\alpha$.

Approximate solution of the optimal Bellman equation. As we showed at the beginning of Sec. 3.3, EVI only computes an approximate solution (g_k, h_k) of the optimality equation i.e., $\|\mathcal{L}^k_\alpha h_k - h_k - g_k e\|_\infty \leq \varepsilon_k$ as opposed to an exact solution (g^*_k, h^*_k) satisfying $\mathcal{L}^k_\alpha h^*_k = h^*_k + g^*_k e$. Jaksch et al. (2010, Section 4.3.1) proved by induction the following proposition (which is a specific case of Thm. 3.3 proved in the next section).

Proposition 3.6

Let L be the optimal Bellman operator of a communicating MDP with diameter D. Consider the sequences of vectors $(v_n)_{n\in\mathbb{N}}$ obtained while executing value iteration (Alg. 3) with operator L and initial vector $v_0 := 0$ as inputs. It holds that for all $\pi \in \Pi^{SR}$, all $s, s' \in \mathcal{S}$ and all $n \geq 0$:

$$v_n(s') - v_n(s) \leq r_{\max} \cdot \mathbb{E}^\pi \left[\tau(s') - 1 | s_1 = s\right] \leq r_{\max} D. \tag{3.30}$$

EVI is run starting from the null vector and so $sp\,(h_k) \leq r_{\max} D^k_\alpha \leq r_{\max} D/\alpha$. Note that in order to apply Prop. 3.6 to the extended MDP, it is essential for the rewards to be contained in $[0, r_{\max}]$.[7]

3.3.2 Refinement of the diameter: travel-budget

The bound $sp\,(h_k) \leq r_{\max} D/\alpha$ derived in Sec. 3.3.1 assumes that while trying to reach a target state, an agent receives *zero* rewards in *all* but the target state (where it receives $r_{\max}$). This can be very *loose* as the agent usually has the opportunity to collect rewards *on the way* to the target state. In this section we introduce a new quantity that better accounts for the reward *discrepancy* in the MDP. We call this new quantity the *travel-budget*[8] and denote it by Λ. We derive theorems analogue to those of Sec. 3.3.1 and show that $sp\,(h_k) \leq \Lambda/\alpha$.

[7] In the original version of UCRL2, (Jaksch et al., 2010) forgot to enforce this constrain.

[8] We acknowledge that Dai and Walter (2019) independently and simultaneously introduced the same quantity with a different name *"maximum expected hitting cost"*.

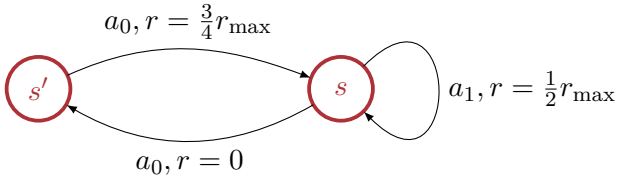


Figure 3.1: Counter-example illustrating the need of $\Pi^{\text{SD}}_{\mapsto s'}$ in Thm. 3.2. Only one action a_0 can be played in s' while two actions a_0, a_1 can be played in s. All transitions are deterministic and M is communicating. It is immediate to verify that the optimal policy corresponds to $\pi^*(s) = a_1$ and moreover $g^* = \frac{1}{2} r_{\max}$, $h^*(s) = 0$ and $h^*(s') = \frac{1}{4} r_{\max}$. We also notice that $\pi^* \notin \Pi^{\text{SD}}_{\mapsto s'}$ and $\mathbb{E}^{\pi^*}\left[\sum_{t=1}^{\tau(s')-1} g^* - r_t \middle| s_1 = s\right] = 0 < \frac{1}{4} r_{\max} = h^*(s') - h^*(s)$ and so (3.33) does not hold.

We first define the set of stationary deterministic policies *reaching a state in finite time* (a.s.) and prove a theorem analogue to the one proved by Bartlett and Tewari (2009) (see Prop. 3.4).

Definition 3.1

For any MDP M, we define for all $s, s' \in \mathcal{S}$

$$\Pi^{SD}_{s \mapsto s'} := \left\{ \pi \in \Pi^{SD} : \ \mathbb{P}^\pi(\tau(s') < +\infty | s_1 = s) = 1 \right\} \tag{3.31}$$

$$\Pi^{SD}_{\mapsto s'} := \bigcap_{s \in \mathcal{S}} \Pi^{SD}_{s \mapsto s'} = \left\{ \pi \in \Pi^{SD} : \ \mathbb{P}^\pi(\tau(s') < +\infty | s_1 = s) = 1, \forall s \in \mathcal{S} \right\} \tag{3.32}$$

where $\tau(s') := \inf\{t \geq 1 : s_t = s'\}$ is the first hitting time *of s'. If M is communicating, then $\Pi^{SD}_{\mapsto s'} \neq \emptyset$ for all $s' \in \mathcal{S}$.*

Proof. We prove the statement by contraposition. If $\mathbb{P}^\pi(\tau(s') = +\infty | s) > 0$ then by the law of total expectations:

$$\begin{aligned}\mathbb{E}^\pi\left[\tau(s') | s_1 = s\right] =& \mathbb{E}^\pi\left[\tau(s') \middle| s_1 = s, \tau(s') < +\infty\right] \cdot \mathbb{P}^\pi(\tau(s') < +\infty | s_1 = s) \\ &+ \underbrace{\mathbb{E}^\pi\left[\tau(s') \middle| s_1 = s, \tau(s') = +\infty\right]}_{=+\infty} \cdot \mathbb{P}^\pi(\tau(s') = +\infty | s_1 = s) = +\infty\end{aligned}$$

Therefore, if $\Pi^{\text{SD}}_{\mapsto s'} = \emptyset$ for at least one $s' \in \mathcal{S}$, then $D = +\infty$. This concludes the proof. ■

Theorem 3.2 (Analogue of Prop. 3.4)

Let M be a communicating MDP with optimal Bellman operator L and $(g^*, h^*) \in \mathbb{R} \times \mathbb{R}^S$ a solution to the optimality equation $h^* + g^* e = L h^*$. For any two states s and s' and any stationary policy $\pi \in \Pi^{SD}_{\mapsto s'}$, we have:

$$h^*(s') - h^*(s) \leq \mathbb{E}^\pi\left[\sum_{t=1}^{\tau(s')-1} g^* - r_t \middle| s_1 = s\right] \tag{3.33}$$

where $\tau(s') := \inf\{t \geq 1 : s_t = s'\}$ is the first hitting time *of s'.*

Proof. The arguments are similar to the one used by Bartlett and Tewari (2009, Theorem 4). The rigorous proof can be found in App. A.1.1. ■

We first notice that Prop. 3.4 can be *deduced* from Thm. 3.2 since when all rewards are non-negative

$$\forall \pi \in \Pi^{\mathrm{SD}}_{\mapsto s'}, \quad \mathbb{E}^{\pi}\left[\sum_{t=1}^{\tau(s')-1} g^* - r_t \middle| s_1 = s\right] \leq g^* \cdot \mathbb{E}^{\pi}\left[\tau(s') - 1 \middle| s_1 = s\right]$$

and if $\pi \notin \Pi^{\mathrm{SD}}_{\mapsto s'}$ then $\mathbb{E}^{\pi}\left[\tau(s') - 1 \middle| s_1 = s\right] = +\infty$ and so the inequality still holds. The difference between $\mathbb{E}^{\pi}\left[\sum_{t=1}^{\tau(s')-1} g^* - r_t \middle| s_1 = s\right]$ and $\mathbb{E}^{\pi}\left[\tau(s') - 1 \middle| s_1 = s\right]$ can be *arbitrarily loose*. For example, when all the rewards are identical, the optimal gain takes the same value and the term on the left handside is 0 while the term on the right handside can be arbitrarily large. When $\pi \notin \Pi^{\mathrm{SD}}_{\mapsto s'}$, the term $\mathbb{E}^{\pi}\left[\sum_{t=1}^{\tau(s')-1} g^* - r_t \middle| s_1 = s\right]$ might be equal to $+\infty$ but when this is the case, Thm. 3.2 still holds and so one might wonder why we need to *restrict attention* to policies belonging to $\Pi^{\mathrm{SD}}_{\mapsto s'}$. In Fig. 3.1 we provide a *counter-example* showing that Thm. 3.2 does not always hold for policies outside $\Pi^{\mathrm{SD}}_{\mapsto s'}$.

Since Thm. 3.2 *refines* Prop. 3.4, we would like to use this theorem to refine the bound on $sp\left(h_k\right)$ derived in Sec. 3.3.1. As we already discussed in Sec. 3.3.1, in general the best upper bound that we have for g^*_k is $r_{\max}$ and so we define the travel-budget as follows:

Definition 3.2

The travel-budget *of a communicating MDP M (denoted Λ) is defined as*

$$\Lambda := \max_{s,s'} \min_{\pi \in \Pi^{SD}} \mathbb{E}^{\pi}\left[\sum_{t=1}^{\tau(s')-1} r_{\max} - r(s_t, a_t) \middle| s_1 = s\right]. \tag{3.34}$$

$\Lambda \geq 0$ and if all the rewards are non-negative, $\Lambda \leq r_{\max} D$.

Proof. The proof is trivial since for all $t \geq 1$, $r_{\max} - r(s_t, \pi(s_t))$ and under the assumption that the rewards are all positive, $-r(s_t, \pi(s_t)) \leq 0$. ■

Notice that in Eq. 4.2 of Def. 3.2, we *do not restrict* the policy space. Instead, we take the minimum over the entire space Π^{SD} and *not* over $\Pi^{\mathrm{SD}}_{\mapsto s'}$. Therefore, $\tau(s')$ might be equal to $+\infty$ with non-zero probability but everything is still well-defined as explained in Sec. 2.1.4. It turns out that despite the counter-example of Fig. 3.1, when replacing g^* by $r_{\max}$ and considering the iterates of value iteration starting from the null vector, this definition is sufficient for our purpose (see Thm. 3.3 below). On Fig. 3.2 we illustrate the difference between $r_{\max} D$ and Λ on a simple MDP.

Similarly to Sec. 3.3.1, we can combine (4.2) with Thm. 3.5 to prove that $\Lambda_k := \Lambda_{\mathcal{M}_k} \leq \Lambda$ for all $k \geq 1$ (where Λ is the travel-budget of the true unknown MDP). Since we no longer restrict the policy space, we can express Λ as a function of the fixed points of some *Bellman shortest path operators* (as we did for D in Sec. 3.3.1).

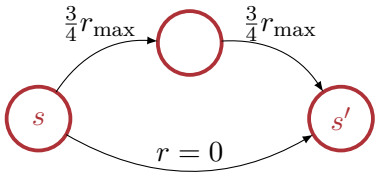


Figure 3.2: Example illustrating the difference between $r_{\max}D$ and Λ. In this example, $\min_\pi \mathbb{E}^\pi \left[\sum_{t=1}^{\tau(s')-1} r_{\max} - r_t | s_1 = s\right] = \frac{1}{2} r_{\max} < r_{\max} = r_{\max} \cdot \min_\pi \mathbb{E}^\pi \left[\tau(s') - 1 | s_1 = s\right]$.

Let's consider $M' = \{\mathcal{S}, \mathcal{A}, r', p\}$ the MDP with identical transition probabilities p than the true MDP M and rewards r' equal to $r - r_{\max} \leq 0$ (for all state and actions). For all $s \in \mathcal{S}$, denote by $L_{\mapsto s}$ the *Bellman shortest path operator* of M' and $h^*_{\mapsto s'}$ its (unique) fixed point (Thm. 3.5). By Prop. 2.8,

$$\forall s' \in \mathcal{S}, \quad -h^*_{\mapsto s}(s') = \min_{\pi \in \Pi^{\mathrm{SD}}(M)} \mathbb{E}^\pi_M \left[\sum_{t=1}^{\tau(s')-1} r_{\max} - r_t \middle| s \right]$$

for all $s, s' \in \mathcal{S}$, and so $\Lambda := \max_s \|h^*_{\mapsto s}\|_\infty$. Similarly to $r_{\max}D$, the travel-budget Λ is obtained as the solution of a *stochastic shortest path* problem where the "lengths" are not always equal to $r_{\max}$ but the actual reward r is *subtracted* i.e., $r_{\max} - r$. Let's denote by $\mathcal{L}^k_{\mapsto s}$ the analogue of $L_{\mapsto s}$ for the extended MDP $\mathcal{M}'_k$ (identical to $\mathcal{M}_k$ with rewards replaced by $r - r_{\max} \leq 0$), and by $h^k_{\mapsto s}$ its fixed point. Under the high probability event of Thm. 3.1, $\mathcal{L}^k_{\mapsto s} h^*_{\mapsto s} \geq L_{\mapsto s} h^*_{\mapsto s} = h^*_{\mapsto s}$ and so by Thm. 3.5, $h^k_{\mapsto s} \geq h^*_{\mapsto s}$. Therefore, $\Lambda_k \leq \Lambda$ and a direct application of Thm. 2.1 shows that $\Lambda^k_\alpha = \Lambda_k / \alpha$ (where Λ^k_α denotes the travel-budget of $\mathcal{M}^k_\alpha$).

Unlike for the diameter, it is difficult to quantify the impact of an η-perturbation of $\mathcal{L}_k$ on the travel-budget Λ_k. This is not surprising since the travel-budget carries much more information about the MDP than the diameter. It also suggests that it is a more relevant quantity to consider for the regret analysis.

We conclude this section with Thm. 3.3 (from which Prop. 3.6 can be deduced).

Theorem 3.3 (Analogue of Prop. 3.6)

Let L be the optimal Bellman operator of a communicating MDP with travel-budget Λ. Consider the sequences of vectors $(v_n)_{n \in \mathbb{N}}$ obtained while executing value iteration (Alg. 3) with operator L and initial vector $v_0 := 0$ as inputs. It holds that for all $\pi \in \Pi^{SR}$, all $s, s' \in \mathcal{S}$ and all $n \geq 0$:

$$v_n(s') - v_n(s) \leq \mathbb{E}^\pi \left[\sum_{t=1}^{\tau(s')-1} r_{\max} - r_t \middle| s_1 = s \right] \leq \Lambda. \tag{3.35}$$

Proof. The argument is similar to the one used by Jaksch et al. (2010, Section 4.3.1). The detailed proof can be found in App. A.1.2. ■

3.4 Regret guarantees for UCRLB

We opened Chap. 3 with a detailed presentation of the *algorithmic structure* of UCRLB (Sec. 3.1). In a nutshell, at every episode k, UCRLB computes an approximate solution $(g_k, h_k) \in \mathbb{R} \times \mathbb{R}^S$ of the Bellman optimality equation of an extended MDP $\mathcal{M}_k$ that is constructed based on *past observations*. In Sec. 3.2 and 3.3 we analysed the *properties* of respectively g_k and h_k. We showed that under a *single* high probability event (Thm. 3.1), $g_k \geq g^*$ and $sp(h_k) \leq \Lambda/\alpha \leq r_{\max} D/\alpha$ where g^*, Λ and D are repsectively the *optimal gain*, *travel-budget* and *diameter* of the *true unknown MDP*. We are now ready to state (and prove) the *main results* of this chapter, namely two *high probability minimax uniform regret bounds* satisfied by UCRLB (Thm. 3.4 and Thm. 3.5). "Uniform" refers to the fact that the high probability bound holds for *all time horizons* $T \geq 1$. Thm. 3.4 and Thm. 3.5 only assume knowledge of the state space $\mathcal{S}$, action space $\mathcal{A}$ and maximal reward $r_{\max}$, even if we already explained in Sec. 3.1 how UCRLB can take advantage of some *additional prior knowledge* about the rewards and transition probabilities. We assume that the initial state s_1 is sampled according to a probability distribution $\mu_1 \in \Delta_S$. For any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, we introduce the notation $\Gamma(s,a)$ for the support of $p(\cdot|s,a)$ i.e.,

$$\Gamma(s,a) := \|p(\cdot|s,a)\|_0 = \sum_{s' \in \mathcal{S}} \mathbb{1}\left\{p(s'|s,a) > 0\right\}.$$

We also denote by Γ the maximal support over all (s,a) i.e., $\Gamma := \max_{s,a \in \mathcal{S} \times \mathcal{A}} \Gamma(s,a)$. Our first regret bound is reported in Thm. 3.4.

Theorem 3.4

There exists a numerical constant $\beta > 0$ such that for any *communicating MDP, with probability at least $1-\delta$, it holds that for all initial state distributions $\mu_1 \in \Delta_S$ and for all time horizons $T > 1$:*

$$\begin{aligned} \Delta(\text{UCRLB}, T) \leq\ & \beta \cdot \max\{r_{\max}, \Lambda\} \sqrt{\left(\sum_{s,a} \Gamma(s,a)\right) T \ln\left(\frac{T}{\delta}\right)} \\ & + \beta \cdot \max\{r_{\max}, \Lambda\} S^2 A \ln\left(\frac{T}{\delta}\right) \ln(T). \end{aligned} \tag{3.36}$$

Jaksch et al. (2010, see Prop. 2.14) showed that up to a multiplicative numerical constant, the regret of UCRL2 is bounded by $r_{\max} DS\sqrt{AT \ln(T/\delta)}$. After noticing that $\Lambda \leq r_{\max} D$ and $\sum_{s,a} \Gamma(s,a) \leq \Gamma SA$ we can simplify the bound in (3.36) as

$$\beta \cdot r_{\max} D \sqrt{\Gamma SAT \ln(T/\delta)} + \beta \cdot r_{\max} DS^2 A \ln(T/\delta) \ln(T)$$

Let's compare the bounds of UCRL2 (Prop. 2.14) and UCRLB (Thm. 3.4) in terms of $\widetilde{\mathcal{O}}$ (i.e., ignoring logarithmic terms, meaning that $\ln(T/\delta)$ is equivalent to a constant). For

$T \leq DS^2A$, a trivial bound on the regret is

$$\Delta(\text{UCRLB}, T) \leq r_{\max} T = r_{\max}\sqrt{T^2} \leq r_{\max} S\sqrt{DAT} \leq r_{\max} DS\sqrt{AT}$$

while for $T \geq DS^2A$ we have $r_{\max} DS\sqrt{AT} \geq r_{\max} S\sqrt{DAT} \geq r_{\max} DS^2A$. Since by definition $\Gamma \leq S$, in either case the regret of UCRLB can be bounded by $r_{\max} DS\sqrt{AT}$ just like the regret of UCRL2. But in general, UCRLB clearly enjoys *better* regret guarantees than UCRL2 as for the dependency in S. This is a consequence of the use of Bernstein bounds for the transition probabilities (Eq. 3.1) instead of Hoeffding/Weissman bounds in UCRL2. This improvement can be quite *significant* in practice since in most MDPs, $\Gamma SA = \Theta(SA)$ or at least $\sum_{s,a} \Gamma(s,a) = \Theta(SA)$ i.e., $\Gamma(s,a) = \mathcal{O}(1)$ for all but only $\mathcal{O}(1)$ state-action pairs. An environment that would satisfy $\Gamma(s,a) = \Omega(S)$ for $\Omega(S)$ state-action pairs would have a very *chaotic dynamics* which is not what we usually observe in *"real-world" environments*. The other improvement brought by Thm. 3.4 compared to the existing literature is the substitution of $r_{\max} D$ by $\max\{r_{\max}, \Lambda\} \leq r_{\max} D$ in the regret bound. Notice however that unlike the improvement in S, the improvement in D is only due to the *analysis* and not to the *algorithm* (improved bound on $sp\,(h_k)$ shown in Sec. 3.3). The same improvement can be shown for UCRL2.

Our second regret bound is reported in Thm. 3.5. This regret bound holds for UCRLB *without any modification of the algorithm*. The difference with Thm. 3.4 is due to a more careful analysis.

Theorem 3.5

There exists a numerical constant $\beta > 0$ such that for any *communicating MDP, with probability at least $1-\delta$, it holds that for all initial state distributions $\mu_1 \in \Delta_S$ and for all time horizons $T > 1$*

$$\begin{aligned}\Delta(\text{UCRLB}, T) \leq\ & \beta \cdot \max\left\{r_{\max}, \sqrt{r_{\max}\Lambda}\right\} \sqrt{\left(\sum_{s,a} \Gamma(s,a)\right) T \ln\left(\frac{T}{\delta}\right) \ln(T)} \\ & + \beta \cdot \max\left\{r_{\max}, \frac{\Lambda^2}{r_{\max}}\right\} S^2 A \ln\left(\frac{T}{\delta}\right) \ln(T)\end{aligned} \tag{3.37}$$

Since the dependency in $r_{\max}$ and Λ of (3.37) may appear non-trivial, we start with a simple *dimensional analysis* to check the consitency of the bound. The regret is always *homogeneous* to a reward and as a consequence, so should be the regret bound. Since both Λ and $r_{\max}$ are *homogeneous* to a reward, it is immediate to see that the bounds of Thm. 3.5 and Thm. 3.4 have the correct dimension. Compared to the bound of Thm. 3.4, the dominant term of Thm. 3.5 has a better dependency in Λ. Indeed, when $r_{\max} \geq \Lambda$, then

$$\max\left\{r_{\max}, \sqrt{r_{\max}\Lambda}\right\} = r_{\max} = \max\{r_{\max}, \Lambda\}$$

and so the dominant terms of (3.36) and (3.37) are the same. However, when $r_{\max} < \Lambda$ then

$$\max\left\{r_{\max}, \sqrt{r_{\max}\Lambda}\right\} = \sqrt{r_{\max}\Lambda} < \Lambda = \max\{r_{\max}, \Lambda\}$$

and so (3.37) is tighter. In conclusion: $\max\left\{r_{\max}, \sqrt{r_{\max}\Lambda}\right\} \leq \max\{r_{\max}, \Lambda\}$. As the improvement in the S-dependency, the improvement in the Λ-dependency of the regret bound is a consequence of the use of Bernstein bounds instead of Hoeffding/Weissman bounds for the transition probabilities. Notice also that in the case where $\Lambda = r_{\max} D$ (worst-case), then

$$\max\left\{r_{\max}, \sqrt{r_{\max}\Lambda}\right\} = r_{\max}\sqrt{D} \leq r_{\max} D = \max\{r_{\max}, \Lambda\}.$$

Symmetrically, $\max\{r_{\max}, \Lambda^2/r_{\max}\} \geq \max\{r_{\max}, \Lambda\}$, meaning that the improvement in the dominant term comes at the expense of an increase in the lower order (logarithmic) term. This term becomes negligible after only $D^2 S^2 A$ steps (ignoring the multiplicative log term). Using the same argument as in the discussion of Thm. 3.4 (see above), we can show that the bound (3.37) can be upper-bounded by $r_{\max} DS\sqrt{AT}$ for all T. In conclusion, the regret of UCRLB grows at most as $r_{\max}\sqrt{D\Gamma SAT}$ for T big enough which is clearly better than the regret of UCRL2 i.e., $r_{\max} DS\sqrt{AT}$. The additional *"burn-in"* of order $(\Lambda^2/r_{\max})S^2 A$ which dominates when T is small is not bigger than the burn-in of UCRL2, but in UCRL2, it is "hidden" by the dominant term $r_{\max} DS\sqrt{AT}$. An additional $\sqrt{\ln(T)}$ multiplicative factor also appears in the dominant term of (3.37) that was not present in (3.36). Whether this extra cost is an artefact of the analysis or cannot be removed is left as an open question.

Impact of the aperiodicity transformation. Neither of the regret bounds (3.36) and (3.37) depend on the aperiodicity parameter α. The $1/\alpha$ factor that appears in the bound of $sp(h_k)$ disappears in the regret proof when introducing the optimality equation (see Sec. 3.5.2). As expected, the aperiodicity transformation has absolutely no impact on the regret, its only impact is on the convergence (and speed of convergence) of EVI as already argued in Sec. 3.1.2.

Comparison with other settings. In the *finite horizon setting*, Azar et al. (2017) derived an algorithm –UCBVI_2– for which they proved a high-probability regret bound scaling as (up to multiplicative numerical constants):

$$r_{\max}\sqrt{HSAT}\ln\left(\frac{T}{\delta}\right) + r_{\max} H^2 S^2 A \ln^2\left(\frac{T}{\delta}\right) + r_{\max} H\sqrt{T\ln\left(\frac{T}{\delta}\right)}$$

where H is the horizon (known to the algorithm). It is common to compare $r_{\max} H$ with $r_{\max} D$ as both terms respectively upper-bound the *range* of the optimal *"value function"* (the bias in the infinite horizon undiscounted case). It is thus natural to compare $r_{\max} H$ with Λ in our case. After substituting the former by the latter, the bound they derived looks very similar to the bound of Thm. 3.5. The first difference is the absence of the support Γ in the dominant term of the regret. The second difference is the presence of an additional $\widetilde{\mathcal{O}}(r_{\max} H\sqrt{T})$ term. When T is big enough, their bound saves a $\sqrt{\Gamma}$ factor compared to ours when $H \leq SA$. It is not clear whether this improvement is *specific* to the finite horizon

setting or not. In particular, extending their proof to the infinite horizon setting does not seem straightforward as the definition of regret differ and several parts of the proof heavily rely on the existence of a known time horizon H. In the same setting, Kakade et al. (2018) introduced vUCQ that achieves a regret of the form $r_{\max}\sqrt{HSAT} + r_{\max}H^5SA$ (ignoring multiplicative logarithmic terms). This bound is similar to the one of Azar et al. (2017) but the time needed to reach the $\sqrt{T}$-regime (burn-in) is of order $\widetilde{\mathcal{O}}(H^5SA)$.

In the *discounted* infinite horizon setting, the common measure of performance of on-line learnig algorithms is the *sample complexity*. A regret bound of the form $C\sqrt{T}$ is usually interpreted as comparable to a sample complexity of order $\frac{C^2}{\varepsilon^2(1-\gamma)^2}$. For the same reason as in the finite horizon setting, it is natural to compare $r_{\max}\Lambda$ with $r_{\max}/(1-\gamma)$ (bound on the discounted value function). Using a UCRL2-like algorithm, Lattimore and Hutter (2012) achieved a sample complexity bound $\frac{r_{\max}SA}{\varepsilon^2(1-\gamma)^3}\ln(1/\delta)$ assuming that $\Gamma \leq 2$ (Lattimore and Hutter, 2012, Assumption 1) and later generalized their result to $\frac{r_{\max}\sum_{s,a}\Gamma(s,a)}{\varepsilon^2(1-\gamma)^3}\ln(1/\delta)$ (Lattimore and Hutter, 2014). This is comparable to the bound of Thm. 3.5.

Finally, Dann and Brunskill (2015) showed that their algorithm UCFH –similar to UCRLB– suffer a sample complexity of order at most $r_{\max}\frac{H^2\Gamma SA}{\varepsilon^2}\ln\left(\frac{1}{\delta}\right)$ where H is the (known) finite horizon. Unlike in the discounted setting, in the finite horizon case a regret bound of the form $C\sqrt{T}$ is usually interpreted as comparable to a sample complexity of order $\frac{C^2}{\varepsilon^2}$. Therefore, the bound of Thm. 3.5 saves a factor H compared to their bound. However, given the similarities between UCFH and UCRLB –both algorithms use Bernstein bounds for the transition probabilities– it is possible this additional H-factor could be removed by a *better analysis* i.e., without requiring any *change in the algorithm*.

In conclusion, the regret bound of Thm. 3.5 is *consistent* with state-of-the-art results in the *discounted setting*, but is *worse* than the state-of-the-art in the *finite horizon setting* by a factor $\sqrt{\Gamma}$.

3.5 First regret proof of UCRLB

We start with the proof of Thm. 3.4 which is both simpler and closer to the proof of Theorem 2 of Jaksch et al. (2010). We follow their proof structure, use similar notations and highlight the main differences. Many arguments will be reused for the proof of Thm. 3.5. In order to increase readibility, we postpone the detailed proof of some intermediate results to the appendix (see App. A). To be able to reuse this material in Chap. 5, we assume π_k my not always be deterministic (i.e., we assume π_k may be stochastic) although this is not strictly needed so far.

We recall two well-known results useful for the proof. We will extensively use *Azuma's inequality* (see for example Jaksch et al. (2010, Lemma 10)) which we recall below (see

Prop. 3.7).

Proposition 3.7 (Azuma's inequality)

Let $(X_n, \mathcal{F}_n)_{n\in\mathbb{N}}$ be an Martingale Difference Sequence (MDS) such that $|X_n| \leq a$ a.s. for all $n \in \mathbb{N}$. Then for all $\delta \in]0,1[$,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq a\sqrt{2n\ln\left(\frac{1}{\delta}\right)}\right) \leq \delta$$

An MDS is a sequence of r.v. X_n that are $\mathcal{F}_n$-integrable for every $n \in \mathbb{N}$, and such that $\mathbb{E}[X_{n+1}|\mathcal{F}_n] = X_n$.

We will also use *Cauchy-Schwarz inequality* several times i.e., $\sum_i |a_i b_i| \leq \sqrt{\sum_i a_i^2 \sum_i b_i^2}$ or equivalently $\sum_i \sqrt{a_i}\sqrt{b_i} \leq \sqrt{\sum_i a_i \sum_i b_i}$ if $a_i, b_i \geq 0$ for all i.

3.5.1 Splitting into episodes

The regret after T time steps is defined as $\Delta(\text{UCRLB}, T) = \sum_{t=1}^{T}\left(g^* - r_t\right)$. To begin with, we replace r_t by its expected value *conditioned* on the current state s_t using the following lemma:

Lemma 3.1

With probability at least $1 - \frac{\delta}{6}$:

$$\forall T \geq 1, \quad -\sum_{t=1}^{T} r_t \leq -\sum_{t=1}^{T}\sum_{a\in\mathcal{A}_{s_t}} \pi_{k_t}(s_t, a) r(s_t, a) + 2r_{\max}\sqrt{T\ln\left(\frac{4T}{\delta}\right)} \tag{3.38}$$

Proof. We use a martingale argument and Prop. 3.7 (see App. A.2 for further details). ■

Lem. 3.1 enables to "remove" from the analysis all the randomness due to the stochasticity of the observed rewards and the executed policy, at the expense of a small $\widetilde{\mathcal{O}}(\sqrt{T})$ term. Jaksch et al. (2010, Section 4.1) use a different argument to obtain a similar bound. They claim that once conditioned on the r.v. $(N_{k_T+1}(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ corresponding to the visit counts in all state-action pairs after T time steps, the r.v. $(r_t(s_t, a_t))_{T\geq t\geq 1}$ are *independent*. Although we *do not claim* that the sampled rewards are not independent conditioned on the visit counts as argued by the atuhors, they never *formally prove* this result and it is *not fully clear* why this property holds. For this reason, we prefer to use a *martingale argument* which is both simple and rigorous.

Let's denote by $\nu_k(s) := \sum_{a\in\mathcal{A}_s} \nu_k(s,a)$ the total number of visits in state s during episode k. Defining $\Delta_k := \sum_{s\in\mathcal{S}} \nu_k(s)\left(g^* - \sum_{a\in\mathcal{A}_{s_t}} \pi_k(a|s) r(s,a)\right)$ the pseudo-regret of episode k, it

holds with probability at least $1-\frac{\delta}{6}$ that for all $T \geq 1$ (using eq. 3.38):

$$\begin{aligned}\Delta(\text{UCRLB}, T) &\leq \sum_{t=1}^{T}\left(g^* - \sum_{a\in\mathcal{A}_{s_t}} \pi_{k_t}(s_t,a) r(s_t,a)\right) + 2r_{\max}\sqrt{T\ln\left(\frac{4T}{\delta}\right)} \\ &= \sum_{k=1}^{k_T}\sum_{s\in\mathcal{S}} \nu_k(s)\left(g^* - \sum_{a\in\mathcal{A}_s} \pi_k(a|s) r(s,a)\right) + 2r_{\max}\sqrt{T\ln\left(\frac{4T}{\delta}\right)} \\ &= \sum_{k=1}^{k_T} \Delta_k + 2r_{\max}\sqrt{T\ln\left(\frac{4T}{\delta}\right)} \end{aligned} \tag{3.39}$$

3.5.2 Plugging the optimistic Bellman optimality equation

In this section we derive a high probability bound for $\sum_{k=1}^{k_T}\Delta_k$. The first step consists in replacing the *true* optimal gain g^* by the optimistic gain g_k. To do this, we rely on the *optimism* property proved in Sec. 3.2. We assume that the complementary event of Thm. 3.1 holds i.e., $M \in \mathcal{M}_k$ for all $T \geq 1$ and for all $k \geq 1$. We will denote this event E in the rest of the regret proof. As shown in Sec. 3.2, under event E, we have that $g_k^* \geq g^*$. Moreover, as shown in Eq. 3.21, $|g_k - g_k^*| \leq \varepsilon_k/2$ implying that $g_k \geq g^* - \varepsilon_k/2$. As a result we can write:

$$\Delta_k \leq \sum_{s\in\mathcal{S}} \nu_k(s)\left(g_k - \sum_{a\in\mathcal{A}_s}\pi_k(a|s) r(s,a) + \frac{\varepsilon_k}{2}\right) \tag{3.40}$$

We will now replace g_k (optimistic gain) by h_k (optimistic bias) using the *optimistic optimality equation*.

We denote by p_k and r_k the transition probabilities and rewards satisfying

$$\forall s\in\mathcal{S},\quad \mathcal{L}_\alpha^k h_k(s) = \sum_{a\in\mathcal{A}_s}\pi_k(a|s) r_k(s,a) + \alpha\sum_{a\in\mathcal{A}_s}\sum_{s'\in\mathcal{S}}\pi_k(a|s)p_k(s'|s,a)h_k(s') + (1-\alpha)h_k(s)$$

As shown in Eq. 3.22, the pair $(g_k, h_k) \in \mathbb{R}\times\mathbb{R}^S$ returned by EVI is an approximate solution to the Bellman optimality equation of $\mathcal{L}_\alpha^k$ i.e., $\|\mathcal{L}_\alpha^k h_k - h_k - g_k e\|_\infty \leq \varepsilon_k$ implying that for all $s \in \mathcal{S}$:

$$\begin{aligned} &-\sum_{a\in\mathcal{A}_s}\pi_k(a|s) r_k(s,a) - \alpha\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}_s}\pi_k(a|s)p_k(s'|s,a)h_k(s') - (1-\alpha)h_k(s) + \cancel{h_k(s)} + g_k \leq \varepsilon_k \\ &\Rightarrow \left(g_k - \sum_{a\in\mathcal{A}_s}\pi_k(a|s) r_k(s,a)\right) + \alpha\left(h_k(s) - \sum_{a\in\mathcal{A}_s}\sum_{s'\in\mathcal{S}}\pi_k(a|s)p_k(s'|s,a)h_k(s')\right) \leq \varepsilon_k \end{aligned} \tag{3.41}$$

Plugging Eq. 3.41 into Eq. 3.40 yields:

$$\begin{aligned}\Delta_k &\leq \sum_{s\in\mathcal{S}}\nu_k(s)\left(g_k - \sum_{a\in\mathcal{A}_s}\pi_k(a|s)r_k(s,a) + \frac{\varepsilon_k}{2}\right) + \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_s}\nu_k(s)\pi_k(a|s)\Big(r_k(s,a) - r(s,a)\Big)\\ &\leq \underbrace{\alpha\sum_{s\in\mathcal{S}}\nu_k(s)\left(\sum_{a\in\mathcal{A}_s}\sum_{s'\in\mathcal{S}}\pi_k(a|s)p_k(s'|s,a)h_k(s') - h_k(s)\right)}_{:=\Delta_k^p}\\ &\quad + \underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_s}\nu_k(s)\pi_k(a|s)\Big(r_k(s,a) - r(s,a)\Big)}_{:=\Delta_k^r} + \frac{3\varepsilon_k}{2}\sum_{s\in\mathcal{S}}\nu_k(s)\end{aligned} \tag{3.42}$$

In the next two sections (Sec. 3.5.3 and 3.5.4), we will bound the sums $\sum_{k=1}^{k_T}\Delta_k^p$ and $\sum_{k=1}^{k_T}\Delta_k^r$.

3.5.3 Bounding the transition probabilities

We start by further decomposing Δ_k^p into two different terms:

$$\begin{aligned}\Delta_k^p &= \underbrace{\alpha\sum_{s,a,s'}\nu_k(s)\pi_k(a|s)\Big(p_k(s'|s,a) - p(s'|s,a)\Big)h_k(s')}_{:=\Delta_k^{p1}}\\ &\quad + \underbrace{\alpha\sum_{s}\nu_k(s)\left(\sum_{a,s'}\pi_k(a|s)p(s'|s,a)h_k(s') - h_k(s)\right)}_{:=\Delta_k^{p2}}\end{aligned} \tag{3.43}$$

Since by construction $\sum_{a,s'\in\mathcal{S}}p_k(s'|s,a)\pi_k(a|s) = \sum_{a,s'\in\mathcal{S}}p(s'|s,a)\pi_k(a|s) = 1$, the terms Δ_k^{p1} and Δ_k^{p2} remain unchanged if h_k is *arbitrarily shifted* by a constant vector, respectively $\lambda_k^1 e$ and $\lambda_k^2 e$ (λ_k^1, $\lambda_k^2 \in \mathbb{R}$ are arbitrary scalars and $e = (1,\dots,1)^\intercal$ is the vector of all ones). To obtain the tightest possible upper bounds, we choose

$$\lambda_k^1 = \lambda_k^2 = -\frac{1}{2}\left(\max_{s\in\mathcal{S}}h_k(s) + \min_{s\in\mathcal{S}}h_k(s)\right)$$

which minimizes the ℓ_∞-norm of $w_k := h_k + \lambda_k^1 e = h_k + \lambda_k^2 e$. Indeed, it is immediate to see that $sp\left(w_k\right) = sp\left(h_k\right)$ and $\|w_k\|_\infty = sp\left(h_k\right)/2$. Under event E, we showed in Sec. 3.3.2 that $sp\left(w_k\right) = sp\left(h_k\right) \leq \Lambda/\alpha$ and so $\|w_k\|_\infty \leq \Lambda/(2\alpha)$. To keep $\sum_{k=1}^{k_T}\Delta_k^{p1}$ under control, we need to replace $\nu_k(s)\pi_k(a|s)$ by $\nu_k(s,a)$ i.e., reintroduce the randomness of the executed policy. To that end, we define $\Delta_k^{p3} := \alpha\sum_{s,a,s'}\nu_k(s,a)\left(p_k(s'|s,a) - p(s'|s,a)\right)h_k(s')$, analogue of Δ_k^{p1} with $\nu_k(s)\pi_k(a|s)$ replaced by $\nu_k(s,a)$, and we use the following lemma:

Lemma 3.2

Under event E, with probability at least $1 - \frac{\delta}{6}$:

$$\forall T \geq 1,\quad \sum_{k=1}^{k_T}\Delta_k^{p1} \leq \sum_{k=1}^{k_T}\Delta_k^{p3} + 4\Lambda\sqrt{T\ln\left(\frac{6T}{\delta}\right)} \tag{3.44}$$

Proof. We use a martingale argument and Prop. 3.7 (see App. A.2 for further details). ■

Using *Hölder's inequality*, the term Δ_k^{p3} can be bounded as follows:

$$\begin{aligned}\Delta_k^{p3} &\leq \cancel{\alpha} \sum_{s,a} \nu_k(s,a) \cdot \|p_k(\cdot|s,a) - p(\cdot|s,a)\|_1 \cdot \underbrace{\|w_k\|_\infty}_{\leq \Lambda/(2\cancel{\alpha})} \\ &\leq \frac{\Lambda}{2} \sum_{s,a} \nu_k(s,a) \cdot \|p_k(\cdot|s,a) - p(\cdot|s,a)\|_1\end{aligned}$$

Using the *triangle inequality*, we can decompose the ℓ_1-norm into two terms:

$$\|p_k(\cdot|s,a) - p(\cdot|s,a)\|_1 \leq \|p_k(\cdot|s,a) - \widehat{p}_k(\cdot|s,a)\|_1 + \|\widehat{p}_k(\cdot|s,a) - p(\cdot|s,a)\|_1 \tag{3.45}$$

By construction, $p_k(\cdot|s,a) \in B_p^k(s,a)$ (see Eq. 3.17) implying that for all states $s' \in \mathcal{S}$, $|p_k(s'|s,a) - \widehat{p}_k(s'|s,a)| \leq \beta_{p,k}^{sas'}$ and so $\|p_k(\cdot|s,a) - \widehat{p}_k(\cdot|s,a)\|_1 \leq \beta_{p,k}^{sa} := \sum_{s' \in \mathcal{S}} \beta_{p,k}^{sas'}$. Similarly, under event E, $p(\cdot|s,a) \in B_p^k(s,a)$ (by definition) and so $|\widehat{p}_k(s'|s,a) - p(s'|s,a)| \leq \beta_{p,k}^{sas'}$ for all $s' \in \mathcal{S}$ implying that $\|p_k(\cdot|s,a) - \widehat{p}_k(\cdot|s,a)\|_1 \leq \beta_{p,k}^{sa}$. In conclusion,

$$\Delta_k^{p3} \leq \Lambda \sum_{s,a} \nu_k(s,a) \cdot \beta_{p,k}^{sa} \tag{3.46}$$

We now focus on the last term Δ_k^{p2} and do the following decomposition:

$$\begin{aligned}\Delta_k^{p2} &= \alpha \sum_{t=t_k}^{t_{k+1}-1} \left(\sum_{a,s'} \pi_k(a|s_t) p(s'|s_t,a) w_k(s') - w_k(s_t) \right) \\ &= \underbrace{\alpha \sum_{t=t_k}^{t_{k+1}-1} \left(\sum_{a,s'} \pi_k(a|s_t) p(s'|s_t,a) w_k(s') - w_k(s_{t+1}) \right)}_{:=\Delta_k^{p4}} + \alpha \underbrace{\sum_{t=t_k}^{t_{k+1}-1} w_k(s_{t+1}) - w_k(s_t)}_{\textit{telescopic sum}} \\ &= \Delta_k^{p4} + \alpha \underbrace{(w_k(s_{t_{k+1}}) - w_k(s_{t_k}))}_{\leq sp(w_k) \leq \Lambda/\alpha} \leq \Delta_k^{p4} + \Lambda\end{aligned}$$

We then notice that $\sum_{k=1}^{k_T} \Delta_k^{p4}$ is an MDS and so

Lemma 3.3

Under event E, with probability at least $1 - \frac{\delta}{6}$:

$$\forall T \geq 1, \quad \sum_{k=1}^{k_T} \Delta_k^{p4} \leq 2\Lambda \sqrt{T \ln\left(\frac{4T}{\delta}\right)} \tag{3.47}$$

Proof. We use a martingale argument and Prop. 3.7 (see App. A.2 for further details). ■

After combining Lem. 3.2, Eq. 3.46 and Lem. 3.3 and taking a union bound, we conclude that with probability at least $1-\frac{\delta}{3}$ (and assuming event E holds):

$$\forall T \geq 1, \quad \sum_{k=1}^{k_T} \Delta_k^p \leq \Lambda \sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a) \beta_{p,k}^{sa} + 6\Lambda \sqrt{T \ln\left(\frac{5T}{\delta}\right)} + \Lambda k_T. \tag{3.48}$$

3.5.4 Bounding the rewards

Similarly to what we did to bound Δ_k^{p1}, we define an analogue of Δ_k^{p3} for the rewards i.e., $\Delta_k^{r1} := \sum_{s,a} \nu_k(s,a) \left(r_k(s,a) - r(s,a)\right)$ (similar to Δ_k^r with $\nu_k(s)\pi_k(a|s)$ replaced by $\nu_k(s,a)$) and we show the following lemma:

Lemma 3.4

With probability at least $1-\frac{\delta}{6}$*:*

$$\forall T \geq 1, \quad \sum_{k=1}^{k_T} \Delta_k^r \leq \sum_{k=1}^{k_T} \Delta_k^{r1} + 4 r_{\max} \sqrt{T \ln\left(\frac{8T}{\delta}\right)}$$

Proof. We use a martingale argument and Prop. 3.7 (see App. A.2 for further details). ■

Similarly to the bound in (3.45), we notice that $r_k(s,a) - r(s,a)$ can be expressed as the sum of $r_k(s,a) - \widehat{r}_k(s,a)$ with $\widehat{r}_k(s,a) - r(s,a)$. Since $r_k(s,a) \in B_r^k(s,a)$ (3.4) by construction, $r_k(s,a) - \widehat{r}_k(s,a) \leq \beta_r^k(s,a)$. Moreover, under event E we have $\widehat{r}_k(s,a) - r(s,a) \leq \beta_r^k(s,a)$ by definition. After summing up the two inequalities we obtain:

$$\Delta_k^{r1} = \sum_{s,a} \nu_k(s,a) \Big(r_k(s,a) - r(s,a)\Big) \leq 2 \sum_{s,a} \nu_k(s,a) \beta_{r,k}^{sa}$$

In conclusion, with probability at least $1-\frac{\delta}{6}$ (and assuming event E holds):

$$\forall T \geq 1, \quad \sum_{k=1}^{k_T} \Delta_k^r \leq 2 \sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a) \beta_{r,k}^{sa} + 4 r_{\max} \sqrt{T \ln\left(\frac{4T}{\delta}\right)} \tag{3.49}$$

3.5.5 Bounding the number of episodes

As in UCRL2, in UCRLB the inequality $\nu_k(s,a) \leq N_k^+(s,a)$ holds for all state-action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$. However, the equality $\nu_k(s,a) = N_k^+(s,a)$ holds true for *exactly* one state-action pair (never more).

In Appendix C.2, Jaksch et al. (2010, Proposition 18) proved that the stopping condition of UCRL2 ensures that when $T \geq SA$, $k_T \leq SA \log_2\left(\frac{8T}{SA}\right)$ a.s. The proof of this result only relies on the fact that there exists at least one (s,a) satisfying $\nu_k(s,a) \geq N_k^+(s,a)$. Since

UCRLB also enjoys this property, the same proof applies and the bound still holds:

Proposition 3.8

For all $T \geq SA$, $k_T \leq SA \log_2\left(\frac{8T}{SA}\right)$.

3.5.6 Summing over episodes

As proved in Thm. 3.1, event E occurs with probability at least $1 - \frac{\delta}{3}$. After taking a union bound and gathering inequalities (3.48) and (3.49) into inequality (3.42) we conclude that with probability at least $1 - \frac{5\delta}{6}$, for all $T \geq SA$:

$$\begin{aligned}\sum_{k=1}^{k_T} \Delta_k \leq\ & \underbrace{\Lambda \sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a) \beta_{p,k}^{sa}}_{(3)} + \underbrace{2 \sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a) \beta_{r,k}^{sa}}_{(2)} + \underbrace{\frac{3}{2} r_{\max} \sum_{k=1}^{k_T} \sum_{s \in \mathcal{S}} \frac{\nu_k(s)}{t_k}}_{(1)} \\ & + 6\Lambda \sqrt{T \ln\left(\frac{5T}{\delta}\right)} + 4 r_{\max} \sqrt{T \ln\left(\frac{5T}{\delta}\right)} + \Lambda SA \log_2\left(\frac{8T}{SA}\right)\end{aligned} \tag{3.50}$$

We will now expand the first three terms appearing in the bound of Eq. 3.50.

(1) Since $t_k \geq N_k^+(s,a)$ for all (s,a), we deduce that:

$$\sum_{k=1}^{k_T} \sum_{s \in \mathcal{S}} \frac{\nu_k(s)}{t_k} = \sum_{s,a} \sum_{k=1}^{k_T} \frac{\nu_k(s,a)}{t_k} \leq \sum_{s,a} \sum_{k=1}^{k_T} \frac{\nu_k(s,a)}{N_k^+(s,a)}$$

(2) Using the definition of $\beta_{r,k}^{sa}$:

$$\begin{aligned}\sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a) \beta_{r,k}^{sa} = \sum_{k=1}^{k_T} \sum_{s,a} \Bigg[& \underbrace{2 \sqrt{\widehat{\sigma}_{r,k}^2(s,a) \ln\left(\frac{6SAN_k^+(s,a)}{\delta}\right)}}_{\leq 2 r_{\max} \sqrt{\ln\left(\frac{6SAT}{\delta}\right)}} \frac{\nu_k(s,a)}{\sqrt{N_k^+(s,a)}} \\ & + 6 r_{\max} \underbrace{\ln\left(\frac{6SAN_k^+(s,a)}{\delta}\right)}_{\leq \ln\left(\frac{6SAT}{\delta}\right)} \frac{\nu_k(s,a)}{N_k^+(s,a)} \Bigg] \\ \leq\ & 2 r_{\max} \left(\sqrt{\ln\left(\frac{6SAT}{\delta}\right)} \sum_{s,a} \sum_{k=1}^{k_T} \frac{\nu_k(s,a)}{\sqrt{N_k^+(s,a)}} + 3 \ln\left(\frac{6SAT}{\delta}\right) \sum_{s,a} \sum_{k=1}^{k_T} \frac{\nu_k(s,a)}{N_k^+(s,a)} \right)\end{aligned}$$

(3) Similarly using the fact that $\beta_{p,k}^{sa} = \sum_{s' \in \mathcal{S}} \beta_{p,k}^{sas'}$:

$$\sum_{k=1}^{k_T}\sum_{s,a}\nu_k(s,a)\beta_{p,k}^{sa} \leq 2\sqrt{\ln\left(\frac{6SAT}{\delta}\right)}\sum_{s,a}\sum_{k=1}^{k_T}\frac{\nu_k(s,a)}{\sqrt{N_k^+(s,a)}}\sum_{s'\in\mathcal{S}}\sqrt{\widehat{p}_k(s'|s,a)(1-\widehat{p}_k(s'|s,a))}$$
$$+6S\ln\left(\frac{6SAT}{\delta}\right)\sum_{s,a}\sum_{k=1}^{k_T}\frac{\nu_k(s,a)}{N_k^+(s,a)}$$

Lemma 3.5

It holds almost surely that for all $k \geq 1$ and for all $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$:

$$\sum_{s'\in\mathcal{S}}\sqrt{\widehat{p}_k(s'|s,a)(1-\widehat{p}_k(s'|s,a))} \leq \sqrt{\Gamma(s,a)-1} \tag{3.51}$$

Proof. The result is a direct consequence of Cauchy-Schwarz inequality (for further details, see App. A.3). ■

As a consequence of Lem. 3.5,

$$\sum_{k=1}^{k_T}\sum_{s,a}\nu_k(s,a)\beta_{p,k}^{sa} \leq 2\sqrt{\ln\left(\frac{6SAT}{\delta}\right)}\sum_{s,a}\sqrt{\Gamma(s,a)}\sum_{k=1}^{k_T}\frac{\nu_k(s,a)}{\sqrt{N_k^+(s,a)}}$$
$$+6S\ln\left(\frac{6SAT}{\delta}\right)\sum_{s,a}\sum_{k=1}^{k_T}\frac{\nu_k(s,a)}{N_k^+(s,a)}$$

Two sums appear in the bounds of the terms (1), (2) and (3):

$$\sum_{k=1}^{k_T}\frac{\nu_k(s,a)}{\sqrt{N_k^+(s,a)}} \quad\text{and}\quad \sum_{k=1}^{k_T}\frac{\nu_k(s,a)}{N_k^+(s,a)}.$$

Lem. 3.6 provides upper-bounds for those sums.

Lemma 3.6

It holds almost surely that for all $k \geq 1$ and for all $(s,a) \in \mathcal{S}\times\mathcal{A}\times$:

$$\sum_{k=1}^{k_T}\frac{\nu_k(s,a)}{\sqrt{N_k^+(s,a)}} \leq 3\sqrt{N_{k_T+1}(s,a)} \quad \textit{and} \quad \sum_{k=1}^{k_T}\frac{\nu_k(s,a)}{N_k^+(s,a)} \leq 2+2\ln\left(N_{k_T+1}^+(s,a)\right) \tag{3.52}$$

Proof. The proof follows from the rate of divergence of the series $\sum_{i=1}^{n}\frac{1}{\sqrt{i}} \sim \sqrt{n}$ and $\sum_{i=1}^{n}\frac{1}{i} \sim \ln(n)$ respectively when $n \to +\infty$. ■

Using Lem. 3.6 together with Cauchy-Schwartz inequality we have:

$$\sum_{s,a}\sqrt{\Gamma(s,a)}\sqrt{N_{k_T+1}(s,a)} \leq \sqrt{\left(\sum_{s,a}\Gamma(s,a)\right)\cdot\sum_{s,a}N_{k_T+1}(s,a)} = \sqrt{\left(\sum_{s,a}\Gamma(s,a)\right)T}.$$

Using Lem. 3.6 together with Jensen's inequality on the concave function $\ln(\cdot)$ (with a normalization factor SA) and the fact that $N^+_{k_T+1}(s,a) \leq T$, we have(for $T, SA \geq 2$):

$$\sum_{s,a} \ln\left(N^+_{k_T+1}(s,a)\right) \leq SA \ln\left(\frac{\sum_{s,a} N^+_{k_T+1}(s,a)}{SA}\right) \leq SA \ln(T)\,.$$

In conclusion, with probability at least $1 - \frac{5\delta}{6}$, for all $T \geq SA$:

$$\begin{aligned}\sum_{k=1}^{k_T} \Delta_k \leq\ & 6\Lambda\sqrt{\left(\sum_{s,a}\Gamma(s,a)\right) T \ln\left(\frac{6SAT}{\delta}\right)} + 12\Lambda S^2 A \ln\left(\frac{6SAT}{\delta}\right)(1+\ln(T)) \\ & + 6r_{\max}\sqrt{SAT\ln\left(\frac{6SAT}{\delta}\right)} + 12 r_{\max} SA \ln\left(\frac{6SAT}{\delta}\right)(1+\ln(T)) \\ & + 4(\Lambda + r_{\max})\sqrt{T\ln\left(\frac{5T}{\delta}\right)} + \Lambda SA \log_2\left(\frac{T}{SA}\right) + 3 r_{\max} SA \cdot (1+\ln(T)).\end{aligned} \tag{3.53}$$

3.5.7 Completing the regret bound of Thm. 3.4

For $T \geq 6SA$ we have $6SAT \leq T^2$ so that $\ln\left(\frac{6SAT}{\delta}\right) \leq \ln\left(\frac{T^2}{\delta}\right) \leq 2\ln\left(\frac{T}{\delta}\right)$ i.e., the logarithmic terms appearing in (3.53) no longer depend on SA but a factor 2 appears. For $T \leq 6SA$ we can use the trivial upper-bound $r_{\max}T$ on the regret (which holds with probability 1) and so

$$\Delta(\text{UCRLB}, T) \leq r_{\max} T = r_{\max}\sqrt{T}\cdot\sqrt{T} \leq r_{\max}\sqrt{6SAT} \leq \sqrt{6\left(\sum_{s,a}\Gamma(s,a)\right)T}.$$

After combining (3.39) with (3.53) and using a union bound, we obtain that there exists an *absolute* numerical constant $\beta > 0$ (i.e., independent of the MDP instance) such that for any MDP M, with probability at least $1-\delta$, for all $T > 1$ the regret of UCRLB after T steps is bounded as

$$\Delta(\text{UCRLB}, T) \leq \beta \cdot \max\{r_{\max}, \Lambda\} \cdot \left(\sqrt{\left(\sum_{s,a}\Gamma(s,a)\right) T \ln\left(\frac{T}{\delta}\right)} + S^2 A \ln\left(\frac{T}{\delta}\right)\ln(T)\right).$$

3.6 Improved regret analysis for UCRLB using variance reduction methods

We now prove Thm. 3.5. In order to improve the dependency of the regret bound in Λ (i.e., replace Λ by $\sqrt{\Lambda}$), we refine our analysis with three key improvements:

1. We leverage on *Freedman's inequality* (Freedman, 1975) instead of Azuma's inequality to bound all MDS. We recall this inequality in Prop. 3.9 below.
2. We use a *tighter bound* than Hölder's inequality to upper-bound the sum $\sum_{k=1}^{k_T} \Delta_k^{p3}$ (see Sec. 3.5.3).

3. We shift the optimistic bias h_{k_t} by a different constant *at every time step* $t \geq 1$ rather than only at every episode $k \geq 1$. More precisely, the optimistic bias is shifted by a different constant for every episode $k \geq 1$ and for every visited state $s \in \mathcal{S}$.

To the best of our knowledge, Thm. 3.5 and its proof are new although it is largely inspired by what is often referred to as *"variance reduction methods"* in the literature (Munos and Moore, 1999; Lattimore and Hutter, 2012, 2014; Azar et al., 2017; Kakade et al., 2018). Similar techniques are used by (Azar et al., 2017) to achieve a similar bound but in the *finite horizon setting*. Our approach also borrows intuitions from the work of Talebi and Maillard (2018a) and Maillard et al. (2014).

Proposition 3.9 (Freedman's inequality)

Let $(X_n, \mathcal{F}_n)_{n\in\mathbb{N}}$ be an MDS such that $|X_n| \leq a$ a.s. for all $n \in \mathbb{N}$. Then for all $\delta \in]0,1[$,

$$\mathbb{P}\left(\forall n \geq 1,\ \left|\sum_{i=1}^{n} X_i\right| \geq 2\sqrt{\left(\sum_{i=1}^{n} \mathbb{V}\left(X_i|\mathcal{F}_{i-1}\right)\right)\cdot \ln\left(\frac{4n}{\delta}\right)} + 4a\ln\left(\frac{4n}{\delta}\right)\right) \leq \delta$$

To bound the rewards $\sum_{k=1}^{k_T} \Delta_k^r$, we keep the same derivation as in Sec. 3.5.4 (see Eq. 3.49). On the other hand, we derive a completely differnt bound for the transition probabilities $\sum_{k=1}^{k_T} \Delta_k^p$. Our new derivation will make appear some sums of *variances*.

For any *vector* $u \in \mathbb{R}^S$, we slightly abuse notation and write $u^2 := u \circ u$ the *Hadamard product* of u with itself. For any probability distribution p over states $\mathcal{S}$ and any vector $u \in \mathbb{R}^S$ we define $\mathbb{V}_p(u) := p^\intercal u^2 - (p^\intercal u)^2 = \mathbb{E}_{X\sim p}[u(X)^2] - \left(\mathbb{E}_{X\sim p}[u(X)]\right)^2$ the *"variance"* of u with respect to p.[9] For the sake of clarity we introduce new notations for the transition probabilities: $p_k(s'|s) := \sum_{a\in\mathcal{A}_s} \pi_k(a|s)p_k(s'|s,a)$, $\overline{p}_k(s'|s) := \sum_{a\in\mathcal{A}_s} \pi_k(a|s)p(s'|s,a)$ and $\widehat{p}_k(s'|s) := \sum_{a\in\mathcal{A}_s} \pi_k(a|s)\widehat{p}_k(s'|s,a)$, for every $s, s' \in \mathcal{S}$ and every $k \geq 1$ (i.e., we drop the summation over a).

We start with a new bound relating Δ_k^{p1} and Δ_k^{p3} (as in Lem. 3.2):

Lemma 3.7 (Analogue of Lem. 3.2)

Under event E, with probability at least $1 - \frac{\delta}{6}$:

$$\forall T \geq 1,\ \sum_{k=1}^{k_T} \Delta_k^{p1} \leq \sum_{k=1}^{k_T} \Delta_k^{p3} + 4\Lambda\ln\left(\frac{24T}{\delta}\right) + 2\sqrt{S\ln\left(\frac{24T}{\delta}\right)}\left(\sqrt{\sum_{t=1}^{T} \mathbb{V}_{p_{k_t}(\cdot|s_t)}(\alpha h_{k_t})} + \sqrt{\sum_{t=1}^{T} \mathbb{V}_{\overline{p}_{k_t}(\cdot|s_t)}(\alpha h_{k_t})}\right) \tag{3.54}$$

Proof. We use a martingale argument and Prop. 3.9 (see App. A.2 for further details). ■

[9] In (Maillard et al., 2014), the authors define the "distribution-norm" of an MDP which is related to the variances $\mathbb{V}_{p(\cdot|s,a)}(h^*)$.

We also *refine* the upper-bound of Δ_k^{p3} derived in Eq. 3.46. Instead of bounding the scalar product $(p_k(\cdot|s,a)-p(\cdot|s,a))^\intercal w_k$ by $\|p_k(\cdot|s,a)-p(\cdot|s,a)\|_1^\intercal \|w_k\|_\infty$ using Hölder's inequality as in Sec. 3.5.3, we bound it by $\sum_{s'} |p_k(s'|s,a)-p(s'|s,a)|\cdot|w_k(s')|$ using the triangle inequality. Since $\sum_{a,s'} p_k(s'|s,a) = \sum_{a,s'} p(s'|s,a) = 1$ we can shift h_k by an arbitrary scalar $\lambda_k^s \in \mathbb{R}$ for all $k \geq 1$ and all $s \in \mathcal{S}$, i.e., $w_k^s := h_k + \lambda_k^s e$. Unlike in Sec. 3.5.3, we choose a *state-dependent* shift, namely $\lambda_k^s := -\sum_{a,s'} \widehat{p}_k(s'|s,a)\pi_k(a|s)h_k(s') = -\widehat{p}_k(\cdot|s)^\intercal h_k$. It is easy to see that $sp\left(w_k^s\right) = sp\left(h_k\right)$ and $\|w_k^s\|_\infty \leq sp\left(h_k\right)$ implying that under event E, $\|w_k^s\|_\infty \leq \Lambda/\alpha$.

Using the triangle inequality and the fact that $p_k(\cdot|s,a) \in B_p^k(s,a)$ by construction and $p(\cdot|s,a) \in B_p^k(s,a)$ under event E:

$$|p_k(s'|s,a) - p(s'|s,a)| \leq |p_k(s'|s,a) - \widehat{p}_k(s'|s,a)| + |\widehat{p}_k(s'|s,a) - p(s'|s,a)| \leq 2\beta_{p,k}^{sas'}.$$

As a result we can write:

$$\begin{aligned}
\Delta_k^{p3} &\leq \alpha \sum_{k=1}^{k_T} \sum_{s,a,s'} \nu_k(s,a)\Big|p_k(s'|s,a) - p(s'|s,a)\Big| \cdot |w_k^s(s')| \\
&\leq 2\alpha \sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a) \sum_{s'} \beta_{p,k}^{sas'} \cdot |w_k^s(s')| \\
&= 4\alpha \sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a) \Bigg[\sqrt{\frac{\ln\left(6SAT/\delta\right)}{N_k^+(s,a)}} \sum_{s'\in\mathcal{S}} \sqrt{\widehat{p}_k(s'|s,a)(1-\widehat{p}_k(s'|s,a))w_k^s(s')^2} \\
&\qquad + \frac{3\ln\left(6SAT/\delta\right)}{N_k^+(s,a)} \sum_{s'} \underbrace{|w_k^{sa}(s')|}_{\leq \Lambda/\alpha} \Bigg]
\end{aligned}$$

We denote by $V_k(s,a) := \alpha^2 \sum_{s'} \widehat{p}_k(s'|s,a) w_k^s(s')^2$. Similarly to Lem. 3.5, we can prove the following inequality:

Lemma 3.8 (Analogue of Lem. 3.5)

It holds almost surely that for all $k \geq 1$ and for all $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$:

$$\alpha \sum_{s'\in\mathcal{S}} \sqrt{\widehat{p}_k(s'|s,a)(1-\widehat{p}_k(s'|s,a))w_k^s(s')^2} \leq \sqrt{V_k(s,a)\cdot(\Gamma(s,a)-1)} \tag{3.55}$$

Proof. The result is a direct consequence of Cauchy-Schwarz inequality (for further details, see App. A.3). ■

As a consequence of Lem. 3.8,

$$\begin{aligned}
\sum_{k=1}^{k_T} \Delta_k^{p3} &\leq 4 \sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a) \left[\sqrt{V_k(s,a)\frac{\Gamma(s,a)}{N_k^+(s,a)}\ln\left(\frac{6SAT}{\delta}\right)} + \frac{3\Lambda S}{N_k^+(s,a)}\ln\left(\frac{6SAT}{\delta}\right)\right] \\
&= 4 \sum_{k=1}^{k_T} \sum_{t=t_k}^{t_{k+1}-1} \left[\sqrt{V_k(s_t,a_t)\frac{\Gamma(s_t,a_t)}{N_k^+(s_t,a_t)}\ln\left(\frac{6SAT}{\delta}\right)} + \frac{3\Lambda S}{N_k^+(s_t,a_t)}\ln\left(\frac{6SAT}{\delta}\right)\right].
\end{aligned}$$

Applying Cauchy-Schwartz gives

$$\sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-1}\sqrt{V_k(s_t,a_t)}\sqrt{\frac{\Gamma(s_t,a_t)}{N_k^+(s_t,a_t)}} \leq \sqrt{\sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-1}\frac{\Gamma(s_t,a_t)}{N_k^+(s_t,a_t)}\sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-1}V_k(s_t,a_t)}$$
$$= \sqrt{\sum_{k=1}^{k_T}\sum_{s,a}\frac{\Gamma(s,a)\nu_k(s,a)}{N_k^+(s,a)}\sum_{t=1}^{T}V_{k_t}(s_t,a_t)}.$$

Using Lem. 3.6, Jensen's inequality and the fact that $N_{k_T+1}^+(s,a) \leq T$ (as in Sec. 3.5.6), we can bound the first sum

$$\begin{aligned}\sum_{s,a}\sum_{k=1}^{k_T}\frac{\Gamma(s,a)\nu_k(s,a)}{N_k^+(s,a)} &\leq 2\sum_{s,a}\Gamma(s,a)\left(1+\ln\left(N_{k_T+1}^+(s,a)\right)\right)\\ &\leq 2\left(1+\ln\left(\frac{\sum_{s,a}\Gamma(s,a)N_{k_T+1}^+(s,a)}{\sum_{s,a}\Gamma(s,a)}\right)\right)\sum_{s,a}\Gamma(s,a)\\ &\leq 2(1+\ln(T))\sum_{s,a}\Gamma(s,a).\end{aligned}$$

To bound the second sum $\sum_{t=1}^{T}V_{k_t}(s_t,a_t)$, we rely on the following Lemma:

Lemma 3.9

Under event E, with probability at least $1-\frac{\delta}{6}$:

$$\forall T\geq 1,\quad \sum_{t=1}^{T}V_{k_t}(s_t,a_t) \leq \sum_{t=1}^{T}\mathbb{V}_{\widehat{p}_{k_t}(\cdot|s_t)}(\alpha h_{k_t}) + 2\Lambda^2\sqrt{T\ln\left(\frac{4T}{\delta}\right)} \tag{3.56}$$

Proof. We notice that for all $k\geq 1$ and $s\in\mathcal{S}$, $\sum_a \pi_k(a|s)V_k(s,a) = \mathbb{V}_{\widehat{p}_k(\cdot|s)}(\alpha h_k)$. The concentration inequality then follows from a martingale argument and Prop. 3.7 (see App. A.2 for further details). ■

From Lem. 3.9 it follows that

$$\begin{aligned}\sum_{k=1}^{k_T}\Delta_k^{p3} \leq & 4\sqrt{2\big(1+\ln(T)\big)\ln\left(\frac{6SAT}{\delta}\right)\left(\sum_{s,a}\Gamma(s,a)\right)\left(\Lambda^2\sqrt{2T\ln\left(\frac{T}{\delta}\right)}+\sum_{t=1}^{T}\mathbb{V}_{\widehat{p}_{k_t}(\cdot|s_t)}(\alpha h_{k_t})\right)}\\ &+ 24\Lambda S^2 A\ln\left(\frac{6SAT}{\delta}\right)(1+\ln(T))\end{aligned} \tag{3.57}$$

It now remains to bound $\sum_{k=1}^{k_T}\Delta_k^{p2}$. As shown in Sec. 3.5.3: $\sum_{k=1}^{k_T}\Delta_k^{p2} \leq \sum_{k=1}^{k_T}\Delta_k^{p4} + \Lambda k_T$. We refine the bound on $\sum_{k=1}^{k_T}\Delta_k^{p4}$ derived in Eq. 3.48 using Freedman's inequality instead of Azuma's.

Lemma 3.10 (Analogue of Lem. 3.3)

Under event E, with probability at least $1-\frac{\delta}{6}$:

$$\forall T\geq 1,\quad \sum_{k=1}^{k_T}\Delta_k^{p4} \leq 2\sqrt{\left(\sum_{t=1}^{T}\mathbb{V}_{\bar{p}_{k_t}(\cdot|s_t)}(\alpha h_{k_t})\right)\cdot\ln\left(\frac{24T}{\delta}\right)} + 4\Lambda\ln\left(\frac{24T}{\delta}\right) \tag{3.58}$$

Proof. We use a martingale argument and Prop. 3.9 (see App. A.2 for further details). ■

3.6.1 Bounding the sum of variances

The main terms appearing respectively in (3.54), (3.57) and (3.58) all have the form of a *sum of variances over time* $\sum_{t=1}^{T} \mathbb{V}_{p_t}(\alpha h_{k_t})$ with p_t a distribution over states (respectively $p_{k_t}(\cdot|s_t)$, $\overline{p}_{k_t}(\cdot|s_t)$ and $\widehat{p}_{k_t}(\cdot|s_t)$), and h_{k_t} the optimistic bias of episode k_t. A first *naïve* upper bound of this sum can be derived using Popoviciu's inequality that we recall in Prop. 3.10.

Proposition 3.10 (Popoviciu's inequality on variances)

Let M and m be upper and lower bounds on the values of a random variable X i.e., $\mathbb{P}(m \le X \le M) = 1$. Then $\mathbb{V}(X) \le \frac{1}{4}(M-m)^2$.

Using Popoviciu's inequality and under event E,

$$\mathbb{V}_{p_t}(\alpha h_{k_t}) \le sp(\alpha h_k)^2/4 = \alpha^2 sp(h_k)^2/4 \le \Lambda^2/4$$

and so $\sum_{t=1}^{T} \mathbb{V}_{p_t}(\alpha h_{k_t}) \le \Lambda^2 T/4$. Unfortunately, this would result in a regret bound scaling as $\widetilde{\mathcal{O}}(\Lambda\sqrt{T})$ (ignoring all other terms like S, A, logarithmic terms, etc.) which is *not better* than the bound of Thm. 3.4. In this section, we show that the cumulative sum of variances only scales as $\widetilde{\mathcal{O}}(\Lambda T + \Lambda^2\sqrt{T})$ resulting in a regret bound of order $\widetilde{\mathcal{O}}\left(\sqrt{\Lambda T} + \Lambda T^{1/4}\right)$ (ignoring all other terms).

We start by analyzing the variance term $\mathbb{V}_{\widehat{p}_k(\cdot|s_t)}(\alpha h_k)$. The other variance terms $\mathbb{V}_{p_k(\cdot|s_t)}(\alpha h_k)$ and $\mathbb{V}_{\overline{p}_k(\cdot|s_t)}(\alpha h_k)$ can be addressed in the same way. We do the following decomposition:

$$\begin{aligned}
\mathbb{V}_{\widehat{p}_k(\cdot|s_t)}(\alpha h_k) &= \alpha^2\left(\widehat{p}_k(\cdot|s_t)^\intercal h_k^2 - (\widehat{p}_k(\cdot|s_t)^\intercal h_k)^2\right) \\
&= \alpha^2\Big(\underbrace{(\widehat{p}_k(\cdot|s_t) - \overline{p_k}(\cdot|s_t))^\intercal h_k^2}_{(1)} + \underbrace{\overline{p_k}(\cdot|s_t)^\intercal h_k^2 - h_k^2(s_{t+1})}_{(2)} + \underbrace{h_k^2(s_{t+1}) - (\widehat{p}_k(\cdot|s_t)^\intercal h_k)^2}_{(3)}\Big)
\end{aligned}$$

Notice that for any r.v. X and any scalar $a \in \mathbb{R}$, $\mathbb{V}(X+a) = \mathbb{V}(X)$. Thus, the term $\mathbb{V}_{\widehat{p}_k(\cdot|s_t)}(\alpha h_k)$ remains unchanged when h_k is shifted by an arbitrary constant vector i.e., when h_k is replaced by $w_k := h_k + \lambda_k e$. As in Sec. 3.5.3, we minimize the ℓ_∞-norm of w_k by choosing $\lambda_k = -\frac{1}{2}(\max_{s\in\mathcal{S}} h_k(s) + \min_{s\in\mathcal{S}} h_k(s))$. We recall that under event E, $\|w_k\|_\infty \le \Lambda/(2\alpha)$ and so $\|w_k^2\|_\infty \le \Lambda^2/(4\alpha^2)$.

(1) The *first term* $\alpha^2 \sum_{k=1}^{k_T} \sum_{t=t_k}^{t_{k+1}-1} (\widehat{p}_k(\cdot|s_t) - \overline{p_k}(\cdot|s_t))^\intercal w_k^2$ is similar to $\sum_{k=1}^{k_T} \Delta_k^{p1}$ (see Sec. 3.5.3) except that αw_k is replaced by $\alpha^2 w_k^2$ and $p_k(\cdot|s_t)$ is replaced by $\widehat{p}_k(\cdot|s_t)$. In Sec. 3.5.3 we had to decompose $p_k(\cdot|s_t) - \overline{p_k}(\cdot|s_t)$ into the sum of $p_k(\cdot|s_t) - \widehat{p}_k(\cdot|s_t)$ and $\widehat{p}_k(\cdot|s_t) - \overline{p_k}(\cdot|s_t)$. Here we no longer need this decomposition and we can use the same derivation with $sp(\alpha^2 w_k^2) \le \Lambda^2/4$ instead. Therefore, with probability at least $1 - \frac{\delta}{6}$ (and

under event E):

$$\alpha^2 \sum_{k=1}^{k_T} \sum_{t=t_k}^{t_{k+1}-1} \left(\widehat{p}_k(\cdot|s_t) - \overline{p_k}(\cdot|s_t)\right)^\intercal w_k^2 \leq \frac{3}{2}\Lambda^2 \sqrt{\left(\sum_{s,a}\Gamma(s,a)\right) T \ln\left(\frac{6SAT}{\delta}\right)} + \Lambda^2 \sqrt{T \ln\left(\frac{5T}{\delta}\right)} + 3\Lambda^2 S^2 A \ln\left(\frac{6SAT}{\delta}\right)(1+\ln(T))$$

(2) The *second term* $\alpha^2 \sum_{k=1}^{k_T} \sum_{t=t_k}^{t_{k+1}-1} \overline{p_k}(\cdot|s_t)^\intercal w_k^2 - w_k^2(s_{t+1})$ is identical to $\sum_{k=1}^{k_T} \Delta_k^{p4}$ (see also Sec. 3.5.3) except that αw_k is replaced by $\alpha^2 w_k^2$. With probability at least $1 - \frac{\delta}{6}$ (and under event E):

$$\alpha^2 \sum_{k=1}^{k_T} \sum_{t=t_k}^{t_{k+1}-1} \overline{p_k}(\cdot|s_t)^\intercal w_k^2 - w_k^2(s_{t+1}) \leq \frac{\Lambda^2}{2} \sqrt{T \ln\left(\frac{5T}{\delta}\right)}$$

(3) The *last term* $\alpha^2 \sum_{k=1}^{k_T} \sum_{t=t_k}^{t_{k+1}-1} w_k^2(s_{t+1}) - \left(\widehat{p}_k(\cdot|s_t)^\intercal w_k\right)^2$ is the *dominant* one and requires more work. Unlike the first two terms, it scales *linearly* with T (instead of $\widetilde{\mathcal{O}}(\sqrt{T})$). We first notice that $\widehat{p}_k(\cdot|s_t)^\intercal w_k = w_k(s_t) + \widehat{p}_k(\cdot|s_t)^\intercal w_k - w_k(s_t)$. Using the fact that $(a+b)^2 = a^2 + b(2a+b)$ with $a = w_k(s_t)$ and $b = \widehat{p}_k(\cdot|s_t)^\intercal w_k - w_k(s_t)$ (and therefore $2a + b = w_k(s_t) + \widehat{p}_k(\cdot|s_t)^\intercal w_k$) we obtain:

$$\left(\widehat{p}_k(\cdot|s_t)^\intercal w_k\right)^2 = w_k^2(s_t) + \left(\widehat{p}_k(\cdot|s_t)^\intercal w_k - w_k(s_t)\right) \cdot \left(w_k(s_t) + \widehat{p}_k(\cdot|s_t)^\intercal w_k\right)$$

and so applying the *reverse triangle inequality*:

$$\left(\widehat{p}_k(\cdot|s_t)^\intercal w_k\right)^2 \geq w_k^2(s_t) - \left|\widehat{p}_k(\cdot|s_t)^\intercal w_k - w_k(s_t)\right| \cdot \left|w_k(s_t) + \widehat{p}_k(\cdot|s_t)^\intercal w_k\right| \tag{3.59}$$

For all $k \geq 1$ and $s \in \mathcal{S}$, we define $r_k(s) := \sum_a \pi_k(a|s) r_k(s,a)$. Using the (near-)optimality equation (see Sec. 3.5.2) we can write:

$$\left|g_k - r_k(s_t) + \alpha\left(w_k(s_t) - p_k(\cdot|s_t)^\intercal w_k\right)\right| = \left|g_k - r_k(s_t) + \alpha\left(h_k(s_t) - p_k(\cdot|s_t)^\intercal h_k\right)\right| \leq \varepsilon_k$$

Moreover, $\varepsilon_k = \frac{r_{\max}}{t_k} \leq r_{\max}$. As a result, since $\alpha > 0$:

$$\begin{aligned}
&\alpha\left|\widehat{p}_k(\cdot|s_t)^\intercal w_k - w_k(s_t)\right| \\
&\quad = \left|g_k - r_k(s_t) + \alpha\left(w_k(s_t) - p_k(\cdot|s_t)^\intercal w_k\right) - g_k + r_k(s_t) + \alpha\left(p_k(\cdot|s_t) - \widehat{p}_k(\cdot|s_t)\right)^\intercal w_k\right| \\
&\quad \leq \underbrace{\left|g_k - r_k(s_t) + \alpha\left(w_k(s_t) - p_k(\cdot|s_t)^\intercal w_k\right)\right|}_{\leq r_{\max}} + \underbrace{\left|r_k(s_t) - g_k\right|}_{\leq r_{\max}} + \alpha\left|\left(p_k(\cdot|s_t) - \widehat{p}_k(\cdot|s_t)\right)^\intercal w_k\right| \\
&\quad \leq 2r_{\max} + \alpha\left|\left(p_k(\cdot|s_t) - \widehat{p}_k(\cdot|s_t)\right)^\intercal w_k\right|
\end{aligned}$$

It is also immediate to see that $\left|w_k(s_t) + \widehat{p}_k(\cdot|s_t)^\intercal w_k\right| \leq 2\|w_k\|_\infty \leq \Lambda/\alpha$. Plugging these inequalities into (3.59) and adding $w_k^2(s_{t+1})$ we obtain:

$$\begin{aligned}
\alpha^2\left(w_k^2(s_{t+1}) - \left(\widehat{p}_k(\cdot|s_t)^\intercal w_k\right)^2\right) \leq &\left(2r_{\max} + \alpha\left|\left(p_k(\cdot|s_t) - \widehat{p}_k(\cdot|s_t)\right)^\intercal w_k\right|\right)\Lambda \\
&+ \alpha^2\left(w_k^2(s_{t+1}) - w_k^2(s_t)\right)
\end{aligned} \tag{3.60}$$

It is easy to bound the telescopic sum

$$\alpha^2 \sum_{t=t_k}^{t_{k+1}-1} w_k^2(s_{t+1}) - w_k^2(s_t) = \alpha^2 \left(w_k^2(s_{t_{k+1}}) - w_k^2(s_{t_k}) \right) \leq \alpha^2 w_k^2(s_{t_{k+1}}) \leq \Lambda^2/4 \tag{3.61}$$

Finally, the sum $\alpha \sum_{k=1}^{k_T} \sum_{t=t_k}^{t_{k+1}-1} |(p_k(\cdot|s_t) - \widehat{p}_k(\cdot|s_t))^\intercal w_k|$ can be bounded in the exact same way as $\sum_{k=1}^{k_T} \Delta_k^{p1}$ (see Sec. 3.5.3). With probability at least $1 - \frac{\delta}{6}$:

$$\begin{aligned}\alpha \sum_{k=1}^{k_T} \sum_{t=t_k}^{t_{k+1}-1} |(p_k(\cdot|s_t) - \widehat{p}_k(\cdot|s_t))^\intercal w_k| \leq & 3\Lambda \sqrt{\left(\sum_{s,a} \Gamma(s,a) \right) T \ln\left(\frac{6SAT}{\delta} \right)} + 4\Lambda \sqrt{T \ln\left(\frac{5T}{\delta} \right)} \\ & + 6\Lambda S^2 A \ln\left(\frac{6SAT}{\delta} \right) (1 + \ln(T))\end{aligned} \tag{3.62}$$

After gathering (3.61) and (3.62) into (3.60)) we conclude that with probability at least $1 - \frac{\delta}{6}$ (and under event E):

$$\alpha^2 \sum_{k=1}^{k_T} \sum_{t=t_k}^{t_{k+1}-1} w_k^2(s_{t+1}) - (\widehat{p}_k(\cdot|s_t)^\intercal w_k)^2 \leq \underbrace{2 r_{\max} \Lambda T}_{\text{main term}} + \frac{k_T \Lambda^2}{4} + \widetilde{\mathcal{O}}\left(\Lambda^2 \sqrt{\left(\sum_{s,a} \Gamma(s,a) \right) T} \right)$$

In conclusion, there exists an *absolute* numerical constant $\beta > 0$ (i.e., independent of the MDP instance) such that with probability at least $1 - \frac{5\delta}{6}$:

$$\sum_{t=1}^{T} \mathbb{V}_{\widehat{p}_{k_t}(\cdot|s_t)}(\alpha h_{k_t}) \leq \beta \cdot \left(r_{\max} \Lambda T + \Lambda^2 \sqrt{\left(\sum_{s,a} \Gamma(s,a) \right) T \ln\left(\frac{T}{\delta} \right)} + \Lambda^2 S^2 A \ln\left(\frac{T}{\delta} \right) \ln(T) \right).$$

We can prove the same bound (possibly with a different multiplicative constant β) for $\sum_{t=1}^{T} \mathbb{V}_{\overline{p}_{k_t}(\cdot|s_t)}(\alpha h_{k_t})$ and $\sum_{t=1}^{T} \mathbb{V}_{p_{k_t}(\cdot|s_t)}(\alpha h_{k_t})$ using the same derivation.

3.6.2 Completing the regret bound of Thm. 3.5

After plugging the bound derived for the sum of variances in the previous section (Sec. 3.6.1) into (3.54), (3.57) and (3.58), we notice that (3.54) and (3.58) can be upper-bounded by (3.57) *up to a multiplicative numerical constant* ans so it is enough to restrict attention to (3.57). The dominant term that we obtain is (ignoring numerical constants):

$$\sqrt{\left(\sum_{s,a} \Gamma(s,a) \right) \ln\left(\frac{T}{\delta} \right) \ln(T) \left(r_{\max} \Lambda T + \Lambda^2 \sqrt{\left(\sum_{s,a} \Gamma(s,a) \right) T \ln\left(\frac{T}{\delta} \right)} + \Lambda^2 S^2 A \ln\left(\frac{T}{\delta} \right) \ln(T) \right)}$$

Using the fact that $\sqrt{\sum_i a_i} \leq \sum_i \sqrt{a_i}$ for any $a_i \geq 0$, we can bound the above square-root

term by the sum of three simpler terms:

(1) A $\sqrt{T}$-term (dominant): $\sqrt{r_{\max}\Lambda\left(\sum_{s,a}\Gamma(s,a)\right)T\ln\left(\frac{T}{\delta}\right)\ln(T)}$

(2) A $T^{1/4}$-term: $\Lambda\left(\sum_{s,a}\Gamma(s,a)\right)^{3/4}T^{1/4}\left(\ln\left(\frac{T}{\delta}\right)\right)^{3/4}\sqrt{\ln(T)}$

(3) A logarithmic term: $\Lambda\sqrt{S^2A\left(\sum_{s,a}\Gamma(s,a)\right)\ln\left(\frac{T}{\delta}\right)\ln(T)} \leq \Lambda S^2 A\ln\left(\frac{T}{\delta}\right)\ln(T)$

When $T \geq \left(\frac{\Lambda}{r_{\max}}\right)^2\left(\sum_{s,a}\Gamma(s,a)\right)\ln\left(\frac{T}{\delta}\right)$, we notice that the $T^{1/4}$-term (2) is actually upper-bounded by the $\sqrt{T}$-term (1), while for $T \leq \left(\frac{\Lambda}{r_{\max}}\right)^2\left(\sum_{s,a}\Gamma(s,a)\right)\ln\left(\frac{T}{\delta}\right)$ we can use the following trivial upper-bound $r_{\max}T$ on the regret:

$$\Delta(\text{UCRLB},T) \leq r_{\max}T \leq \frac{\Lambda^2}{r_{\max}}\left(\sum_{s,a}\Gamma(s,a)\right)\ln\left(\frac{T}{\delta}\right) \leq \frac{\Lambda^2}{r_{\max}}S^2A\ln\left(\frac{T}{\delta}\right).$$

To complete the regret bound of Thm. 3.5 we also need to take into consideration (3.39) and (3.49) as well as the *lower order terms* of (3.54), (3.57) and (3.58). It turns out that the only terms that are not already upper-bounded by (1), (2) and (3) (up to multiplicative numerical constants) sum as:

$$r_{\max}\sqrt{SAT\ln\left(\frac{T}{\delta}\right)} + r_{\max}SA\ln\left(\frac{T}{\delta}\right)\ln(T) + \Lambda S^2A\ln\left(\frac{T}{\delta}\right)\ln(T)$$

If $\Lambda \leq r_{\max}$ then $\Lambda^2/r_{\max} \leq \Lambda \leq r_{\max}$, while if $\Lambda \geq r_{\max}$ then $\Lambda^2/r_{\max} \geq \Lambda \geq r_{\max}$. Therefore, alll the above logarithmic terms can be bounded by: $\max\left\{r_{\max}, \frac{\Lambda^2}{r_{\max}}\right\}S^2A\ln\left(\frac{T}{\delta}\right)\ln(T)$. Moreover, all the $\sqrt{T}$-terms can be bounded by

$$\max\left\{r_{\max}, \sqrt{r_{\max}\Lambda}\right\}\sqrt{\left(\sum_{s,a}\Gamma(s,a)\right)T\ln\left(\frac{T}{\delta}\right)\ln(T)}$$

To conclude, we only need to *adjust* δ to obtain an event of probability at least $1-\delta$. This will *only* impact the multiplicative numerical constants of the above terms.

3.7 Comparison between upper and lower-bounds

We recall the minimax lower-bound of Prop. 2.12: for *any* learning algorithm, it is possible to find a specific *worst-case* MDP for which the regret suffered is at least $\Omega(r_{\max}\sqrt{DSAT})$ *on expectation*. The intermediate MDP constructed by Jaksch et al. (2010, Figure 3) to prove Prop. 2.12 satisfies $\Lambda = r_{\max}D$ and so Prop. 2.12 can also be written as

$$\mathbb{E}\left[\Delta(M,\mathfrak{A},\mu_1,T)\right] \geq 0.015\cdot\sqrt{r_{\max}\Lambda}\sqrt{SAT} \tag{3.63}$$

i.e., D can be replaced by Λ.

The upper bound of Theorem. 3.5 (see Sec. 3.4) holds with probability $1-\delta$ but it is possible to obtain the same bound *in expectation* using the *law of total expectations* and setting $\delta = 1/\sqrt{T}$:

$$\mathbb{E}_M\left[\Delta_M(\mathfrak{A}, T)\right] = \widetilde{\mathcal{O}}\Bigg(\underbrace{(1-\delta)}_{\leq 1} \cdot \max\{r_{\max}, \sqrt{r_{\max}\Lambda}\}\sqrt{\sum_{s,a}\Gamma(s,a)T}\Bigg) + \underbrace{\delta \cdot r_{\max} T}_{\leq r_{\max}\sqrt{T}} \tag{3.64}$$

If we ignore multiplicative numerical constants and logarithmic terms, (3.63) matches the dominant term of (3.64) up to a factor $\sqrt{\Gamma}$. Unlike UCRL2, UCRLB is *minimax optimal* in Λ (or D). To the best of our knwoledge, it is the *first* bound with this property for the undiscounted infinite horizon setting. Although the dependency in S dropped from S (UCRL2) to $\sqrt{\Gamma S} \leq S$ (UCRLB), it is still not matching the lower bound (3.63).

Until very recently, it was still an *open question* of the literature whether $\sqrt{S}$ is achievable when $\Gamma = \Omega(S)$. Quite remarkably, the same question remained open in the discounted setting. Lattimore and Hutter (2014) indeed proved a $\widetilde{\Omega}\left(\frac{SA}{\varepsilon^2(1-\gamma)^3}\right)$ lower-bound on the sample complexity and derived an upper-bound matching the lower-bound up to a factor Γ.[10]

In the *finite horizon setting*, this question was answered by Azar et al. (2017); Kakade et al. (2018) who proved a regret bound of order $\widetilde{\mathcal{O}}\left(\sqrt{HSAT}\right)$ for their algorithm. Unfortunately, it is not easy to extend their approach to the infinite horizon case as it seems to heavily rely on the existence of a known horizon H.

In the *infinite horizon undiscounted setting*, there had been several notable attempts to try to fill the gap between lower and upper-bounds. For example, Agrawal and Jia (2017) initially claimed that the optimistic version of PSRL they designed incurs a regret bounded by $\widetilde{\mathcal{O}}\left(r_{\max} D\sqrt{SAT}\right)$. This improvement was obtained thanks to the use of tighter concentration inequalities proved by the same authors (Agrawal and Jia, 2017, Lemma C.1 & C.2). To better understand the main challenge of the proof, it is important to recall that the term $\sqrt{\Gamma}$ appears when bounding Δ_k^{p3} in the regret decomposition (see Sec. 3.5 and 3.6). Bounding this term requires to bound $(\widehat{p}_k(\cdot|s) - \overline{p_k}(\cdot|s))^\intercal h_k$ where $\widehat{p}_k$ is the estimated transition probability under policy π_k, $\overline{p}_k$ is the true transition probability under π_k and h_k is the optimistic bias at episode k. While for a fixed vector v, $(\widehat{p}_k(\cdot|s) - \overline{p_k}(\cdot|s))^\intercal v \lesssim sp\,(v)\sqrt{\frac{1}{N_k^+}}$ (Hoeffding bound), this concentration inequality may no longer hold when v and $\widehat{p}_k$ are correlated (which is the case for $v = h_k$). To overcome this issue, in the regret proof we used a *worst-case* bound:

$$\max_{sp(v)\leq D} (\widehat{p}_k(\cdot|s) - \overline{p_k}(\cdot|s))^\intercal v \lesssim sp\,(v)\sqrt{\frac{\Gamma}{N_k^+}}$$

which introduces $\sqrt{\Gamma}$ in the final regret bound. Agrawal and Jia (2017, Lemma C.2) claimed that the $\sqrt{\Gamma}$ could be removed in the above bound. Unfortunately, there seem to be a major mistake in the proof of both Lemma C.1 and Lemma C.2. We showed both theoretically

[10] We recall that a regret bound of order $C\sqrt{T}$ should be compared with a sample complexity bound of order $\frac{C^2}{(1-\gamma)^3\varepsilon^2}$ and D is comparable to $\frac{1}{1-\gamma}$.

and empirically an anti-concentration scaling linearly with $\sqrt{S}$ when $\Gamma = \Omega(S)$ (Qian et al., 2018a). This anti-concentration suggested that in order to remove the $\sqrt{\Gamma}$ factor, new arguments were needed that do not involve bounding $\max_{sp(v)\leq D} (\widehat{p}_k(\cdot|s) - \overline{p_k}(\cdot|s))^\intercal v$.

Despite the failed attempts, Tossou et al. (2019) seem to have finally solved this problem (the paper is still unpublished). Not long before, Ortner (2018) derived a Γ-free bound for ergodic MDPs.

Posterior sampling vs optimism. Agrawal and Jia (2017) points out that their Lemma C.1 is essentially Lemma 3 of Osband and Roy (2017) re-written. Osband and Roy (2017) used Lemma 3 to show a bound $\widetilde{\mathcal{O}}\left(H\sqrt{SAT}\right)$ on the *Bayesian regret* of PSRL for finite horizon problems. Unfortunately, the proof of Lemma 3 is also *mistaken* and our anti-concentration result also applies here. Osband and Roy (2017) further claimed that the improved S-dependency of their bound illustrates the *superiority* of *posterior sampling methods* over *OFU methods*: the latter will always suffer a regret scaling linearly with S while the former suffers a regret scaling linearly with $\sqrt{S}$. Our result *questions the validity* of this claim. Osband and Roy (2017) showed that their claim is empirically verified. However, they run UCRL2 which indeed suffers S due to the use of Hoeffding/Weissman bounds. Moreover, they run experiments on a family of MDPs (known as "River Swim") with increasing S but $\Gamma = 2$ in all MDPs. Therefore, on this specific family of MDPs, the regret of UCRLB will empirically grow as $\Theta(\sqrt{S})$ just like the regret of PSRL. The problem of the S dependency in the regret bound does not seem to be linked to the family of algorithm used (posterior sampling vs OFU).

3.8 Conclusion

In this chapter we introduced UCRLB, a variant of UCRL2 that leverages Bernstein concentration inequality to construct the confidence bounds used in the definition of the extended MDP. We showed that this simple modification allows to save a $\sqrt{DS/\Gamma}$ factor in the regret bound, implying that the best known minimax lower bound (Prop. 2.12) is somehow tight. We also generalized the notion of diameter by introducing the concept of travel-budget and made several contributions to the proof techniques used in the regret analysis of UCRL2-like algorithms. In the rest of the thesis, we will make an extensive use of all the material presented in this chapter in different contexts.

For future work, it would be very helpful to simplify and understand better the proof of Thm. 3.5 (second regret bound). For example, it could be insightful to provide a unified view of varince reduction methods in RL by relating our analysis to the other works mentioned in Sec. 3.6.

4 Exploration–exploitation in MDPs with infinite diameter

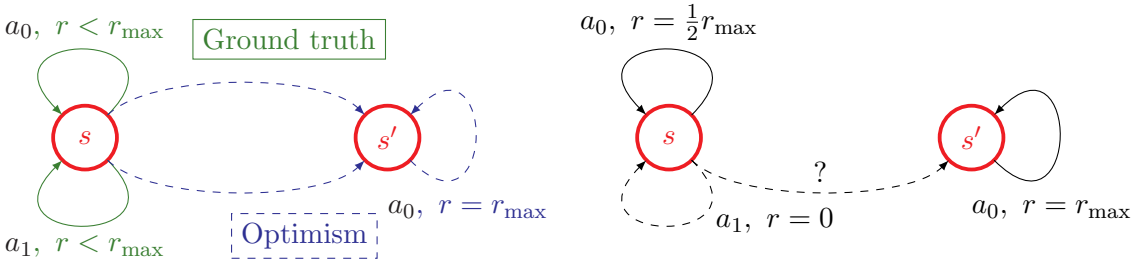
4.1 Introduction

4.1.1 Motivations

In the undiscounted infinite horizon setting, a major limitation of UCRL2-like algorithms is that the true unknown MDP M needs to be *communicating* i.e., its diameter D (see Def. 3.1) should be *finite*. For example, when $D = +\infty$, the regret bounds of Thm. 3.4 and 3.5 are worthless[1]. This is not just an artefact of the regret analysis as whenever $D = +\infty$, UCRLB (as well as UCRL2 and its variants) will indeed suffer a *linear* regret i.e., will *never learn*. One can easily verify this claim by running the algorithms on any non-communicating MDP, but this behaviour is more easily understood by looking at Example 1 of Ortner (2008). Their example (see Fig. 4.1a) is a slight modification of the stochastic *Multi-Armed Bandit* problem with only two arms/actions in state s –a_0 and a_1– that both have a reward strictly bounded by $r_{\max}$, and a third action that can only be played in a different state s'. If s' is not reachable from s, s' will never be visited and any UCRL2-like algorithm will expect to receive maximal reward $r_{\max}$ in that state (by optimism). As a consequence, it will always choose a model assigning as much probability mass as allowed by the confidence intervals to go from s to s' (by optimism). The "best" action to play in this optimistic model is the one that is expeced to cause a transition to s' with highest probability. In the optimistic model, the probability to go to s' when playing action a_i ($i \in \{0, 1\}$) decreases as the number of times the action is played ($N(s, a_i)$) increases. Therefore, the "best" action keeps changing: it is a_0 half of the time (when $N(s, a_0) < N(s, a_1)$), a_1 the other half (when $N(s, a_0) > N(s, a_1)$). The regret incured is therefore linear whenever the problem is non-trivial i.e., whenever $r(s, a_0) \neq r(s, a_1)$.

One might be tempted to think that the poor performance of UCRLB in the example of Fig. 4.1a is only a drawback of the algorithm and that the problem is not *intrinsically* more difficult than any RL task where $D < +\infty$. Let's slighltly modify the previous example (see

[1] When $D = +\infty$, there exists at least one state $s \in \mathcal{S}$ such that $\Pi^{\mathrm{SD}}_{\mapsto s} = \emptyset$ (see Def. 3.2) and so $\Lambda = +\infty$. Since the bounds of Thm. 3.4 and 3.5 scale linearly with respectively Λ and $\sqrt{\Lambda}$, they are worthless.



(a) Failure of UCRL2-like algorithms.

(b) Challenging exploration–exploitation.

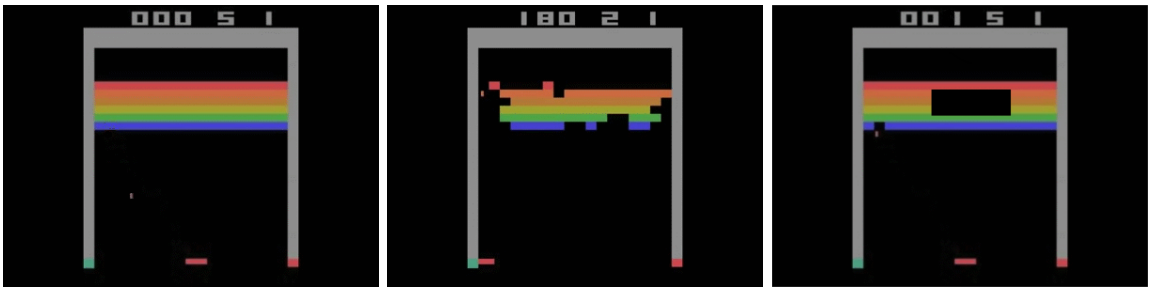
Figure 4.1: Examples inspired by (Ortner, 2008, Example 1). Fig. 4.1a illustrates why UCRLB fail to *learn* when some states are not reachable. Fig. 4.1b ilustrates the additional difficulty of the *exploration–exploitation* dilemma when the diameter is potentially *infinite*. In both examples, two actions can be played in state s (a_0 and a_1) and only one in state s' (a_0).

Fig. 4.1b): in state s, action a_0 yields reward $\frac{1}{2} r_{\max}$ and action a_1 yields reward 0 ; in state s', action a_0 yields reward $r_{\max}$. We further assume that the learning agent knows all the parameters of the MDP except the transition probability to move from s to s' after playing a_1 (dashed arrows on Fig. 4.1). State s' can only be *reached* when playing a_1 in s but might also *not* be reachable at all. If the probability to go to s' is *non-zero*, the agent should play a_1 in order to move to s' as quickly as possible. On the other hand, if the probability to go to s' is *zero*, then playing a_1 only increases the regret and a_0 should be played instead. Unfortunately, as long as no transition to s' has ever been observed, and no matter how many times action a_1 has already been played, the statement *"the probability to go to s' is non-zero"* can never be *refuted* as this probability can be *arbitrarily small*. In other words, while in state s and *independently* of past observations, it is impossible for the agent to *distinguish* between the two scenarios: arbitrarily low probability *versus absence* of a transition to s'. This is not specific to an algorithm but it is a *fundamental difficulty* of the learning problem. In the example of Fig. 4.1, an "efficient" algorithm should carefully balance the exploration of a_1 with the exploitation of a_0 while in s. When in addition the other parameters of the MDP are unknown, this comes as an extra "cost" compared to the usual exploration–exploitation trade-off that occurs when $D < +\infty$. In conclusion, the exploration-exploitation dilemma becomes *intrinsically* more challenging in non-communicating MDPs.

Notice that the problems descibed in Fig. 4.1 does not occur in the *discounted* or *finite horizon* settings since the exploration is directly tailored to the states that are reachable *within* the known horizon[2]. Then, it does not matter whether the transition to s' exists or not. It is sufficient to test whether the probability to go to s' is smaller than $1 - \gamma$ (discounted) or $1/H$ (finite horizon). This only requires to play a_1 for a *finite* number of times. The problem of Fig. 4.1 can also be overcome by leveraging on additional *prior knowledge* about the MDP (s.t. knowledge of the value of the smallest probability of transition, etc.) given to the learning agent. In this Chapter, we will assume that no such knowledge is available to the learning agent and we will analyse the general problem.

One might wonder whether the example of Fig. 4.1 is not *artificial* and whether MDPs with

[2]The discount factor γ implicitly defines an "horizon" of order $\frac{1}{1-\gamma}$.

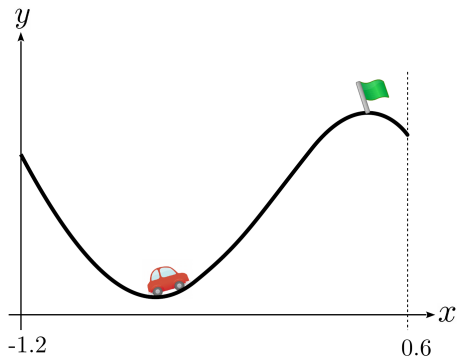


(a) Initial state s_1 (b) State reachable from s_1 (c) State *not* reachable from s_1

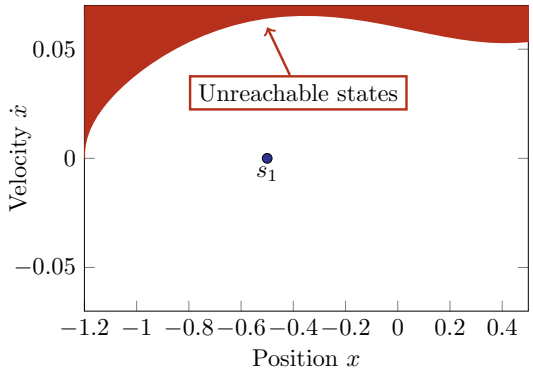
Figure 4.2: Example of *non-communicating* RL environment: game of Breakout (Mnih et al., 2015a). The state space is the set of all possible configurations of the brick wall, the paddle and the ball. Fig. a: The initial state of the game. Fig. b: An example of state reachable after playing the game for some time. Fig. c: An example of state *not* reachable from the initial state due to the presence of a *"hole"* in the brick wall.

non-reachable states (like s' on Fig. 4.1) are *frequently* encountered in RL. While assuming that all states are reachable may seem a reasonable assumption at first sight, it is rarely verified *in practice*. In fact, it requires a designer to carefully define a state space $\mathcal{S}$ that *contains* all reachable states (otherwise it may not be possible to learn the optimal policy), but that *excludes* unreachable states (otherwise the resulting MDP would be non-communicating). This requires a considerable amount of knowledge about the environment and its dynamics, and may be against the main purpose of RL which is to learn in an *unknown* environment with limited human supervision. Consider for example a problem where we learn from images e.g., the Atari Breakout game (Mnih et al., 2015a). A somehow simple *"intuitive"* state space could be the set of all *"plausible"* configurations of the brick wall, ball and paddle. The situation in which the wall has an hole in the middle is a valid state (e.g., as an initial state) but it cannot be observed/reached starting from a dense wall (see Fig. 4.2). As such, it should be removed to obtain a *"well-designed"* state space. While it may be possible to design a suitable set of reachable states that define a communicating MDP, this is often a difficult and tedious task, sometimes even impossible. Now consider a continuous domain e.g., the Mountain Car problem (Moore, 1990). The state is decribed by the position x and velocity $\dot{x}$ of the car along the x-axis. The state space of this domain is usually defined as the cartesian product $(x, \dot{x}) \in [-1.2, 0.6] \times [-0.07, 0.07]$. Unfortunately, this set contains configurations that are not physically reachable as shown on Fig. 4.3. The *dynamics* of the system is constrained by the *evolution equations* (law of motion). Therefore, the car can not go arbitrarily fast. On the leftmost position ($x = -1.2$) the speed $\dot{x}$ cannot exceed 0 because this position can be reached only with velocity $\dot{x} \leq 0$. To reach a higher velocity, the car would need to acquire momentum from further left (i.e., $x < -1.2$) which is impossible by design (-1.2 is the left-boundary of the position domain). The maximal speed reachable for $x > -1.2$ can be attained by applying the maximum acceleration at any time step starting from the state $(x, \dot{x}) = (-1.2, 0)$. This identifies the boundary of an *unreachable region* (red area on Fig. 4.3). Note that other states may not be reachable either.

As shown on the example of Fig. 4.1a, whenever the state space is *"misspecified"* or the MDP is non-communicating (i.e., $D = +\infty$), OFU-based algorithms (e.g., UCRLB) *optimistically* attribute large rewards and non-zero probability to reach states that have never



(a) Mountain Car domain



(b) Mountain Car state space

Figure 4.3: Example of *non-communicating* RL environment: Mountain Car (Moore, 1990; Brockman et al., 2016). Fig. a: The (red) car needs to reach the (green) flag on the top of the hill. The car does not have enough power and first needs to acquire momentum by reversing. Fig. 4.3b: x and $\dot{x}$ denote respectively the position and velocity of the car along the x-axis. The state space is defined by $(x, \dot{x}) \in [-1.2, 0.6] \times [-0.07, 0.07]$. The state labeled s_1 (in blue) corresponds to the initial state (car at the bottom of the hill at rest). From the *law of motion* of the car, it is possible to show that the sates in the red area can never be reached from s_1.

been observed, and thus they tend to repeatedly attempt to *explore* unreachable states. This results in poor performance and linear regret. In this chapter, we will describe and analyse an "efficient" algorithm that achieves a *sublinear* regret in both communicating and non-communicating MDPs without any *prior knowledge* on the diameter.

4.1.2 Previous work

Surprisingly, the problem of infinite diameter has received very little attention in the RL literature. The few papers dealing with this issue do not focus explicitly on this problem. They incidentally –and only partially– address it by attempting to solve a different –often more general– problem. Unfortunately, most of this literature is either incomplete (i.e., leaves a lot of open questions) or not fully accurate (e.g., makes questionable assumptions).

A first attempt to overcome the case $D = +\infty$ is Regal.C (Bartlett and Tewari, 2009) which requires prior knowledge of an upper-bound $c \geq 0$ to the span (i.e., range) of the optimal bias function h^* (this stting will be the focus of Chap. 5). The optimism of UCRL2 is then "constrained" to policies whose bias has span smaller than c. This implicitly "removes" non-reachable states, whose large optimistic reward would cause the span to become too large. Unfortunately, an accurate knowledge of the bias span may not be easier to obtain than designing a well-specified state space. Bartlett and Tewari (2009) proposed an alternative algorithm – Regal.D– that leverages on the *doubling trick* (Auer et al., 1995; Cesa-Bianchi and Lugosi, 2006) to avoid any prior knowledge on the span. Nonetheless, we noticed a major flaw in the proof of Bartlett and Tewari (2009, Theorem 3) that questions the validity of the algorithm (Fruit et al., 2018a, Appendix A). PS-based algorithms also suffer from similar issues. To the best of our knowledge, the only regret guarantees available

in the literature for this setting are[3] (Abbasi-Yadkori and Szepesvári, 2015; Ouyang et al., 2017b; Theocharous et al., 2017). However, the counter-example of Osband and Roy (2016) invalidates the result of Abbasi-Yadkori and Szepesvári (2015). On the other hand, Ouyang et al. (2017b) and Theocharous et al. (2017) present PS algorithms with expected *Bayesian* regret scaling linearly with c, where c is an upper-bound on the optimal bias spans of all the MDPs that can be drawn from the prior distribution ((Ouyang et al., 2017b, Asm. 1) and (Theocharous et al., 2017, Sec. 5)). Ouyang et al. (2017b, Remark 1) claim that their algorithm does not require the knowledge of c to derive the regret bound. However, in (Fruit et al., 2018a, Appendix B) we show on a very simple example that for most continuous prior distributions (e.g., commonly used uninformative priors like Dirichlet), it is very likely that $c = +\infty$ implying that the regret bound may not hold (and similarly for the work of Theocharous et al. (2017)). As a result, similarly to REGAL.C, the prior distribution should contain prior knowledge on the bias span to avoid poor performance.

In this chapter, we present TUCRL, an algorithm designed to trade-off exploration and exploitation in *weakly-communicating* and *multi-chain MDPs* (e.g., MDPs with misspecified state space) without any prior knowledge and under the only assumption that the agent starts from a state in a communicating subset of the MDP (Sec. 4.2). In communicating MDPs, TUCRL eventually (after a finite number of steps) performs as UCRL2, thus achieving problem-dependent logarithmic regret (Prop. 2.13). When the true MDP is weakly-communicating, we prove that TUCRL achieves a $\widetilde{O}(\sqrt{T})$ regret with *polynomial* dependency on the MDP parameters. We also show that it is not possible to design an algorithm achieving logarithmic regret in weakly-communicating MDPs without having an exponential dependence on the MDP parameters (see Sec. 4.5). TUCRL is the first *computationally tractable* algorithm in the OFU literature that is able to adapt to the MDP nature without any prior knowledge. The theoretical findings are supported by experiments on several domains (see Sec. 4.4).

The work presented in this chapter extends the conference paper (Fruit et al., 2018a).

4.2 Truncated Upper-Confidence RL (TUCRL)

4.2.1 Formalisation of the problem

In all this chapter, we relax the assumption that the true MDP M should be *communicating* (see Chap. 3). Instead, we only assume that M is *weakly communicating*. This is more general as communicating implies weakly communicating but not conversely. We recall the definition of a weakly communicating MDP in Def. 4.1 below (Puterman, 1994, Section 8.3.1

[3] We recall that the problem of weakly-communicating MDPs and misspecified states does not hold in the more restrictive setting of finite horizon (e.g., Osband et al., 2013) since exploration is directly tailored to the states that are reachable *within* the known horizon, or under the assumption of the existence of a recurrent state (e.g., Gopalan and Mannor, 2015). Therefore, we ignored this part of the literature.

and Proposition 8.3.1).

Definition 4.1 (Weakly communicating MDP)

An MDP $M = \{\mathcal{S}, \mathcal{A}, r, p\}$ is said to be weakly communicating *if the state space $\mathcal{S}$ can be* partioned *into two subsets, $\mathcal{S}^C$ and $\mathcal{S}^T$ (i.e., $\mathcal{S}^C \cap \mathcal{S}^T = \emptyset$ and $\mathcal{S}^C \cup \mathcal{S}^T = \mathcal{S}$), such that:*

1. *Every state in $\mathcal{S}^C$ is* accessible *from every other state in $\mathcal{S}^C$ under at least one deterministic stationary policy,*
2. *Either $\mathcal{S}^T$ is empty or every state in $\mathcal{S}^T$ is* transient *under every policy.*

Equivalently, M is weakly communicating if and only if the Markov Chain induced by any stationary policy that plays every action with non-zero probability is unichain. *Under such policy, all states in $\mathcal{S}^C$ are* recurrent *while, all states in $\mathcal{S}^T$ are* transient.

By definition, the states in $\mathcal{S}^{\mathtt{T}}$ are not accessible from the states in $\mathcal{S}^{\mathtt{C}}$ and so it is possible to *restrict* the state space $\mathcal{S}$ to $\mathcal{S}^{\mathtt{C}}$ while still preserving the "properties" of an MDP. The MDP defined on the restricted state space $\mathcal{S}^{\mathtt{C}}$ is always communicating by definition and we denote by $D^{\mathtt{C}}$ its diameter i.e.,

$$D^{\mathtt{C}} := \max_{(s,s') \in \mathcal{S}^{\mathtt{C}} \times \mathcal{S}^{\mathtt{C}}} \min_{\pi \in \Pi^{\mathrm{SD}}} \mathbb{E}^{\pi}[\tau(s') | s_1 = s] - 1 \tag{4.1}$$

where $\tau(s') := \inf\{t \geq 1 : s_t = s'\}$ is the first hitting time of s' (see Sec. 3.3). Similarly, we denote by Λ its travel-budget i.e.,

$$\Lambda^{\mathtt{C}} := \max_{s,s' \in \mathcal{S}^{\mathtt{C}} \times \mathcal{S}^{\mathtt{C}}} \min_{\pi \in \Pi^{\mathrm{SD}}} \mathbb{E}^{\pi} \left[\sum_{t=1}^{\tau(s')-1} r_{\max} - r(s_t, \pi(s_t)) \middle| s_1 = s \right] \tag{4.2}$$

where the sum should be interpreted as a Cesaro limit when $\mathbb{P}(\tau(s') = +\infty | s) < 1$. We denote by $S^{\mathtt{C}} = |\mathcal{S}^{\mathtt{C}}|$ (resp. $S^{\mathtt{T}} = |\mathcal{S}^{\mathtt{T}}|$) the number of states in $\mathcal{S}^{\mathtt{C}}$ (resp. $\mathcal{S}^{\mathtt{T}}$). $\Gamma^{\mathtt{C}} = \max_{s \in \mathcal{S}^{\mathtt{C}}, a \in \mathcal{A}} \|p(\cdot|s,a)\|_0$ is the maximum support of all transition probabilities $p(\cdot|s,a)$ with $s \in \mathcal{S}^{\mathtt{C}}$. As in Chap. 3, the state and action spaces –$\mathcal{S}$ and $\mathcal{A}$– are still assumed to be finite, and the rewards are assumed to lie in $[0, r_{\max}]$.

Learning problem. Similarly to Chap. 3, we consider the learning problem where $\mathcal{S}$, $\mathcal{A}$ and $r_{\max}$ are *known*, while sets $\mathcal{S}^{\mathtt{C}}$ and $\mathcal{S}^{\mathtt{T}}$, rewards r and transition probabilities p are *unknown* and need to be estimated on-line. As shown by Puterman (1994, Theorem 8.3.2), the following proposition holds

Proposition 4.1

In any weakly-communicating MDP, the optimal gain g^* is state independent.

Since g^* is *state-independent*, we can still evaluate the performance of a learning algorithm $\mathfrak{A}$ by its cumulative *regret* $\Delta(\mathfrak{A}, T) = \sum_{t=1}^{T} g^* - r_t$. Furthermore, we state the following assumption:

Assumption 4.1

The initial state s_1 belongs to the communicating subset of states, i.e., $s_1 \in \mathcal{S}^C$.

While this assumption somehow restricts the scenario we consider, it is fairly common in practice. For example, all the domains that are characterized by the presence of a resetting distribution (e.g., episodic problems) satisfy this assumption (e.g., Mountain Car, Cart Pole, Atari games, taxi, etc.). Under Asm. 4.1, $D^{\mathtt{C}} < +\infty$.

Multi-chain MDPs. While we consider weakly-communicating MDPs for ease of notation, all the results presented in this Chapter extend to the more general case of multi-chain MDPs.[4] In this case, there may be multiple communicating and transient sets of states and the optimal gain g^* is different in each communicating subset. We then define $\mathcal{S}^{\mathtt{C}}$ as the set of states that are accessible –with non-zero probability– from the initial state s_1 (s_1 included) under some stationary deterministic policy. $\mathcal{S}^{\mathtt{T}}$ is defined as the complement of $\mathcal{S}^{\mathtt{C}}$ in $\mathcal{S}$ i.e., $\mathcal{S}^{\mathtt{T}} := \mathcal{S} \setminus \mathcal{S}^{\mathtt{C}}$. With these new definitions of $\mathcal{S}^{\mathtt{C}}$ and $\mathcal{S}^{\mathtt{T}}$, Asm. 4.1 needs to be reformulated as follows:

Assumption 4.2 (Equivalent of Asm. 4.1 for Multi-chain MDPs.)

The initial state s_1 is accessible from any other state in $\mathcal{S}^{\mathtt{C}}$ under some stationary deterministic policy. Equivalently, $\mathcal{S}^{\mathtt{C}}$ is a communicating set of states (i.e., $D^{\mathtt{C}} < +\infty$).

Note that the states belonging to $\mathcal{S}^{\mathtt{T}}$ can either be transient or belong to other communicating subsets of the MDP disjoint from $\mathcal{S}^{\mathtt{C}}$. It does not really matter because the states in $\mathcal{S}^{\mathtt{T}}$ will never be visited by definition. As a result, the regret is still defined as before, where the learning performance is compared to the optimal gain $g^*(s_1)$ related to the communicating set of states $\mathcal{S}^{\mathtt{C}} \ni s_1$. We highlight that $g^*(s_1) = g^*(s)$ for all $s \in \mathcal{S}^{\mathtt{C}}$.

4.2.2 Algorithm

In this section we present our solution to the problem of learning in an MDP with infinite diameter. We introduce Truncated Upper-Confidence for Reinforcement Learning (TUCRL), an optimistic online RL algorithm that efficiently balances exploration and exploitation in non-communicating MDPs without prior knowledge. Because TUCRL is very similarly to UCRLB (same structure, confidence bounds, etc.), we do not repeat the full pseudo-code of Alg. 5 and only stress the differences between the two algorithms which i.e., the extended MDP constructed at each episode and the stopping condition of an episode. We recall that the extended MDP constructed by UCRLB is denoted $\mathcal{M}_k$ (Eq. 3.3 and 3.4).

Estimation of reachable states. UCRLB is *optimistic* w.r.t. the confidence intervals so that for all states s that have never been visited (i.e., s.t. $\sum_a N_k(s,a) = 0$), the optimistic reward $r_k(s,a)$ will automatically be set to $r_{\max}$ *by optimism* (see example on Fig. 4.1a), while all transitions to s are set to the largest value compatible with B_p^k. Unfortunately, some of

[4] This is the most general category of MDPs that we can define (Puterman, 1994, Section 8.3.1)). It includes all possible MDPs.

the states with $\sum_a N_k(s,a) = 0$ may actually be *unreachable* (i.e., $s \in \mathcal{S}^{\mathtt{T}}$) and UCRLB would uniformly explore the policy space with the hope that at least one policy reaches those (optimistically desirable) states with non-zero probability (see example on Fig. 4.1a). TUCRL addresses this issue by first constructing empirical estimates of $\mathcal{S}^{\mathtt{C}}$ and $\mathcal{S}^{\mathtt{T}}$ (i.e., the set of communicating and transient states, see Sec. 4.2.1) using the states that have been visited so far, that is

$$\mathcal{S}_k^{\mathtt{C}} := \left\{ s \in \mathcal{S} : \sum_{a \in \mathcal{A}_s} N_k(s,a) > 0 \right\} \cup \{s_{t_k}\} \quad \text{and} \quad \mathcal{S}_k^{\mathtt{T}} := \mathcal{S} \setminus \mathcal{S}_k^{\mathtt{C}} \tag{4.3}$$

where we recall that t_k is the starting time of episode k (see Eq. 3.11). All states in $\mathcal{S}_k^{\mathtt{C}}$ are *for sure* reachable from s_1 and so under Asm. 4.1 (or Asm. 4.2), $\mathcal{S}_k^{\mathtt{C}} \subseteq \mathcal{S}^{\mathtt{C}}$. In the rest of this chapter, we will denote by $S_k^{\mathtt{C}}$ (resp. $S_k^{\mathtt{T}}$) the cardinal of $\mathcal{S}_k^{\mathtt{C}}$ (resp. $\mathcal{S}_k^{\mathtt{T}}$).

Truncated transition probabilities. In order to avoid that optimism drives the algorithm into attempting to reach unreachable states, we could simply execute UCRLB on $\mathcal{S}_k^{\mathtt{C}}$, which is *guaranteed* (by design and under Asm. 4.1 or 4.2) to contain only states in the communicating set $\mathcal{S}^{\mathtt{C}}$. Nonetheless, with such a strategy, the algorithm could *under-explore* some state-action pairs that would allow discovering other states in $\mathcal{S}^{\mathtt{C}}$, thus getting *stuck* in a *strict subset* of $\mathcal{S}^{\mathtt{C}}$ and suffering *linear regret*. While the states in $\mathcal{S}_k^{\mathtt{C}}$ are guaranteed to be in $\mathcal{S}^{\mathtt{C}}$, it is not possible to know whether the states in $\mathcal{S}_k^{\mathtt{T}}$ are actually reachable from $\mathcal{S}_k^{\mathtt{C}}$ or not (see the example of Fig. 4.1b and the impossibility to distinguish between a zero and arbitrarily small transition probability). To account for the eventuality that some states in $\mathcal{S}_k^{\mathtt{T}}$ actually belong to $\mathcal{S}^{\mathtt{C}}$, TUCRL first *"guesses"* a lower bound on the probability of transition from states $s \in \mathcal{S}_k^{\mathtt{C}}$ to $s' \in \mathcal{S}_k^{\mathtt{T}}$ and whenever the maximum transition probability from s to s' compatible with the confidence intervals (i.e., $\min\{1, \widehat{p}_k(s'|s,a) + \beta_{p,k}^{sas'}\}$, see Alg. 5) is below the lower bound, it assumes that such transition is not possible. This strategy is based on the intuition that a transition either does not exist or it should have a sufficiently "big" mass. However, these transitions should be *periodically* reconsidered in order to avoid *under-exploration* issues. More formally, let $(\rho_t(s,a))_{t \geq 1}$ be positive non-increasing sequences to be defined later. For all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we define $p_k^+(s'|s,a)$ to be the *largest* (i.e., most optimistic) probability of transition from s to s' through action a that belongs to $B_p^k(s,a,s')$ (see Eq. 3.3 in Alg. 5) i.e.,

$$p_k^+(s'|s,a) := \max_{p \in B_p^k(s,a,s')} \{p\}\,. \tag{4.4}$$

For all $s' \in \mathcal{S}_k^{\mathtt{T}}$, $s \in \mathcal{S}_k^{\mathtt{C}}$ and $a \in \mathcal{A}_s$, the empirical mean $\widehat{p}_k(s'|s,a)$ and variance $\widehat{\sigma}_{p,k}^2(s'|s,a)$ are by definition zero (since this transition has never been observed so far, see Eq. 3.7 and 3.9), so that $p_k^+(s'|s,a) = \min\left\{1, \frac{6\ln(6SAN_k^+(s,a)/\delta)}{N_k^+(s,a)}\right\}$ (see Eq. 3.1 and. 3.3). Since in that case $p_k^+(s'|s,a)$ does not depend on s', we will drop the dependency on the next state and write $p_k^+(s,a) := \min\left\{1, \frac{6\ln(6SAN_k^+(s,a)/\delta)}{N_k^+(s,a)}\right\}$. For all $(s,a) \in \mathcal{S}_k^{\mathtt{C}} \times \mathcal{A}_s$, TUCRL compares $p_k^+(s,a)$ to $\rho_{t_k}(s,a)$ and, whenever the latter is strictly bigger than the former, forces all transition probabilities to $\mathcal{S}_k^{\mathtt{T}}$ to be zero (i.e., whenever $p_k^+(s,a) < \rho_{t_k}(s,a)$, $p_k(s'|s,a) \leftarrow 0$ for all $s' \in$

$\mathcal{S}_k^{\mathsf{T}}$). The confidence intervals of all other transitions are kept unchanged. This corresponds to constructing the alternative *restricted* confidence intervals

$$\overline{B}_p^k(s,a,s') := \begin{cases} \{0\} \text{ if } s \in \mathcal{S}_k^{\mathsf{C}},\ p_k^+(s,a) < \rho_{t_k}(s,a), \text{ and } s' \in \mathcal{S}_k^{\mathsf{T}}, \text{ otherwise:} \\ B_p^k(s,a,s') = \left[\widehat{p}_k(s'|s,a) - \beta_{p,k}^{sas'}, \widehat{p}_k(s'|s,a) + \beta_{p,k}^{sas'}\right] \cap [0,1]. \end{cases} \tag{4.5}$$

With the new confidence sets $\overline{B}_p^k(s,a,s')$, Thm. 3.1 of Chap. 3 no longer holds as some of the probabilities set to 0 might actually be non-zero in the true MDP. In this case, it may be difficult to relate the optimistic bias h_k with the travel-budget of the true MDP (see Sec. 3.3). To overcome this issue, we slightly increase the confidence intervals $B_p^k(s,a,s')$. For all $s \in \mathcal{S}$ and $a \in \mathcal{A}_s$ we define

$$\zeta_{p,k}^{sa} := \sum_{s' \in \mathcal{S}_k^{\mathsf{T}}} p_k^+(s'|s,a) = S_k^{\mathsf{T}} \cdot p_k^+(s,a) \tag{4.6}$$

$\zeta_{p,k}^{sa}$ simply corresponds to the *maximal cumulative probability mass* that could be assigned to the transition $(s,a) \to \mathcal{S}_k^{\mathsf{T}}$ if we were using the same confidence intervals $B_p^k(s,a,s')$ as in UCRLB. In TUCRL, for all $(s,a) \in \mathcal{S}_k^{\mathsf{C}} \times \mathcal{A}$ such that $p_k^+(s,a) < \rho_{t_k}(s,a)$, the probability $p_k(s'|s,a)$ is set to 0 for all $s' \in \mathcal{S}_k^{\mathsf{T}}$. We thus *redistribute* this "optimistic probability mass" on all other states. This amounts to defining the following confidence intervals (for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$)

$$\overline{Z}_p^k(s,a,s') :=:= \begin{cases} B_p^k(s,a,s') \text{ if } s \in \mathcal{S}_k^{\mathsf{T}}, \\ B_p^k(s,a,s') \text{ if } s \in \mathcal{S}_k^{\mathsf{C}} \text{ and } p_k^+(s,a) \geq \rho_{t_k}(s,a), \\ \{0\} \text{ if } s \in \mathcal{S}_k^{\mathsf{C}},\ p_k^+(s,a) < \rho_{t_k}(s,a), \text{ and } s' \in \mathcal{S}_k^{\mathsf{T}}, \\ \left[\widehat{p}_k(s'|s,a) - \beta_{p,k}^{sas'}, \widehat{p}_k(s'|s,a) + \beta_{p,k}^{sas'} + \zeta_{p,k}^{sa}\right] \cap [0,1] \text{ otherwise.} \end{cases} \tag{4.7}$$

With the new confidence intervals $\overline{Z}_p^k(s,a,s')$, we will show that the travel-budget of the associated extended MDP is bounded by the travel-budget of $\mathcal{M}_k$ (the extended MDP constructed by UCRLB) which is itself bounded (with high probability) by the travel-budget of the true MDP as shown in Chap. 3. Moreover, the increase of $\zeta_{p,k}^{sa}$ in the confidence bounds only impacts the logarithmic terms of the regret bound since $\zeta_{p,k}^{sa} \leq \frac{6S\ln\left(6SAN_k^+(s,a)/\delta\right)}{N_k^+(s,a)}$. Finally, the confidence intervals of the rewards $B_r^k(s,a)$ will remain unchanged.

Extended value iteration. With some transitions set to 0, it is possible that the associated extended MDP is not *communicating* and not even *weakly-communicating*. The gain of such an MDP is *not* necessarily *state-independent*. Therefore, Lem. 2.7 (see Sec. 3.1.2) no longer holds and the stopping condition of EVI (Alg. 3) can no longer be used. This problem can be fixed by *restricting* the state space of the extended MDP to the set $\mathcal{S}_k^{\text{EVI}}$ defined as the set of states that are *reachable* from the communicating set $\mathcal{S}_k^{\mathsf{C}}$. Since by design (see Eq. 4.7) all states in $\mathcal{S}$ are reachable from $\mathcal{S}_k^{\mathsf{T}}$, in practice there are only *two possible cases*: either all the transitions from $\mathcal{S}_k^{\mathsf{C}}$ to $\mathcal{S}_k^{\mathsf{T}}$ are forbidden in which case $\mathcal{S}_k^{\text{EVI}} = \mathcal{S}_k^{\mathsf{C}}$, otherwise $\mathcal{S}_k^{\text{EVI}} = \mathcal{S}$. Formally, we have:

$$\mathcal{S}_k^{\mathrm{EVI}} := \begin{cases} \mathcal{S}_k^{\complement} \text{ if for all } (s,a) \in \mathcal{S}_k^{\complement} \times \mathcal{A}_s,\ p_k^+(s,a) < \rho_{t_k}(s,a) \\ \mathcal{S} \text{ otherwise} \end{cases} \tag{4.8}$$

We can now define the extended MDP $\overline{\mathcal{M}}_k$ as

$$\overline{\mathcal{M}}_k := \left\{ \mathcal{S}_k^{\mathrm{EVI}},\ \mathcal{A},\ r_k(s,a) \in B_r^k(s,a),\ p_k(s'|s,a) \in \overline{Z}_p^k(s,a,s') \right\} \tag{4.9}$$

Compared to the extended MDP $\mathcal{M}_k$ constructed by UCRLB, only the state space (4.8) and the confidence intervals of transition probabilities (4.7) change. By construction, $\overline{\mathcal{M}}_k$ is always communicating and so its optimal gain is constant, EVI is guaranteed to converge and Lem. 2.7 applies. TUCRL executes EVI on the extended MDP $\overline{\mathcal{M}}_k$. In TUCRL, line 9 of Alg. 5 (Eq. 3.5) is replaced by

$$(g_k, h_k, \pi_k) := \mathrm{EVI}\left(\overline{\mathcal{L}}_\alpha^k, \overline{\mathcal{G}}_\alpha^k, \frac{r_{\max}}{t_k}, 0, s_1 \right) \tag{4.10}$$

where $\overline{\mathcal{L}}_\alpha^k$ denotes the optimal Bellman operator of $\overline{\mathcal{M}}_k$ with aperiodicity transformation of parameter $\alpha \in]0,1]$ (and $\overline{\mathcal{G}}_\alpha^k$ is the associated greedy operator). We will also denote by p_k and r_k the transition probabilities and rewards satisfying

$$\forall s \in \mathcal{S},\quad \overline{\mathcal{L}}_\alpha^k h_k(s) = \sum_{a \in \mathcal{A}_s} \pi_k(s,a) r_k(s,a) + \alpha \sum_{a \in \mathcal{A}_s} \sum_{s' \in \mathcal{S}} \pi_k(s,a) p_k(s'|s,a) h_k(s') + (1-\alpha) h_k(s).$$

Finally, we denote by (g_k^*, h_k^*) a solution of the Bellman optimality equation $\overline{\mathcal{L}}_\alpha^k h_k^* = h_k^* + h_k^* e$. Since Lem. 2.7 holds, $g_k \geq g_k^* - \frac{r_{\max}}{t_k}$.

Stopping condition of episodes. Besides the change in the *definition of* B_p^k, the *stopping condition* of episodes is also slightly modified compared to UCRLB (line 12 in Alg. 5). In addition to ending the current episode as soon as $\nu_k(s_t,a_t) \geq N_k^+(s_t,a_t)$, TUCRL also stops whenever $\sum_a N_k(s_{t+1},a) = 0$. Equivalently, TUCRL forces an episode to terminate as soon as a state previously in $\mathcal{S}_k^{\mathsf{T}}$ is visited (the state is then added to $\mathcal{S}_k^{\complement}$). In TUCRL, line 12 of Alg. 5 is then rewritten as:

$$\textbf{if } \nu_k(s_t,a_t) \geq N_k^+(s_t,a_t) \textbf{ or } \sum_a N_k(s_{t+1},a) = 0 \ \left(\iff \ s_{t+1} \in \mathcal{S}_k^{\mathsf{T}} \right) \tag{4.11}$$

This minor change guarantees that for every episode $k \geq 1$ and for all the states $s \in \mathcal{S}_k^{\mathsf{T}}$ and all actions $a \in \mathcal{A}_s$, we have $N_k(s,a) = 0$ (when the condition is about to be violated, episode k stops). Furthermore, the number of epsiodes is hardly impacted as we will see.

Communicating MDPs. In the next section, we will show that under Asm. 4.1 (or Asm. 4.2), and with a carefully tuned sequences $(\rho_t(s,a))_{t \geq 1}$, TUCRL is always able to learn i.e., to achieve sublinear regret: $\Delta(\mathrm{TUCRL}, T) = o(T)$. When the true MDP is *communicating*, this means that *all* states are eventually *visited* at least once and so there exists an episode

$\overline{k}$ s.t. for all $k \geq \overline{k}$, $\mathcal{S}_k^{\mathsf{T}} = \emptyset$. When this condition is met, we notice that $\mathcal{S}_k^{\mathrm{EVI}} = \mathcal{S}$ and $\overline{Z}_p^k(s,a,s') = B_p^k(s,a,s')$ for all 3-tuple $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ (since $\zeta_{p,k}^{sa} = 0$). So for all $k \geq \overline{k}$, $\overline{\mathcal{M}}_k = \mathcal{M}_k$ (see Eq. 4.9) and the condition $\sum_a N_k(s_{t+1},a) = 0$ is always false meaning that the stopping condition of episodes implemented by TUCRL (4.11) is the same as the one implemented by UCRLB. For all $k \geq \overline{k}$, TUCRL naturally *reduces* to UCRLB. This seems reasonable since UCRLB is known to *efficiently* learn under the *prior knowledge* that the MDP is communicating. When $\mathcal{S}_k^{\mathsf{T}} = \emptyset$, this prior knowledge is not needed and is *automatically deduced* from the observations.

Sequences of thresholds. In practice, we set

$$\rho_t(s,a) := \min\left\{1, \frac{6\ln\left(\frac{6SAN_{k_t}^+(s,a)}{\delta}\right)}{N_{k_t}^+(s,a)}\right\} \cdot N_{k_t}^+(s,a) \cdot \sqrt{\frac{SA}{t}} \tag{4.12}$$

for all $t \geq 1$, so that the *condition to remove transition* reduces to $N_k^+(s,a) > \sqrt{t_k/SA}$. This shows that only transitions from state-action pairs that have been *poorly visited* so far are enabled, while if the state-action pair has already been tried *often* and yet no transition to $s' \in \mathcal{S}_k^{\mathsf{T}}$ is observed, then it is assumed that s' is *not reachable* from (s,a). When the number of visits in (s,a) is big, the transitions to *"unvisited"* states ($\mathcal{S}_k^{\mathsf{T}}$) should be discarded because if the transition actually exists, it is most likely extremely *small* and so it is worth exploring *other parts* of the MDP first. Symmetrically, when the number of visits in (s,a) is small, the transitions to "unvisited" states should be *enabled* because the transitions are quite *plausible* and the algorithm should try to explore the outcome of taking action a in s and possibly reach states in $\mathcal{S}_k^{\mathsf{T}}$. We denote the set of state-action pairs that are not sufficiently explored by

$$\mathcal{E}_k := \left\{(s,a) \in \mathcal{S}_k^{\mathsf{C}} \times \mathcal{A} \,:\, N_k^+(s,a) \leq \sqrt{\frac{t_k}{SA}}\right\}. \tag{4.13}$$

Executed policy π_k. The policy π_k may be stochastic but all actions that are played with non-zero probability satisfy the (near-)optimality equation. This will simplify the regret proof compared to Chap. 3.

4.3 Analysis of TUCRL

4.3.1 Optimistic gain and bias

Gain-optimism

The first technical difficulty in the analysis of TUCRL is that whenever some transitions are *disabled* (i.e., forced to be 0), the plausible set of MDPs $\overline{\mathcal{M}}_k$ may actually be *biased* and not

contain the true MDP M. In other words, Thm. 3.1 does not hold for $\overline{\mathcal{M}}_k$ (i.e., it is possible that $M \notin \overline{\mathcal{M}}_k$ for at least one $k \geq 1$ with probability strictly bigger than $\frac{\delta}{3}$). However, since $\mathcal{M}_k$ is still defined as in Chap. 3, Thm. 3.1 still holds for $\mathcal{M}_k$ i.e., $M \in \mathcal{M}_k$ for all $k \geq 1$ with probability at least $1 - \frac{\delta}{3}$. We denote by E this high probability event as in Chap. 3. In this section we prove that TUCRL is always gain-optimistic (i.e., $g_k^* \geq g^*$) despite "wrong" confidence intervals $\overline{Z}_p^k$ (4.7). A first approach would be to use Prop. 3.3 as suggested in Sec. 3.2.1. Intuitively, the "truncation" of the confidence intervals operated by TUCRL only *perturbs* the vector $\overline{\mathcal{L}}_\alpha^k h^*$ by a term of order $\eta_k \sim sp(h^*)\sqrt{\frac{SA}{t_k}\ln\left(\frac{SAt_k}{\delta}\right)}$ compared to $\mathcal{L}_\alpha^k h^*$ i.e., $\overline{\mathcal{L}}_\alpha^k h^* \geq Lh^* - \eta_k e$ and so $g_k^* \geq g^* - \eta_k$ (see Sec. 3.2.1). The problem is that the additional regret created by the term $\sum_{k=1}^{k_T}(t_{k+1} - t_k)\eta_k$ is of order $\Theta\left(sp(h^*)\, S^2 A\sqrt{T\ln\left(\frac{T}{\delta}\right)}\right)$ in the worst case. In order to avoid such a bad dependency in S and A in the regret bound, we rely on completely different arguments to prove optimism. The following lemma helps to identify the possible scenarios that TUCRL can produce.

Lemma 4.1

Let episode k be such that $M \in \mathcal{M}_k$, $\mathcal{S}_k^T \neq \emptyset$ and

$$t_k \geq C_k := 36 \cdot \left(D^{\mathtt{C}}\right)^2 \cdot SA \cdot \left(S_k^T\right)^2 \cdot \ln\left(\frac{6SAt_k}{\delta}\right)^2. \tag{4.14}$$

Then, either $\mathcal{S}_k^T = \mathcal{S}^T$ *(case I)* or $\mathcal{E}_k \neq \emptyset$, *i.e.,* $\exists (s,a) \in \mathcal{S}_k^{\mathtt{C}} \times \mathcal{A}$ *for which transitions to* $\mathcal{S}_k^T$ *are allowed (case II).*

Proof. We prove the result by showing that under the assumptions of Lem. 4.1, we have the implication $\mathcal{E}_k = \emptyset \implies \mathcal{S}_k^{\mathtt{T}} = \mathcal{S}^{\mathtt{T}}$. Assume that episode k is such that inequality (4.14) holds and that $M \in \mathcal{M}_k$, $\mathcal{S}_k^{\mathtt{T}} \neq \emptyset$ and $\mathcal{E}_k = \emptyset$ i.e., for any state-action pair $(s,a) \in \mathcal{S}_k^{\mathtt{C}} \times \mathcal{A}_s$

$$N_k^+(s,a) > \sqrt{\frac{t_k}{SA}} \geq \sqrt{\frac{C_k}{SA}} = 6D^{\mathtt{C}} S_k^{\mathtt{T}} \ln\left(\frac{6SAt_k}{\delta}\right).$$

Since $\mathcal{S}_k^{\mathtt{T}} \neq \emptyset$ and $M \in \mathcal{M}_k$, for any $(s,a,s') \in \mathcal{S}_k^{\mathtt{C}} \times \mathcal{A}_s \times \mathcal{S}_k^{\mathtt{T}}$, $p(s'|s,a) \in B_p^k(s,a,s')$ implying

$$\underbrace{p(s'|s,a)}_{\text{transition probability in } M} \leq \underbrace{\widehat{p}_k(s'|s,a)}_{=0} + \beta_{p,k}^{sas'} = \underbrace{2\sqrt{\frac{\widehat{\sigma}_{p,k}^2(s'|s,a)\ln(6SAt_k/\delta)}{N_k^+(s,a)}}}_{=0} + \frac{6\ln\left(\frac{6SAN_k^+(s,a)}{\delta}\right)}{N_k^+(s,a)}$$

$$\leq \frac{6\ln\left(\frac{6SAt_k}{\delta}\right)}{N_k^+(s,a)} < \frac{1}{D^{\mathtt{C}} S_k^{\mathtt{T}}}$$

where we have exploited the fact that $\widehat{p}(s'|s,a) = 0$ and $\widehat{\sigma}_{p,k}^2(s'|s,a) = 0$ for any state $s' \in \mathcal{S}_k^{\mathtt{T}}$ ($N_k(s,a,s') = 0$, see (3.13)), and the fact that $t_k \geq N_k^+(s,a) > 6D^{\mathtt{C}} S_k^{\mathtt{T}} \ln\left(\frac{6SAt_k}{\delta}\right)$.

As in Sec. 2.1.4, for all $s \in \mathcal{S}$, we denote by $h_{\mapsto s}^*$ the maximal non-positive fixed point of the Bellman shortest path operator $L_{\mapsto s}$ of the true MDP M where all rewards are set to -1 (see Thm. 2.8). As shown in Sec. 2.1.4, for all $(s,a,s') \in \mathcal{S} \times \mathcal{A}_s \times \mathcal{S}$, $-h_{\mapsto s'}^*(s) = \min_{\pi \in \Pi^{\mathrm{SR}}(M)} \mathbb{E}_M^\pi[\tau(s')|s_1 = s] - 1$ i.e., $h_{\mapsto s'}^*(s)$ is the expected length of the stochastic shortest

path going from s to s' in the true MDP M. Fix an arbitrary target state $\overline{s} \in \mathcal{S}_k^{\mathtt{T}}$ and define $h_{\max}(\overline{s}) := \max_{s \in \mathcal{S}_k^{\mathtt{C}}} h^*_{\mapsto \overline{s}}(s)$. By construction, $h^*_{\mapsto \overline{s}}(\overline{s}) = 0$ and for all $s \in \mathcal{S}_k^{\mathtt{C}}$

$$
\begin{aligned}
h^*_{\mapsto \overline{s}}(s) &= \max_{a \in \mathcal{A}_s} \left\{ -1 + \sum_{s' \in \mathcal{S}} \underbrace{p(s'|s,a) h^*_{\mapsto \overline{s}}(s')}_{\leq 0} \right\} \leq -1 + \max_{a \in \mathcal{A}_s} \left\{ \sum_{s' \in \mathcal{S}_k^{\mathtt{C}}} p(s'|s,a) \underbrace{h^*_{\mapsto \overline{s}}(s')}_{\leq h_{\max}(\overline{s})} \right\} \\
&\leq -1 + h_{\max}(\overline{s}) \cdot \min_{a \in \mathcal{A}} \left\{ \sum_{s' \in \mathcal{S}_k^{\mathtt{C}}} p(s'|s,a) \right\} = -1 + h_{\max}(\overline{s}) \cdot \min_{a \in \mathcal{A}} \left\{ 1 - \sum_{s' \in \mathcal{S}_k^{\mathtt{T}}} \underbrace{p(s'|s,a)}_{< \frac{1}{D^{\mathtt{C}} S_k^{\mathtt{T}}}} \right\} \\
&< -1 + h_{\max}(\overline{s}) \cdot \left(1 - \sum_{s' \in \mathcal{S}_k^{\mathtt{T}}} \frac{1}{D^{\mathtt{C}} S_k^{\mathtt{T}}} \right) = -1 + h_{\max}(\overline{s}) \cdot \left(1 - \frac{1}{D^{\mathtt{C}}} \right)
\end{aligned}
$$

Applying the above inequality to the state $s \in \mathcal{S}_k^{\mathtt{C}}$ achieving $h^*_{\mapsto \overline{s}}(s) = h_{\max}(\overline{s})$ we obtain $-h_{\max}(\overline{s}) > D^{\mathtt{C}}$. By definition, $-h_{\max}(\overline{s})$ is the minimum expected time it takes to go from $\mathcal{S}_k^{\mathtt{C}}$ to $\overline{s}$ in M. Therefore, the shortest path between any state $s \in \mathcal{S}_k^{\mathtt{C}} \subseteq \mathcal{S}^{\mathtt{C}}$ and any state in $\overline{s} \in \mathcal{S}_k^{\mathtt{T}}$ is strictly longer than $D^{\mathtt{C}}$ in expectation. But by definition $D^{\mathtt{C}}$ is the longest shortest path between any pair of states in $\mathcal{S}^{\mathtt{C}}$. Therefore, $\overline{s} \in \mathcal{S}^{\mathtt{T}}$. Since $\overline{s} \in \mathcal{S}_k^{\mathtt{T}}$ was chosen arbitrarily, then $\mathcal{S}_k^{\mathtt{T}} = \mathcal{S}^{\mathtt{T}}$. ■

Lem. 4.1 basically excludes the case where $\mathcal{S}^{\mathtt{T}} \subsetneq \mathcal{S}_k^{\mathtt{T}}$ (i.e., some states in $\mathcal{S}^{\mathtt{C}}$ have not been visited yet). Let's assume that event E holds i.e., $M \in \mathcal{M}_k$ for all $k \geq 1$. As pointed out in Sec. 4.2.2 (paragraph on "Communicating MDPs"), when $\mathcal{S}_k^{\mathtt{T}} = \emptyset$, $\mathcal{M}_k = \overline{\mathcal{M}}_k$ and so $M \in \overline{\mathcal{M}}_k$. Using the same argument as in Sec. 3.2, we have that $g_k^* \geq g^*$. We now analyze separately the two cases of Lem. 4.1.

Case 1. If $\mathcal{S}_k^{\mathtt{T}} = \mathcal{S}^{\mathtt{T}}$ then $M \in \overline{\mathcal{M}}_k$ (under event E) because TUCRL only forbids transitions that indeed do not exist in M itself. Formally, for any $(s,a,s') \in \mathcal{S}_k^{\mathtt{C}} \times \mathcal{A}_s \times \mathcal{S}_k^{\mathtt{T}}$ we have $p(s'|s,a) = p_k(s'|s,a) = 0$ and $M \in \mathcal{M}_k$ so $p(s'|s,a) \in \overline{Z}_p^k(s,a,s')$ for all $(s,a,s') \in \mathcal{S} \times \mathcal{A}_s \times \mathcal{S}$. In conclusion, $g_k^* \geq g^*$.

Case 2. If $\mathcal{E}_k = \emptyset$, $\mathcal{S}_k^{\mathrm{EVI}} = \mathcal{S}$ and every state in $\mathcal{S}$ is accessible from any other state in $\mathcal{S}$ (in the extended MDP $\overline{\mathcal{M}}_k$). Thus, g_k^* is the optimal gain of all the states in $\mathcal{S}$ and in particular the states in $\mathcal{S}_k^{\mathtt{T}}$ (Puterman, 1994, Theorem 8.3.2). For all $(s,a) \in \mathcal{S}_k^{\mathtt{T}} \times \mathcal{A}_s$, $\overline{Z}_p^k(s,a,s) = [0,1]$ and $B_r^k(s,a) = [0, r_{\max}]$ meaning that we can set $p_k(s|s,a) \leftarrow 1$ and $r_k(s,a) = r_{\max}$. Therefore the optimal gain in such states is clearly $r_{\max}$ and so $g_k^* = r_{\max}$.

In conclusion, under event E and for $t_k \geq C_k$, TUCRL is always *optimistic* i.e., $g_k^* \geq g^*$. Note that Lem. 4.1 is not true with $D^{\mathtt{C}}$ replaced by $\Lambda^{\mathtt{C}}/r_{\max}$ (take for example $\Lambda^{\mathtt{C}} = 0$ i.e., all rewards equal to $r_{\max}$).

Range of the optimistic bias

The second technical difficulty in the analysis of TUCRL is to bound the range of h_k i.e., $sp(h_k)$. While in communicating MDPs, it is possible to bound this quantity by the travel-budget of the MDP as $sp(h_k) \leq \Lambda$ (see Sec. 3.3.2), in weakly-communicating MDPs $\Lambda = +\infty$, thus making this bound *uninformative*. As a result, we need to *restrict* our attention to the subset of communicating states $\mathcal{S}^{\mathsf{c}}$, where the travel-budget Λ^{c} is finite. We will actually see in the regret proof that we only need to bound the range of h_k on the subset of states $\mathcal{S}_k^{\mathsf{c}}$ i.e., $sp_{\mathcal{S}_k^{\mathsf{c}}}(h_k) := \max_{s\in\mathcal{S}_k^{\mathsf{c}}}\{h_k(s)\} - \min_{s\in\mathcal{S}_k^{\mathsf{c}}}\{h_k(s)\}$. Since the true MDP M belongs to the extended MDP $\mathcal{M}_k$ w.h.p. but may not belong to $\overline{\mathcal{M}}_k$, we bound $sp_{\mathcal{S}_k^{\mathsf{c}}}(h_k)$ by first comparing the Bellman shortest path operators of $\mathcal{M}_k$ and $\overline{\mathcal{M}}_k$ (rather than directly comparing the operators of M and $\overline{\mathcal{M}}_k$ like in Sec. 3.3).

Define the extended MDP $\mathcal{M}'_k := \left\{\mathcal{S},\ \mathcal{A},\ r(s,a) \in B_r^{k\prime}(s,a),\ p(s'|s,a) \in B_p^k(s,a,s')\right\}$ where $B_r^{k\prime}(s,a) := \left\{r - r_{\max} \text{ s.t. } r \in B_r^k(s,a)\right\}$ and $B_r^k(s,a)$ (3.4) and $B_p^k(s,a,s')$ (3.3) are the confidence intervals used to construct $\mathcal{M}_k$ in UCRLB. We define $\overline{\mathcal{M}}'_k$ similarly where $B_p^k(s,a,s')$ is replaced by $\overline{Z}_p^k(s,a,s')$ (4.7). For any state in $\mathcal{S}_k^{\mathsf{c}}$, we denote by $\mathcal{L}_{\mapsto s}^k$ (resp. $\overline{\mathcal{L}}_{\mapsto s}^k$) the Bellman shortest path operator to s in $\mathcal{M}'_k$ (resp. $\overline{\mathcal{M}}'_k$) as defined in Thm. 3.5. We also denote by $h_{\mapsto s}^k$ (resp. $\overline{h}_{\mapsto s}^k$) the fixed point of $\mathcal{L}_{\mapsto s}^k$ (resp. $\overline{\mathcal{L}}_{\mapsto s}^k$). The fixed points exist and are unique because every state in $\mathcal{S}_k^{\mathsf{c}}$ is accessible from any state in $\mathcal{S}$ (see Thm. 3.5). Furthermore, we prove the following lemma:

Lemma 4.2

For all $s \in \mathcal{S}_k^{\mathsf{c}}$ we have $\overline{h}_{\mapsto s}^k \geq h_{\mapsto s}^k$ (component-wise).

Proof. $h_{\mapsto s}^k$ is a fixed point of $\mathcal{L}_{\mapsto s}^k$ and so for all $x \in \mathcal{S}\setminus\{s\}$,

$$h_{\mapsto s}^k(x) = \mathcal{L}_{\mapsto s}^k h_{\mapsto s}^k(x) = \max_{a\in\mathcal{A}_s}\left\{\max_{r\in B_r^k(x,a)}\{r\} - r_{\max} + \max_{p\in B_p^k(x,a)}\left\{\sum_{s'\neq s} p(s')h_{\mapsto s}^k(s')\right\}\right\}$$

where $B_p^k(x,a) = \left\{p \in \Delta_S :\ p(s') \in B_p^k(x,a,s'),\ \forall s' \in \mathcal{S}\right\}$ (see Eq. 3.17). Similarly, we define $\overline{Z}_p^k(x,a) := \left\{p \in \Delta_S :\ p(s') \in \overline{Z}_p^k(x,a,s'),\ \forall s' \in \mathcal{S}\right\}$. Our goal is to show that for all $x \in \mathcal{S}$, $\overline{\mathcal{L}}_{\mapsto s}^k h_{\mapsto s}^k(x) \geq h_{\mapsto s}^k(x)$ where by definition

$$\overline{\mathcal{L}}_{\mapsto s}^k h_{\mapsto s}^k(x) = \max_{a\in\mathcal{A}_s}\left\{\max_{r\in B_r^k(x,a)}\{r\} - r_{\max} + \max_{p\in \overline{Z}_p^k(x,a)}\left\{\sum_{s'\neq s} p(s')h_{\mapsto s}^k(s')\right\}\right\}$$

Denote by $p'_k(\cdot|x,a) \in B_p^k(x,a)$ the probability distribution achieving the maximum in the fixed point equation of $h_{\mapsto s}^k$ i.e.,

$$\sum_{s'\neq s} p'_k(s'|x,a)h_{\mapsto s}^k(s') := \max_{p\in B_p^k(x,a)}\left\{\sum_{s'\neq s} p(s')h_{\mapsto s}^k(s')\right\} \tag{4.15}$$

and denote by $\overline{p}'_k(\cdot|x,a) \in \overline{Z}^k_p(x,a)$ the analogue of $p'_k(\cdot|x,a)$ for $\overline{Z}^k_p(x,a)$ i.e.,

$$\sum_{s' \neq s} \overline{p}'_k(s'|x,a) h^k_{\mapsto s}(s') := \max_{p \in \overline{Z}^k_p(x,a)} \left\{ \sum_{s' \neq s} p(s') h^k_{\mapsto s}(s') \right\} \tag{4.16}$$

If $x \in \mathcal{S}^{\mathtt{T}}_k$ or $p^+_k(s,a) \geq \rho_{t_k}(s,a)$ then $\overline{Z}^k_p(x,a) = B^k_p(x,a)$ by definition (4.7), and so (4.15) and (4.16) are equal. On the other hand, if $x \in \mathcal{S}^{\mathtt{C}}_k$ and $p^+_k(x,a) < \rho_{t_k}(x,a)$ (see Eq. 4.7), then we might not have equality. Define $\widetilde{p}'_k(\cdot|x,a)$ as

$$\widetilde{p}'_k(s'|x,a) := \begin{cases} p'_k(s'|x,a) \text{ if } s' \in \mathcal{S}^{\mathtt{C}}_k \setminus \{s\} \\ 0 \text{ if } s' \in \mathcal{S}^{\mathtt{T}}_k \\ p'_k(s'|x,a) + \sum_{y \in \mathcal{S}^{\mathtt{T}}_k} p'_k(y|x,a) \text{ if } s' = s \end{cases}$$

$p'_k(y|x,a) \in B^k_p(x,a,y)$ and it is clear from the definition of $B^k_p(x,a,y)$ for $y \in \mathcal{S}^{\mathtt{T}}_k$ that

$$\sum_{y \in \mathcal{S}^{\mathtt{T}}_k} p'_k(y|x,a) \leq \sum_{y \in \mathcal{S}^{\mathtt{T}}_k} p^+_k(y|x,a) = S^{\mathtt{T}}_k \cdot p^+_k(s,a) = \zeta^{sa}_{p,k}$$

and so $\widetilde{p}'_k(s'|x,a) \in \overline{Z}^k_p(x,a)$ when $x \in \mathcal{S}^{\mathtt{C}}_k$ and $p^+_k(x,a) < \rho_{t_k}(x,a)$. Moreover, by construction (see Sec. 3.3) $h^k_{\mapsto s}(s') \leq 0$ for all $s' \in \mathcal{S}$. In conclusion we can write:

$$\sum_{s' \neq s} p'_k(s'|x,a) h^k_{\mapsto s}(s') \leq \sum_{s' \neq s} \widetilde{p}'_k(s'|x,a) h^k_{\mapsto s}(s') \leq \sum_{s' \neq s} \overline{p}'_k(s'|x,a) h^k_{\mapsto s}(s')$$

and as a consequence, $\overline{\mathcal{L}}^k_{\mapsto s} h^k_{\mapsto s}(x) \geq \mathcal{L}^k_{\mapsto s} h^k_{\mapsto s}(x)$. This proves that $\overline{\mathcal{L}}^k_{\mapsto s} h^k_{\mapsto s} \geq \mathcal{L}^k_{\mapsto s} h^k_{\mapsto s} = h^k_{\mapsto s}$ and Thm. 3.5 implies that $\overline{h}^k_{\mapsto s} \geq h^k_{\mapsto s}$. ∎

From Sec. 3.3.2 we know that $sp_{\mathcal{S}^{\mathtt{C}}_k}(h_k) \leq \max_{s,x \in \mathcal{S}^{\mathtt{C}}_k} \left| \overline{h}^k_{\mapsto s}(x) \right|$ (see Eq. 3.35 of Thm. 3.3). Applying Lem. 4.2 and since $h^k_{\mapsto s} \leq 0$ we have that $\left| \overline{h}^k_{\mapsto s}(x) \right| \leq \left| h^k_{\mapsto s}(x) \right|$ and so $sp_{\mathcal{S}^{\mathtt{C}}_k}(h_k) \leq \max_{s,x \in \mathcal{S}^{\mathtt{C}}_k} \left| h^k_{\mapsto s}(x) \right|$. Finally, we already showed in Sec. 3.3.2 that $\left| h^k_{\mapsto s}(x) \right| \leq |h^*_{\mapsto s}(x)|$ (where $h^*_{\mapsto s}$ is the fixed point of the Bellman shortest path operator in the true MDP M). In conclusion, since $\mathcal{S}^{\mathtt{C}}_k \subseteq \mathcal{S}^{\mathtt{C}}$, $sp_{\mathcal{S}^{\mathtt{C}}_k}(h_k) \leq \max_{s,x \in \mathcal{S}^{\mathtt{C}}_k} |h^*_{\mapsto s}(x)| \leq \max_{s,x \in \mathcal{S}^{\mathtt{C}}} |h^*_{\mapsto s}(x)| := \Lambda^{\mathtt{C}}$.

4.3.2 Regret guarantees

We prove that the regret of TUCRL is bounded as follows.

Theorem 4.1 (Analogue to Thm. 3.4)

There exists a numerical constant $\beta > 0$ such that for any *weakly-communicating MDP (resp. multi-chain MDP), with probability at least $1-\delta$, it holds that for all initial state distribution $\mu_1 \in \Delta_S$ satisfying Asm. 4.1 (resp. Asm. 4.2) and for all time horizons $T > 1$*

$$\begin{aligned}\Delta(\mathrm{TUCRL}, T) = \; & \beta \cdot \max\{r_{\max}, \Lambda^{\mathsf{c}}\} \sqrt{\sum_{s \in \mathcal{S}^{\mathsf{c}}, a \in \mathcal{A}_s} \Gamma(s,a) T \ln\left(\frac{T}{\delta}\right)} \\ & + \beta \cdot r_{\max} \left(D^{\mathsf{c}}\right)^2 S^3 A \ln^2\left(\frac{T}{\delta}\right)\end{aligned} \tag{4.17}$$

The first term in the regret shows the ability of TUCRL to adapt to the communicating part of the true MDP M by scaling with the *communicating* travel-budget Λ^{c} and MDP parameters S^{c} and Γ^{c} (more precisely the sum of all $\Gamma(s,a)$ with $s \in \mathcal{S}^{\mathsf{c}}$). The second term mainly corresponds to the regret incurred in the *early stage* where the regret grows *linearly*. When M is communicating, we match the square-root term of UCRLB (first term) since $\Lambda^{\mathsf{c}} = \Lambda$ and $\mathcal{S}^{\mathsf{c}} = \mathcal{S}$ while in the worst-case where $\Lambda = r_{\max} D$, the second term is bigger than the one appearing in UCRLB by a multiplicative factor DS (ignoring logarithmic terms). It is not clear whether D^{c} can be replaced by Λ^{c} in general.

Unfortunately, we were not able to adapt the proof techniques of Thm. 3.5 to show a $\widetilde{\mathcal{O}}\left(\sqrt{r_{\max} \Lambda S^{\mathsf{c}} \Gamma^{\mathsf{c}} A T}\right)$ regret bound in general. Perhaps surprisingly, the problem is not coming from variance reduction methods or any new tool that we introduced in Sec. 3.6. All the steps of Sec. 3.6 are still valid but the dependency in Λ^{c} cannot be trivially improved. The linear (instead of square-root) dependency in Λ^{c} arises because the *telescopic sum* appearing in the decomposition of the term Δ_k^{p2} no longer telescops in our analysis, and can only be bounded by a $\widetilde{\mathcal{O}}\left(\Lambda^{\mathsf{c}} \sqrt{S^{\mathsf{c}} A T}\right)$ term. At first sight, this may seem to be an artefact of the proof, but it could also be an intrinsic limitation of the algorithm, or even an intrinsic limitation of the setting (i.e., infinite diameter). In order to avoid spending too much time attempting to visit unreachable states, TUCRL periodically ignores some transitions that have never been observed but may lead to highly rewarding state. Yet, TUCRL eventually takes these transitions into account again if they have not been visited enough (less than $\sqrt{T/SA}$ times) so as to prevent under-exploration. By doing so, the algorithm may move back and forth multiple times in the environment (even when only nonexistent transitions have not been observed), each time suffering a regret of order $sp(h^*)$ (in the worst case). The frequency at which useless transitions are considered is of order $\sqrt{T/SA}$. This may be the cause for the unavoidable linear dependency in Λ^{c}. We leave this problem as an open question. Note that if all states have been visited, then TUCRL eventually becomes completely equivalent to UCRLB and so the regret scales with $\sqrt{\Lambda^{\mathsf{c}}}$ instead.

4.3.3 Regret proofs

We now provide a sketch of the proof of Thm. 4.1. In order to preserve readability, all following inequalities should be interpreted up to minor approximations and in high probability.

We follow the same steps as the first regret proof of UCRLB (Sec. 3.5).

Isolating poorly visited state-action pairs. For any state-action pair (s,a), we denote by $\mathbb{1}_{\mathcal{E}_k}\{s,a\} := \mathbb{1}\{(s,a) \in \mathcal{E}_k\}$ the indicator function equal to 1 if and only if $(s,a) \in \mathcal{E}_k$ and 0 otherwise (see Eq. 4.13 for the definition of $\mathcal{E}_k$). We also denote by $\mathbb{1}_{\overline{\mathcal{E}_k}}\{s,a\} := \mathbb{1}\{(s,a) \notin \mathcal{E}_k\}$ the complement of $\mathbb{1}_{\mathcal{E}_k}\{s,a\}$ i.e., $\mathbb{1}_{\mathcal{E}_k}\{s,a\} + \mathbb{1}_{\overline{\mathcal{E}_k}}\{s,a\} = 1$. We use this equality to decompose the regret as:

$$\Delta(\text{TUCRL}, T) = \sum_{t=1}^{T} (g^* - r_t(s_t,a_t)) \mathbb{1}_{\mathcal{E}_{k_t}}\{s_t,a_t\} + \sum_{t=1}^{T} (g^* - r_t(s_t,a_t)) \mathbb{1}_{\overline{\mathcal{E}_{k_t}}}\{s_t,a_t\} \tag{4.18}$$

The first term isolates state-action pairs that have been visited a small number of times i.e., such that $N_k^+(s,a) \leq \sqrt{\frac{t_k}{SA}}$. Whenever such a sate-action pair is visited, the corresponding visit count $N_k^+(s,a)$ will be incremented by 1 at the end of episode k. But if $N_k^+(s,a)$ is incremented too much, we will eventually have $N_k^+(s,a) > \sqrt{\frac{t_k}{SA}}$ and so intuitively $\mathbb{1}_{\mathcal{E}_{k_t}}\{s_t,a_t\}$ cannot be equal to 1 too often. Lem. 4.3 indeed shows that the number of times $\mathbb{1}_{\mathcal{E}_{k_t}}\{s_t,a_t\} = 1$ occurs is cumulatively "small".

Lemma 4.3

For any $T \geq 1$ and any sequence of states and actions $\{s_1, a_1, \ldots\ldots s_T, a_T\}$ we have:

$$\sum_{t=1}^{T} \mathbb{1}_{\mathcal{E}_{k_t}}\{s_t,a_t\} \leq 2\sqrt{S^c AT}. \tag{4.19}$$

Proof. We first notice that by definition $t_{k_t} \leq t$ where $k_t := \sup\{k \geq 1 : t_k \leq t\}$ is the current episode at time t. As a result,

$$\mathbb{1}_{\mathcal{E}_{k_t}}\{s_t,a_t\} := \mathbb{1}\left\{N_{k_t}^+(s_t,a_t) \leq \sqrt{t_{k_t}/SA}\right\} \leq \mathbb{1}\left\{N_{k_t}^+(s_t,a_t) \leq \sqrt{t/SA}\right\}.$$

Instead of directly bounding $\sum_{t=1}^{T} \mathbb{1}_{\mathcal{E}_{k_t}}\{s_t,a_t\}$ we will bound the number of visits Z_T in state-action pairs that have been visited less than $\sqrt{t/SA}$ times

$$Z_T := \sum_{t=1}^{T} \mathbb{1}\left\{N_{k_t}^+(s_t,a_t) \leq \sqrt{t/SA}\right\}.$$

We recall that the quantity $N_k(s,a)$ is updated only after the end of episode k and the stopping condition of episodes used by TUCRL implies that

$$\forall k \geq 1,\ \forall (s,a) \in \mathcal{S} \times \mathcal{A},\quad \nu_k(s,a) \leq N_k^+(s,a). \tag{4.20}$$

Moreover, for all $(s,a) \notin \mathcal{S}^{\mathsf{c}} \times \mathcal{A}$, $\nu_k(s,a) = 0$ implying that only the states $s \in S^{\mathsf{c}}$ should be taken care of. We first decompose Z_T as:

$$Z_T := \sum_{s,a} \sum_{t=1}^{T} \mathbb{1}\Big\{N_{k_t}^+(s,a) \le \sqrt{t/SA}\Big\} \cdot \mathbb{1}\Big\{(s_t,a_t) = (s,a)\Big\} = \sum_{s \in S^{\mathsf{c}}} \sum_{a} Z_T(s,a)$$
$$\text{where} \quad Z_T(s,a) := \sum_{t=1}^{T} \mathbb{1}\Big\{N_{k_t}^+(s,a) \le \sqrt{t/SA}\Big\} \cdot \mathbb{1}\Big\{(s_t,a_t) = (s,a)\Big\}.$$

Using the fact that for all $\tau \ge 1$, $t_{k_\tau} \le \tau \le t_{k_\tau+1} - 1$ and eq. 3.13 and 3.14 we have:

$$\begin{aligned}\forall T \ge \tau \ge 1, \quad Z_\tau(s,a) &= \sum_{t=1}^{\tau} \underbrace{\mathbb{1}\Big\{N_{k_t}^+(s,a) \le \sqrt{t/SA}\Big\}}_{\le 1} \cdot \underbrace{\mathbb{1}\{(s_t,a_t) = (s,a)\}}_{\ge 0} \\ &\le \sum_{t=1}^{\tau} \mathbb{1}\{(s_t,a_t) = (s,a)\} \le \sum_{t=1}^{t_{k_\tau+1}-1} \mathbb{1}\{(s_t,a_t) = (s,a)\} \\ &= N_{k_\tau+1}(s,a) \end{aligned} \tag{4.21}$$

Let's define $t_{s,a}$ as the last time that $Z_t(s,a)$ was incremented by 1:

$$\begin{aligned} t_{s,a} &:= \max\Big\{T \ge t \ge 1 : \ N_{k_t}^+(s,a) \le \sqrt{t/SA} \ \text{ and } \ (s_t,a_t) = (s,a)\Big\} \\ &= \min\Big\{T \ge t \ge 1 : \ Z_t(s,a) = Z_T(s,a)\Big\}. \end{aligned}$$

We denote by $m_{s,a} := k_{t_{s,a}}$ the corresponding episode. By definition and using (4.21),

$$Z_T(s,a) = Z_{t_{s,a}}(s,a) \le N_{m_{s,a}+1}(s,a) \ \text{ and } N_{m_{s,a}}^+(s,a) \le \sqrt{t_{s,a}/SA}. \tag{4.22}$$

Moreover, by definition of $N_k(s,a)$ (see eq. 3.13 and 3.14) and (4.20):

$$N_{m_{s,a}+1}(s,a) = \underbrace{N_{m_{s,a}}(s,a)}_{\le N_{m_{s,a}}^+(s,a)} + \underbrace{\nu_{m_{s,a}}(s,a)}_{\le N_{m_{s,a}}^+(s,a)} \le 2N_{m_{s,a}}^+(s,a). \tag{4.23}$$

Gathering (4.22), and (4.23) we obtain:

$$\begin{aligned} Z_T(s,a) = Z_{t_{s,a}}(s,a) \le N_{m_{s,a}+1}(s,a) \le 2N_{m_{s,a}}^+(s,a) \le 2\sqrt{\frac{t_{s,a}}{SA}} \le 2\sqrt{\frac{T}{SA}} \\ \implies Z_T = \sum_{s \in S^{\mathsf{c}}} \sum_{a} Z_T(s,a) \le 2\sqrt{S^{\mathsf{c}} AT} \end{aligned}$$

where for the last inequality we used the fact that $S^{\mathsf{c}} \le S$ (by definition) implying $S^{\mathsf{c}}/\sqrt{S} = \sqrt{S^{\mathsf{c}}/S} \cdot \sqrt{S^{\mathsf{c}}} \le \sqrt{S^{\mathsf{c}}}$. ■

When $\mathbb{1}_{\mathcal{E}_{k_t}}\{s_t,a_t\} = 1$, TUCRL suffers at most the maximum per-step regret $r_{\max} \ge$

$g^* - r(s,a)$ and so combined with Lem. 4.3:

$$\sum_{t=1}^{T} \underbrace{(g^* - r_t(s_t,a_t))}_{\leq r_{\max}} \underbrace{\mathbb{1}_{\mathcal{E}_{k_t}}\{s_t,a_t\}}_{\geq 0} \leq r_{\max} \sum_{t=1}^{T} \mathbb{1}_{\mathcal{E}_{k_t}}\{s_t,a_t\} \leq 2r_{\max}\sqrt{S^{\mathtt{C}}AT} \tag{4.24}$$

We now have to deal with the second term appearing in the inequality (4.18) i.e., state-action pairs that have been frequently visited. The whole purpose of restricting attention to those pairs will be clear later after we expand this term. We slightly change the definition of the per-episode regret Δ_k compared to Sec. 3.5 to account for $\mathbb{1}_{\overline{\mathcal{E}_k}}\{s,a\}$:

$$\begin{aligned}\Delta_k &:= \sum_{s\in\mathcal{S},a\in\mathcal{A}} \nu_k(s,a)\,(g^* - r(s,a)) \cdot \mathbb{1}_{\overline{\mathcal{E}_k}}\{s,a\} \\ &= \sum_{s\in\mathcal{S}_k^{\mathtt{C}},a\in\mathcal{A}} \nu_k(s,a)\,(g^* - r(s,a)) \cdot \mathbb{1}_{\overline{\mathcal{E}_k}}\{s,a\}\,.\end{aligned}$$

The reason why the sum over all states can be *restricted* to a sum over states in $\mathcal{S}_k^{\mathtt{C}}$ is because $\nu_k(s) = 0$ for all $s \in \mathcal{S}_k^{\mathtt{T}}$ by definition of the stopping condition of episode k (4.11). Furthermore, inequality (3.1) in Lem. 3.38 (Sec. 3.5) is based on an MDS argument and remains valid even with the additional multiplicative factor $\mathbb{1}_{\overline{\mathcal{E}_{k_t}}}\{s_t,a_t\}$ and if we keep $\nu_k(s,a)$ instead of taking the conditional expectation $\nu_k(s)\pi_k(a|s)$. In the end,

$$\sum_{t=1}^{T}(g^* - r_t(s_t,a_t))\mathbb{1}_{\overline{\mathcal{E}_{k_t}}}\{s_t,a_t\} \leq \sum_{k=1}^{k_T}\Delta_k + 2r_{\max}\sqrt{T\ln\left(\frac{4T}{\delta}\right)}.$$

Isolating non-optimistic episodes. In order to be able to use the *optimism* property proved in Sec. 4.3.1, we need to separate the episodes where $t_k < C_k$ (C_k is defined in Eq. 4.14 of Lem. 4.1) from the other episodes i.e., we decompose the sum of Δ_k as

$$\sum_{k=1}^{k_T}\Delta_k \leq \sum_{k=1}^{k_T}\Delta_k \cdot \mathbb{1}\{t_k < C_k\} + \sum_{k=1}^{k_T}\Delta_k \cdot \mathbb{1}\{t_k \geq C_k\}.$$

The episodes where $t_k < C_k$ define a full exploratory phase, where the agent may suffer *linear regret*. However, this phase is somehow "short". Define $\overline{k_T} := \max\{k_T \geq k \geq 1 :\ t_k < C_k\}$ to be the last episode $k_T \geq k \geq 1$ satisfying $t_k < C_k$. Because of the stopping condition of episodes (4.11), $\nu_k(s,a) \leq 2N_k^+(s,a)$ for all (s,a) and so $t_{k+1} \leq 2t_k$ implying that

$$\begin{aligned}\sum_{k=1}^{k_T}\Delta_k \cdot \mathbb{1}\{t_k < C_k\} = \sum_{k=1}^{\overline{k_T}}\Delta_k &\leq r_{\max}\sum_{k=1}^{\overline{k_T}}(t_{k+1} - t_k) = r_{\max}t_{\overline{k_T}+1} \leq 2r_{\max}t_{\overline{k_T}} < 2r_{\max}C_{\overline{k_T}} \\ &\leq 72r_{\max}\left(D^{\mathtt{C}}\right)^2 S^3 A \ln\left(\frac{6SAT}{\delta}\right)^2 \end{aligned} \tag{4.25}$$

where the last inequality follows from the definition of C_k.

Per-episode regret. It now remains to bound the dominant term $\sum_{k=1}^{k_T}\Delta_k \cdot \mathbb{1}\{t_k \geq C_k\} = \sum_{k=\overline{k_T}+1}^{k_T}\Delta_k$. We do this by first analyzing individually the regret Δ_k of each episode k (as

we did for UCRLB). We proceed as in Sec. 3.5.2: we bound g^* by $g_k + \varepsilon_k/2$ and plug-in the (approximate) optimality equation of the extended MDP $\overline{\mathcal{M}}_k$ involving g_k, h_k, r_k and p_k (we recall that all the actions played with non-zero probability satisfy an optimality equation). The same terms Δ_k^p and Δ_k^r appear except that the sum is over $s \in \mathcal{S}_k^{\mathsf{C}}$ and a multiplicative factor $\mathbb{1}_{\overline{\mathcal{E}_k}}\{s,a\}$ appears e.g.,

$$\Delta_k^p := \alpha \sum_{\substack{s\in\mathcal{S}_k^{\mathsf{C}} \\ a\in\mathcal{A}_s}} \nu_k(s,a) \left(\sum_{s'\in\mathcal{S}} p_k(s'|s,a) h_k(s') - h_k(s) \right) \mathbb{1}_{\overline{\mathcal{E}_k}}\{s,a\}$$

and similarly for Δ_k^r. We further notice that for all $(s,a) \notin \mathcal{E}_k$ (i.e., satisfying $\mathbb{1}_{\overline{\mathcal{E}_k}}\{s,a\} \neq 0$) and $s' \in \mathcal{S}_k^{\mathsf{T}}$, we have $p_k(s'|s,a) = 0$ by *construction* of $\overline{\mathcal{M}}_k$ (see Eq. 4.7). The whole point of having $\mathbb{1}_{\overline{\mathcal{E}_k}}\{s,a\}$ in factor is that the sum over $s' \in \mathcal{S}$ can be *restricted* to a sum over $s' \in \mathcal{S}_k^{\mathsf{C}}$ i.e.,

$$\Delta_k^p = \alpha \sum_{\substack{s\in\mathcal{S}_k^{\mathsf{C}} \\ a\in\mathcal{A}_s}} \nu_k(s,a) \left(\sum_{s'\in\mathcal{S}_k^{\mathsf{C}}} p_k(s'|s,a) h_k(s') - h_k(s) \right) \mathbb{1}_{\overline{\mathcal{E}_k}}\{s,a\}\,.$$

In the case of UCRLB, the travel-budget of the *whole* MDP Λ appears because the range of h_k can only be bounded by Λ. But since in the case of TUCRL s' lies in $\mathcal{S}_k^{\mathsf{C}}$, only the range of h_k *on this subset* matters. We already proved in Sec. 4.1 that (under event E) $sp_{\mathcal{S}_k^{\mathsf{C}}}(h_k) \leq \Lambda^{\mathsf{C}}$. This explains why Λ^{C} appears instead of Λ when bounding the regret of TUCRL.

We now proceed as in eq. 3.43 i.e., we add and subtract the term

$$\alpha \sum_{\substack{s\in\mathcal{S}_k^{\mathsf{C}} \\ a\in\mathcal{A}_s}} \nu_k(s,a) \sum_{s'\in\mathcal{S}_k^{\mathsf{C}}} p(s'|s,a) h_k(s') \mathbb{1}_{\overline{\mathcal{E}_k}}\{s,a\}$$

in order to obtain two terms Δ_k^{p1} and Δ_k^{p2}. Note that s and s' are summed over $\mathcal{S}_k^{\mathsf{C}}$ (as we just explained) and there is an additional indicator function $\mathbb{1}_{\overline{\mathcal{E}_k}}\{s,a\}$ compared to Sec. 3.5.3. The indicator function does not impact the bound of Δ_k^{p1} (the same analysis as for UCRLB can be carried out, where we enventually bound $\mathbb{1}_{\overline{\mathcal{E}_k}}\{s,a\} \leq 1$ once the difference $p_k - p$ has been bounded by a positive term). However, the term Δ_k^{p2} is more problematic. We decompose this term as follows:

$$\Delta_k^{p2} = \underbrace{\alpha \sum_{t=t_k}^{t_{k+1}-1} \left(\sum_{s'\in\mathcal{S}_k^{\mathsf{C}}} p(s'|s_t,a_t) w_k(s') - w_k(s_{t+1}) \cdot \mathbb{1}\left\{ s_{t+1} \in \mathcal{S}_k^{\mathsf{C}} \right\} \right) \cdot \mathbb{1}_{\overline{\mathcal{E}_k}}\{s_t,a_t\}}_{:=\Delta_k^{p4}}$$
$$+ \underbrace{\alpha \sum_{t=t_k}^{t_{k+1}-1} \left(w_k(s_{t+1}) \cdot \mathbb{1}\left\{ s_{t+1} \in \mathcal{S}_k^{\mathsf{C}} \right\} - w_k(s_t) \right) \cdot \mathbb{1}_{\overline{\mathcal{E}_k}}\{s_t,a_t\}\,.}_{\textit{not telescopic!}}$$

Despite the indicator functions $\mathbb{1}_{\overline{\mathcal{E}_k}}\{s_t, a_t\}$ and $\mathbb{1}\{s_{t+1} \in \mathcal{S}_k^{\mathsf{c}}\}$, the term Δ_k^{p4} is still an MDS because k_t (the episode at time t) is $\mathcal{F}_{t-1}$-measurable where $\mathcal{F}_{t-1} := \sigma(s_1, a_1, r_1, \dots, s_t)$ (see App. A.2) and moreover

$$\mathbb{E}\left[w_{k_t}(s_{t+1})\mathbb{1}_{\overline{\mathcal{E}_k}}\{s_t, a_t\}\,\mathbb{1}\left\{s_{t+1} \in \mathcal{S}_{k_t}^{\mathsf{c}}\right\}\middle|\mathcal{F}_{t-1}\right] = \underbrace{\sum_{s' \in \mathcal{S}_{k_t}^{\mathsf{c}}} p(s'|s_t, a_t) w_{k_t}(s') \mathbb{1}_{\overline{\mathcal{E}_k}}\{s_t, a_t\}}_{\mathcal{F}_{t-1}-\text{measurable}}.$$

As a result, Lem. 3.3 still applies. However, the second term is no longer a *telescopic sum* although the problem is not coming from the indicator function $\mathbb{1}\{s_{t+1} \in \mathcal{S}_k^{\mathsf{c}}\}$. Indeed, due to the new stopping condition implemented by TUCRL, for all episodes $k \geq 1$ and time steps $t_k \leq t < t_{k+1} - 1$, $s_t \notin \mathcal{S}_k^{\mathsf{c}}$ and so $w_k(s_t) = w_k(s_t) \cdot \mathbb{1}\{s_t \in \mathcal{S}_k^{\mathsf{c}}\}$. On the other hand, the presence of the second indicator function $\mathbb{1}_{\overline{\mathcal{E}_k}}\{s_t, a_t\}$ is an issue. Using the fact that $\mathbb{1}_{\overline{\mathcal{E}_k}}\{s_t, a_t\} = 1 - \mathbb{1}_{\mathcal{E}_k}\{s_t, a_t\}$ we can make a telescopic sum appear:

$$\underbrace{\sum_{t=t_k}^{t_{k+1}-1} w_k(s_{t+1}) \cdot \mathbb{1}\left\{s_{t+1} \in \mathcal{S}_k^{\mathsf{c}}\right\} - w_k(s_t) \cdot \mathbb{1}\left\{s_t \in \mathcal{S}_k^{\mathsf{c}}\right\}}_{\leq \Lambda^{\mathsf{c}} \text{ (telescopic sum)}}$$
$$+ \sum_{t=t_k}^{t_{k+1}-1} \underbrace{\left(w_k(s_{t+1}) \cdot \mathbb{1}\left\{s_{t+1} \in \mathcal{S}_k^{\mathsf{c}}\right\} - w_k(s_t)\mathbb{1}\left\{s_t \in \mathcal{S}_k^{\mathsf{c}}\right\}\right)}_{\leq \Lambda^{\mathsf{c}}} \cdot \mathbb{1}_{\overline{\mathcal{E}_k}}\{s_t, a_t\}$$

Using Lem. 4.3, this term can be bounded by $\Lambda^{\mathsf{c}} + 2\Lambda^{\mathsf{c}}\sqrt{\mathcal{S}^{\mathsf{c}} AT}$. The presence of this term is the reason why we were *not able* to obtain a regret bound scaling linearly with $\sqrt{\Lambda^{\mathsf{c}}}$ instead of Λ^{c}. All our attempts to either refine the current analysis, or modify the algorithm to improve the dependency in Λ^{c} have failed so far.

The fact that the sum $\sum_{k=\overline{k_T}+1}^{k_T} \Delta_k$ starts from $k = \overline{k_T} + 1$ (instead of $k = 1$) has no impact on the final bound and the increase in the number of episodes due to the modification of the stopping condition of UCRLB is negligible.

The final regret bounds in Thm. 4.1 is then obtained by combining all different terms (4.24), (4.25) and the bound on the sum $\sum_{k=\overline{k_T}+1}^{k_T} \Delta_k$.

4.4 Experiments

In this section, we present experiments to validate the theoretical findings of Sec. 4.3 (Thm. 4.1). We compare TUCRL against UCRLB. To the best of out knowledge, there exists no implementable algorithm to solve the optimization step of REGAL and REGAL.D and so we do not report any experiments with these algorithms. We are not aware of any other algorithm that addresses the problem of infinite diameter to compare with.

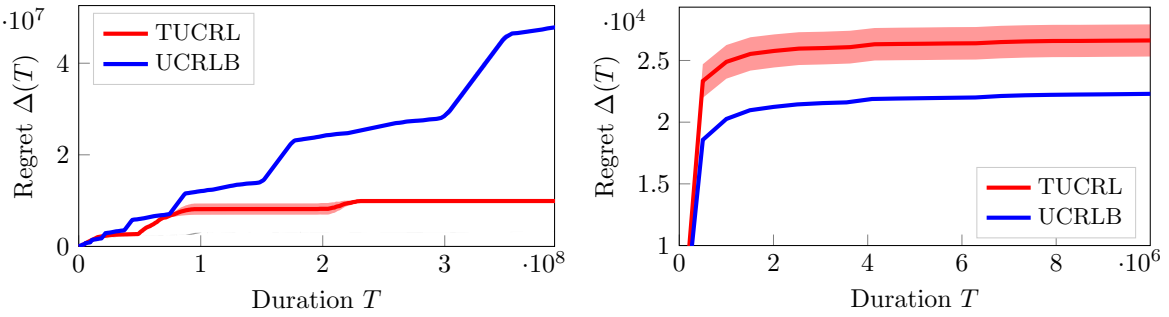


Figure 4.4: Cumulative regret in the taxi with misspecified states (Fig. 4.4a) and in the communicating taxi (Fig. 4.4b). Confidence intervals $\beta_{r,k}$ and $\beta_{p,k}$ are respectively shrunk by a factor 0.05 and 0.01. Results are averaged over 20 runs and 95% confidence intervals are reported.

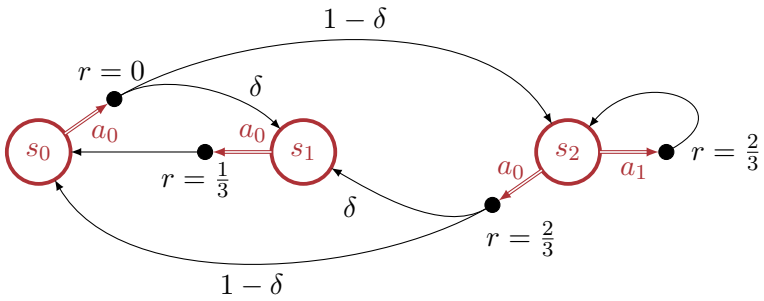


Figure 4.5: Family of three-state MDPs characterized by a single parameter δ. When $\delta > 0$, the MDP is communicating, when $\delta = 0$ it is weakly-communicating. Only two stationary deterministic policies can be played (corresponding to the two actions available in s_2).

Taxi Problem. We first consider the taxi problem (**?**) implemented in OpenAI Gym (Brockman et al., 2016). Even such a simple domain contains *misspecified states*. The state space is constructed as the outer product of the taxi position, the passenger position and the destination and this leads to states that cannot be reached from any possible starting configuration (all the starting states belong to $\mathcal{S}^{\mathsf{c}}$). More precisely, out of 500 states in $\mathcal{S}$, 100 are non-reachable. On Fig. 4.4 we compare the regret of UCRLB and TUCRL when the misspecified states are present (Fig. 4.4a) and when they are removed from the definition of the state space (Fig. 4.4b). In the presence of misspecified states (Fig. 4.4a), the regret of UCRLB clearly grows linearly with T (as expected, see Sec. 4.1) while TUCRL is able to *learn* as expected. On the other hand, when the MDP is communicating (Fig. 4.4b) TUCRL performs similarly to UCRLB. The small loss in performance is most likely due to the initial exploration phase during which the confidence intervals on the transition probabilities used by UCRLB (extended MDP $\mathcal{M}_k$) are tighter than those used by TUCRL (extended MDP $\overline{\mathcal{M}}_k$). Indeed, TUCRL slightly increases some confidence bounds by $\zeta^{sa}_{p,k}$ (4.6) compared to UCRLB (see Eq. 4.7).

Simple three-state domain. In order to better understand the empirical behaviour of the algorithm, We further study the regret of TUCRL in the simpler three-state domain of Fig. 4.5. The environment is composed of only three states (s_0, s_1 and s_2) and one action

per state, except in s_2 where two actions are available. As a result, the agent only has the choice between two possible policies. We first consider the case the MDP is communicating by defining $\delta = 0.005 > 0$. Fig. 4.6a shows that, as expected, TUCRL behaves similarly to UCRLB. In this example it is able to outperform UCRLB since the preliminary phase in which transitions to non-observed states are forbidden leads to a less explorative behaviour that, due to the structure of the problem (s_1 is difficult to reach but it is also non-optimal), results in a smaller regret.

Fig. 4.6b shows the cumulative regret achieved by TUCRL when the diameter is *infinite* i.e., $\mathcal{S}^{\mathtt{C}} = \{s_0, s_2\}$ and $\mathcal{S}^{\mathtt{T}} = \{s_1\}$. Similarly to the taxi problem, UCRLB fails to learn in this setting (i.e., suffers linear regret) and for the sake of clarity, we do not report its regret on the figure. TUCRL quickly achieves sub-linear regret as predicted by theory. However, TUCRL seem to achieve different regret growth rates depending on whether s_1 is removed or not. While the regret curve of Fig. 4.6b quickly achieves an asymptotic regime (slow logartihmic increase), the regret curve of Fig. 4.6b)seems to keep growing as $\sqrt{T}$ (no matter for how long we run the experiment), with *periodic* "jumps" that are increasingly distant (in time) from each other. The time between two consecutive "jumps" grows exponentially fast and the increase in regret at every "jump" also grows exponentially fast. This can be explained by the way the algorithm works: while most of the time TUCRL is optimistic on the restricted state space $\mathcal{S}^{\mathtt{C}} = \{s_0, s_2\}$ (i.e., $\mathcal{S}_k^{\mathtt{C}} = \mathcal{S}^{\mathtt{C}}$), it *periodically* allows transitions to the set $\mathcal{S}^{\mathtt{T}} = \{s_1\}$ (i.e., $\mathcal{S}_k^{\mathtt{C}} = \mathcal{S}$), which is indeed not reachable. Enabling these transitions triggers *"aggressive" exploration* during an entire episode. The policy played is then sub-optimal creating a *"jump"* in the regret. At the end of this *exploratory episode*, $\mathcal{S}_k^{\mathtt{C}}$ will be set again to $\mathcal{S}^{\mathtt{C}}$ and the regret will stop increasing until the condition $N_k^+ \leq \sqrt{t_k/SA}$ occurs again. The cumulative regret incurred during exploratory episodes (when transitions to $\mathcal{S}^{\mathtt{T}}$ are allowed) can be bounded by the term plotted in green on Fig. 4.6b ($\sum_{t=1}^{T} \mathbb{1}_{\mathcal{E}_{k_t}}\{s_t, a_t\}$). In Lem. 4.3 we proved that this term is always bounded by $O(\sqrt{S^{\mathtt{C}}AT})$. Therefore, it is not surprising to observe a $\sqrt{T}$ increase of both the green and red curves.

Unfortunately, the growth rate of the regret will keep increasing as $\sqrt{T}$ and will never become logarithmic unlike when the MDP is communicating (in which case both UCRLB and TUCRL seem to perform equally well). This is because the condition $N_k^+ \leq \sqrt{t_k/SA}$ will always be triggered $\Theta(\sqrt{T})$ times for any T. When $\mathcal{S}^{\mathtt{T}} \neq \emptyset$, TUCRL will *restrict* the extended MDP every time the condition is triggered while when $\mathcal{S}^{\mathtt{T}} = \emptyset$, all state-action pairs will eventually be visited and so this condition will no longer be used to restrict the extended MDP. In Sec. 4.5 we show that this is not just a drawback specific to TUCRL, but it is rather an *intrinsic limitation* of learning in weakly-communicating MDPs.

Note that the big periodic jumps observed in Fig. 4.6b appear because the domain contains only one state in $\mathcal{S}^{\mathtt{T}}$ and *deterministic* transitions (only the rewards are random). For more complex environments (with random transitions) it is very difficult to predict in advance what the behaviour of TUCRL will be. However, for MDPs with high randomness in the transitions, it is likely that we do not observe "jumps" and just a smooth $\sqrt{T}$ increase (the green/red curves should always be of the same order of the orange curve as proved by Lem. 4.3, but they can be arbitrarily smooth or sharp).

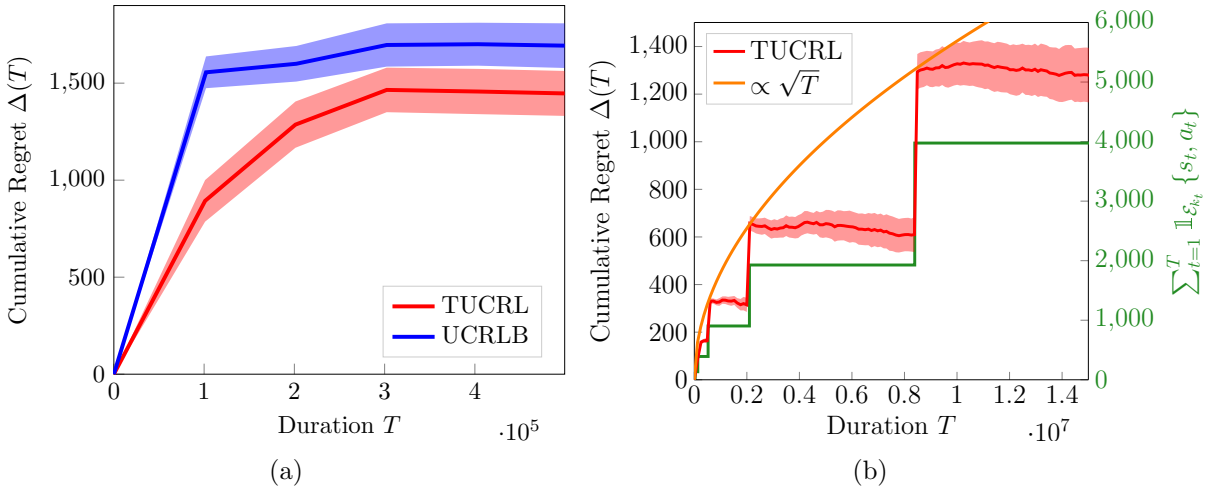


Figure 4.6: Cumulative regret of TUCRL and UCRLB on the MDPs of Fig. 4.5. Fig. 4.6a corresponds to the case where $\delta = 0.005 > 0$. Fig. 4.6a corresponds to the case where $\delta = 0$.

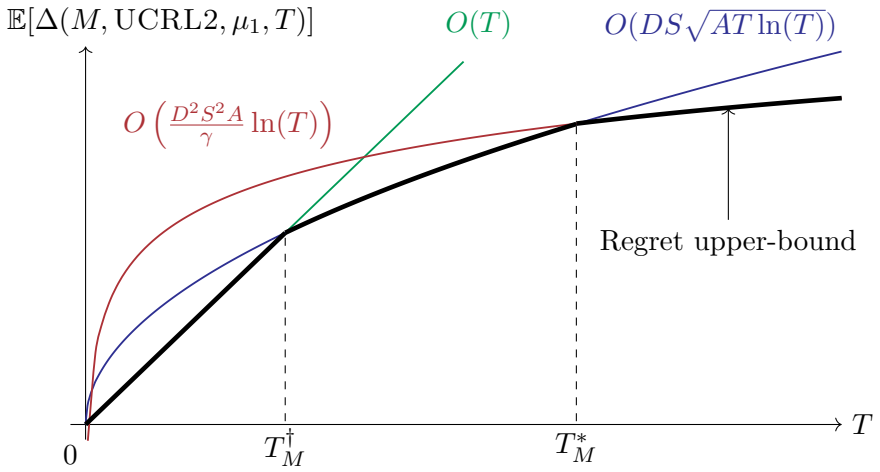


Figure 4.7: Expected regret of UCRL2 (with known horizon T given as input) as a function of T.

4.5 Learning limitations with infinite diameter

In this section we further investigate the empirical difference in the regret growth of TUCRL when the diameter is finite and infinite. We prove an impossibility result *characterizing* the *exploration-exploitation dilemma* when the diameter is *infinite*.

We first recall that the expected regret $\mathbb{E}[\Delta(M, \text{UCRL2}, \mu_1, T)]$ of UCRL2 (with input parameter $\delta = 1/3T$) after $T \geq 1$ time steps and for any finite MDP M can be bounded in several ways:

$$\mathbb{E}[\Delta(M, \text{UCRL2}, \mu_1, T)] \leq \begin{cases} r_{\max} T & \text{(by definition)} \\ 34 \cdot r_{\max} DS\sqrt{AT\ln(3T^2)} + \frac{1}{3} & \text{(Prop. 2.14)} \\ 34^2 \cdot r_{\max} \frac{D^2S^2A}{\delta_g}\ln(T) + C(M) & \text{(Prop. 2.13).} \end{cases} \tag{4.26}$$

Note that D can be replaced by Λ without changing the algorithm. The three different

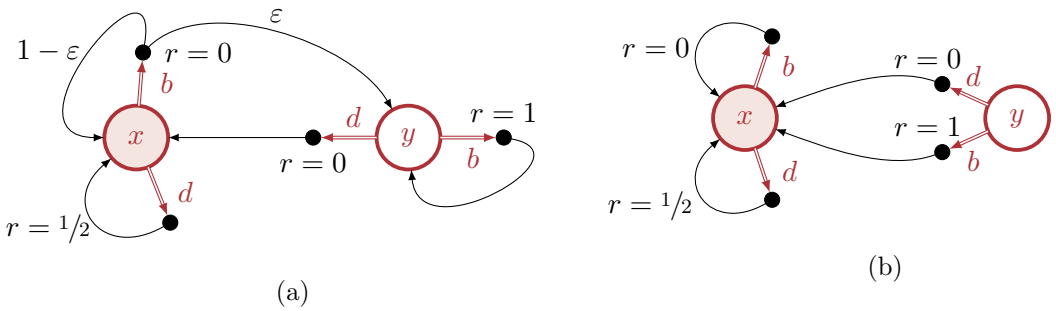


Figure 4.8: Toy example illustrating the difficulty of learning non-communicating MDPs. We represent a family of possible MDPs $\mathcal{M} = (M_\varepsilon)_{\varepsilon\in[0,1]}$ where the probability ε to go from x to y lies in $[0,1]$.

bounds lead to three different *growth rates* for the function $T \longmapsto \mathbb{E}[\Delta(M, \text{UCRL2}, \mu_1, T)]$ (see Fig. 4.7):

1. for $T_M^\dagger \geq T \geq 0$, the expected regret is linear in T,
2. for $T_M^* \geq T \geq T_M^\dagger$ the expected regret grows as $\sqrt{T}$,
3. finally for $T \geq T_M^*$, the increase in regret is only logarithmic in T.

These different *"regimes"* can also be observed empirically (both for UCRL2 and UCRLB). Using (4.26), it is easy to show that the time it takes for UCRL2 to achieve sub-linear regret is at most $T_M^\dagger = \widetilde{O}(D^2S^2A)$. We say that a learning algorithm is *efficient* when it achieves sublinear regret after a number of steps that is polynomial in the parameters of the MDP i.e., both UCRL2 and UCRLB are *efficient*. We now show with an example –similar to the example of Fig. 4.1b presented in introduction of this chapter– that *without prior knowledge*, any *efficient* learning algorithm must satisfy $T_M^* = +\infty$ when M has *infinite diameter* (i.e., it cannot achieve logarithmic regret if $D = +\infty$).

Example We consider a family of weakly-communicating MDPs $\mathcal{M} = (M_\varepsilon)_{\varepsilon\in[0,1]}$ represented on Fig. 4.8. Every MDP instance in $\mathcal{M}$ is characterised by a specific value of $\varepsilon \in [0,1]$ which corresponds to the probability to go from x to y. For $\varepsilon > 0$ (Fig. 4.8a), the optimal policy of M_ε is such that $\pi^*(x) = b$ and the optimal gain is $g_\varepsilon^* = 1$ while for $\varepsilon = 0$ (Fig. 4.8b) the optimal policy is such that $\pi^*(x) = d$ and the optimal gain is $g_0^* = 1/2$. We assume that the learning agent knows that the true MDP M belongs to $\mathcal{M}$ but does not know the specific value ε associated to $M = M_{\varepsilon^*}$. We assume that all rewards are deterministic and that the agent starts in state x (coloured).

Theorem 4.2

Let $C_1, C_2, \alpha, \beta > 0$ be positive real numbers and f a function defined for all $\varepsilon \in]0,1]$ by $f(\varepsilon) = C_1(1/\varepsilon)^\alpha$. There exists no learning algorithm $\mathfrak{A}_T$ (with known horizon T) satisfying both

1. *for all $\varepsilon \in]0,1]$, there exists $T_\varepsilon^\dagger \leq f(\varepsilon)$ such that $\mathbb{E}[\Delta(M_\varepsilon, \mathfrak{A}_T, x, T)] < 1/6 \cdot T$ for all $T \geq T_\varepsilon^\dagger$,*
2. *and there exists $T_0^* < +\infty$ such that $\mathbb{E}[\Delta(M_0, \mathfrak{A}_T, x, T)] \leq C_2(\ln(T))^\beta$ for all $T \geq T_0^*$.*

Proof. We prove the statement by contradiction: we assume that there exists a learning algorithm denoted $\mathfrak{A}_T$ satisfying

1. for all $\varepsilon \in]0,1]$, there exists $T_\varepsilon^\dagger \leq f(\varepsilon)$ such that $\mathbb{E}[\Delta(M_\varepsilon, \mathfrak{A}_T, x, T)] < 1/6 \cdot T$ for all $T \geq T_\varepsilon^\dagger$,

2. there exists $T_0^* < +\infty$ such that $\mathbb{E}[\Delta(M_0, \mathfrak{A}_T, x, T)] \leq C_2(\ln(T))^\beta$ for all $T \geq T_0^*$.

Any randomised strategy for choosing an action at time t is equivalent to an (a priori) random choice from the set of all deterministic strategies. Thus, it is sufficient to show a contradiction when the action played by $\mathfrak{A}_T$ at any time t is a deterministic function of the past trajectory $h_t := \{s_1, a_1, r_1, \ldots, s_t\}$. In the rest of the proof we assume that $\mathfrak{A}_T$ maps any sequence of observations $h_t = \{s_1, a_1, r_1, \ldots, s_t\}$ to a (single) action a_t.

By trivial induction it is easy to see that as long as state y has not been visited, the history h_t is independent of ε ($\mathfrak{A}_T$ cannot distinguish between different values of ε and plays exactly the same action when the past history is the same).

Let's define $N_T^0(x,b) := \sum_{t=1}^T \mathbb{1}\{(s_t, a_t) = (x,b)\}$ the number of visits in (x,b) with $a_t = \mathfrak{A}_T(h_t)$ and $\varepsilon = 0$. Note that $N_T^0(x,b)$ is not random since when $\varepsilon = 0$ both action b and action d loop on x with probability 1. For any $\varepsilon \in [0,1]$ and any horizon T define the event:

$$F(T,\varepsilon) := \bigcap_{1 \leq t \leq T} \{s_t \neq y\}$$

where the sequence of states s_t is obtained by executing $\mathfrak{A}_T$ on MDP M_ε. We will denote by $\overline{F(T,\varepsilon)}$ the complement of $F(T,\varepsilon)$.

For any horizon T, and independently of ε, there is only one possible trajectory $h_T = \{s_1, a_1, r_1, \ldots, s_T\}$ that never goes to y and which corresponds to the trajectory observed when $\varepsilon = 0$. When $\varepsilon = 0$, the probability of this trajectory is 1 and so $\mathbb{P}(F(T,0)) = 1$ (recall that everything is deterministic in this case) while in general we have (using the Markov property):

$$\forall T \geq 1, \ \forall \varepsilon \in [0,1], \quad \mathbb{P}(F(T,\varepsilon)) = (1-\varepsilon)^{N_T^0(x,b)}. \tag{4.27}$$

We now prove by contradiction that

$$\lim_{T\to+\infty} N_T^0(x,b) = +\infty. \tag{4.28}$$

Let's assume that $C := \max\{10, \max_{T\geq 1}\{N_T^0(x,b)\}\} < +\infty$. Taking $\varepsilon = 1/C$ and applying the law of total expectation we obtain:

$$\begin{aligned}\forall T \geq 1, \quad \mathbb{E}[\Delta(M_{1/C}, \mathfrak{A}_T, x, T)] &= \underbrace{\mathbb{E}\left[\Delta(M_{1/C}, \mathfrak{A}_T, x, T) | F(T, 1/C)\right]}_{=T/2+1/2\cdot N_T^0(x,b)\geq T/2} \cdot \underbrace{\mathbb{P}\left(F(T,1/C)\right)}_{=(1-1/C)^{N_T^0(x,b)}} \\ &\quad + \underbrace{\mathbb{E}\left[\Delta(M_{1/C}, \mathfrak{A}_T, x, T) | \overline{F(T,1/C)}\right]}_{\geq 0} \cdot \mathbb{P}\left(\overline{F(T,1/C)}\right) \\ &\geq \frac{T}{2}\cdot\left(1-\frac{1}{C}\right)^{N_T^0(x,b)} \geq \frac{T}{2}\cdot \underbrace{\left(1-\frac{1}{C}\right)^{C}}_{\geq 1/3 \quad \text{by Lem. B.1}} \geq \frac{T}{6}\end{aligned}$$

where we used the fact that

- $N_T^0(x,b) \leq C$ and $(1-1/C) \in [0,1]$ by definition, implying $\left(1-\frac{1}{C}\right)^{N_T^0(x,b)} \leq \left(1-\frac{1}{C}\right)^{C}$,
- since $C \geq 10$ we have $\left(1-\frac{1}{C}\right)^{C} \geq 1/3$ by Lem. B.1 (App. B.2) applied to $x = 1/C$,
- and finally under event $F(T, 1/C)$, the regret incurred is exactly $T/2 + 1/2 \cdot N_T^0(x,b) \geq T/2$.

This contradicts our assumption that there exists $T_{1/C}^\dagger < +\infty$ such that for all $T \geq T_{1/C}^\dagger$, $\mathbb{E}[\Delta(\mathfrak{A}_T, M_{1/C}, x, T)] < T/6$ and so (4.28) holds.

Since $\lim_{T\to+\infty} N_T^0(x,b) = +\infty$, it is possible to construct a strictly increasing sequence $(T_n)_{n\in\mathbb{N}}$ such that:

$$\forall n \in \mathbb{N},\ N_{T_{n+1}}^0(x,b) > N_{T_n}^0(x,b), \quad T_0 = T_0^*, \quad T_1 \geq C_2, \quad T_1 \geq C_2(\ln(T_1))^\beta \quad \text{and} \quad N_{T_1}^0(x,b) \geq 10$$

We also define the (strictly decreasing) sequence: $\varepsilon_n := 1/N_{T_n}^0(x,b),\ \forall n \geq 1$. By the law of total expectation:

$$\begin{aligned}\mathbb{E}[\Delta(\mathfrak{A}_{T_n}, M_{\varepsilon_n}, x, T_n)] &= \underbrace{\mathbb{E}\left[\Delta(\mathfrak{A}_{T_n}, M_{\varepsilon_n}, x, T_n) | F(T_n, \varepsilon_n)\right]}_{\geq T_n/2} \cdot \underbrace{\mathbb{P}\left(F(T_n,\varepsilon_n)\right)}_{=(1-\varepsilon_n)^{N_{T_n}^0(x,b)}} \\ &\quad + \underbrace{\mathbb{E}\left[\Delta(\mathfrak{A}_{T_n}, M_{\varepsilon_n}, x, T_n) | \overline{F(T_n,\varepsilon_n)}\right]}_{\geq 0} \cdot \mathbb{P}\left(\overline{F(T_n,\varepsilon_n)}\right) \\ &\geq \frac{T_n}{2}\cdot(1-\varepsilon_n)^{N_{T_n}^0(x,b)} = \frac{T_n}{2}\cdot \underbrace{(1-\varepsilon_n)^{1/\varepsilon_n}}_{\geq 1/3 \quad \text{by Lem. B.1}} \geq \frac{T_n}{6}\end{aligned} \tag{4.29}$$

where we applied Lem. B.1 (App. B.2) to $x = \varepsilon_n \leq 1/10$ since $N_{T_n}^0(x,b) \geq 10$ for all $n \geq 1$.

Moreover, since by construction for all $n \geq 1$, $T_n > T_0 = T_0^*$ we have by assumption that

$$\forall n \geq 1, \quad \mathbb{E}[\Delta(\mathfrak{A}_{T_n}, M_0, x, T_n)] = \frac{1}{2} N_{T_n}^0(x, b) = \frac{1}{2\varepsilon_n} \leq C_2 (\ln(T_n))^\beta$$
$$\implies T_n \geq \exp\left(\frac{1}{(2C_2 \cdot \varepsilon_n)^{1/\beta}}\right)$$

Since $\lim_{n \to +\infty} 1/\varepsilon_n = +\infty$ and $\lim_{x \to +\infty} \exp\left(x^{1/\beta}\right)/x^\alpha = +\infty$ there exists $N \in \mathbb{N}$ such that for all $n \geq N$, $T_n \geq f(\varepsilon_n)$. By assumption, for all $n \geq N$,

$$\mathbb{E}[\Delta(\mathfrak{A}_{T_n}, M_{\varepsilon_n}, x, T_n)] < \frac{T_n}{6}$$

which contradicts (4.29) therefore concluding the proof. ■

Note that point 1. in Lem. 4.2 formalizes the concept of *"efficient learnability"* introduced by Sutton and Barto (2018, Section 11.6) i.e.,*"learnable within a polynomial rather than exponential number of time steps"*. All the MDPs in $\mathcal{M}$ share the same number of states $S = 2$, number of actions $A = 2$, and gap in average reward $\gamma = 1/2$. As a result, any function of S, A and δ_g will be considered as constant. For $\varepsilon > 0$, the diameter and travel-budget coincide with the optimal bias span of the MDP and $\Lambda = D = sp(h^*) = 1/\varepsilon < +\infty$, while for $\varepsilon = 0$, $\Lambda = D = +\infty$ but $sp(h^*) = 1/2$. As shown in Eq. 4.26 and Thm. 4.1, UCRL2, UCRLB and TUCRL satisfy property 1. of Lem. 4.2 with $\alpha = 2$ and $C_1 = O(S^2 A)$ but do not satisfy 2. Lem. 4.2 proves that no algorithm can actually achieve both 1. and 2. As a result, since TUCRL satisfies 1., it cannot satisfy 2. This matches the empirical results presented in Sec. 4.4 where we observed that when the diameter is infinite, the growth rates of the regret of TUCRL is of order $\Theta(\sqrt{T})$. An algorithm that does not satisfy 1. could potentially satisfy 2. but, by definition of 1., it would suffer linear regret for a number of steps that is more than *polynomial* in the parameters of the MDP (more precisely, $e^{D^{1/\beta}}$). This is not a very desirable property and we claim that an *efficient* learning algorithm should always prefer *finite time guarantees* (1.) over *asymptotic guarantees* (2.) when both cannot be *accommodated*.

4.6 Conclusion

In this chapter we introduced TUCRL, an algorithm that efficiently balances exploration and exploitation in weakly-communicating and multi-chain MDPs, when the starting state s_1 belongs to a communicating set (Asm. 4.1). We showed that TUCRL achieves a square-root regret bound scaling with parameters (D^{c}, S^{c}, Γ^{c}) of the communicating part of the MDP and that, in the general case, it is not possible to design algorithm with logarithmic regret and polynomial dependence on the MDP parameters. Several questions remain open:

1. relaxing Asm. 4.1 by considering a transient initial state (i.e., $s_1 \in \mathcal{S}^{\mathsf{T}}$),
2. investigating whether a regret scaling as $\widetilde{\mathcal{O}}\left(\sqrt{\Lambda^{\mathsf{c}} S^{\mathsf{c}} \Gamma^{\mathsf{c}} A T}\right)$ is achievable,

3. refining the lower bound of Jaksch et al. (2010) to finally understand whether it is possible to scale with $sp(h^*)$ (at least in communicating MDPs) instead of $\Lambda \geq sp(h^*)$ (the flaw in REGAL.D may suggest it is indeed impossible).

In the next chapter, we will show that achieving a regret scaling with $sp(h^*)$ instead of Λ is at least possible when the value $sp(h^*)$ is known and given as input to the learning algorithm.

5 Exploration–exploitation with prior knowledge on the optimal bias span

5.1 Introduction

5.1.1 Bias span versus travel-budget

While the travel-budget Λ quantifies the total "cost" incurred to "recover" from a bad state *in the worst case* (i.e., when $g^* = r_{\max}$), the actual regret incurred while "recovering" is related to the difference in potential reward between "bad" and "good" states, which is accurately measured by the span (i.e., the range) $sp(h^*)$ of the optimal bias function h^*. While the travel-budget is an upper bound on the bias span (Sec. 3.3.2), it could be arbitrarily larger (e.g., weakly-communicating MDPs may have finite span and infinite travel-budget) thus suggesting that algorithms whose regret scales with the span may perform significantly better.[1] Building on the idea that the OFU principle should be mitigated by the bias span of the optimistic solution, Bartlett and Tewari (2009) proposed three different algorithms (referred to as Regal) achieving regret scaling with $sp(h^*)$ instead of Λ. The first algorithm defines a span regularized problem, where the regularization constant needs to be carefully tuned depending on the state-action pairs visited in the future, which makes it unfeasible in practice. Alternatively, they propose a constrained variant, called Regal.C, where the regularized problem is replaced by a constraint on the span. Assuming that an upper-bound c on the bias span of the optimal policy is known (i.e., $sp(h^*) \leq c$), Regal.C achieves a regret upper-bounded by $\widetilde{\mathcal{O}}(cS\sqrt{AT})$. Unfortunately, they do not propose any computationally tractable algorithm solving the constrained optimization problem, which may even be ill-posed in some cases. Finally, Regal.D avoids the need of knowing the future visits by using a doubling trick, but we argued in Chap. 4 that the analysis is flawed and probably difficult to fix.

In this chapter, we take inspiration from Regal.C and propose a constrained optimization problem for which we derive a computationally efficient algorithm, called ScOpt (analogue

[1] The proof of the minimax lower-bound (Prop. 2.12) relies on the construction of an MDP whose travel-budget actually coincides with the bias span (up to a factor 2), thus leaving the open question whether the "actual" lower-bound depends on Λ or the bias span (or an even tighter quantity).

to EVI). We identify conditions under which ScOpt converges to the optimal solution and propose a suitable stopping criterion to achieve an ε-optimal policy. Finally, we show that the convergence conditions are always satisfied and the learning algorithm obtained by integrating ScOpt into a UCRL2-like scheme (resulting into SCAL) achieves regret scaling as $\widetilde{\mathcal{O}}(\sqrt{\min\{\Lambda, c\}\Gamma SAT})$ when an upper-bound c on the optimal bias span is available.

5.1.2 Exploration bonus

In Sec. 5.7, we build on ScOpt to derive SCAL$^+$, a variant of Regal.C which enforces optimism through the use of an exploration bonus rather than an extended MDP. Regal.C estimates the true MDP (rewards and transition probabilities) and adds a state-action dependent high probability confidence bound to the reward function (not the transition probability). Strehl and Littman (2008b) were the first to exploit the idea of enforcing exploration in RL by using a "bonus" on the reward. They analysed the *infinite-horizon γ-discounted setting* and introduced the Model Based Interval Estimation with Exploration Bonus (MBIE-EB) algorithm. MBIE-EB plays the optimal policy of the empirically estimated MDP where for each state-action pair (s, a), a bonus $b(s, a)$ is added to the empirical average reward $\widehat{r}(s, a)$ i.e., the immediate reward associated to (s, a) is $\widehat{r}(s, a) + b(s, a)$. The goal of RL is to find a policy maximizing the cumulative reward i.e., the Q-function. Therefore, the bonus needs to account for the uncertainty in both the rewards and transition probabilities and so $b(s, a) = \widetilde{\Theta}\left(\frac{r_{\max}}{1-\gamma}\sqrt{\frac{1}{N(s,a)}}\right)$ where $\frac{r_{\max}}{1-\gamma}$ is the range of the Q-function. Strehl and Littman (2008b) also derived PAC guarantees on the sample complexity of MBIE-EB. More recently, *count-based methods* (e.g., Bellemare et al., 2016; Tang et al., 2017; Ostrovski et al., 2017; Martin et al., 2017) tried to combine the idea of MBIE-EB with Deep RL (DRL) techniques to achieve a good exploration-exploitation trade off in high dimensional problems. The exploration bonus usually used has a similar form $\widetilde{\Theta}\left(\frac{\beta}{\sqrt{N}}\right)$ where β is now an hyper-parameter tuned for the specific task at hand, and the visit count N is approximated using discretization (e.g., hashing) or density estimation methods.

Exploration bonuses have also been successfully applied to *finite-horizon problems* (Azar et al., 2017; Kakade et al., 2018; Jin et al., 2018). In this setting, the planning horizon H is known to the learning agent and the range of the Q-function is $r_{\max}H$. A natural choice for the bonus is then $b(s, a) = \widetilde{\Theta}(r_{\max}H/\sqrt{N(s, a)})$. UCBVI_1 introduced by Azar et al. (2017) uses such a bonus and achieves near-optimal regret guarantees $\widetilde{O}(H\sqrt{SAT})$. Extensions of UCBVI_1 exploiting the variance instead of the range of the Q-function achieve a better regret bound $\widetilde{O}(\sqrt{HSAT})$ (Azar et al., 2017; Kakade et al., 2018; Jin et al., 2018).

Both the finite horizon setting and infinite horizon discounted setting assume that there exists an *intrinsic horizon* (respectively H and $\frac{1}{1-\gamma}$) known to the learning agent. Unfortunately, in many common RL problems it is not clear how to define H or $\frac{1}{1-\gamma}$ and it is often desirable to set them as big as possible (e.g., in episodic problem, the time to the goal is not known in advance and random in general). As H tends to infinity the regret (of UCBVI_1, etc.) will become linear while as γ tends to to 1 the sample complexity (of MBIE-EB, etc.) tends to infinity (not to mention the numerical instabilities that may arise). In this chapter

we analyze the exploration bonus approach in the infinite horizon undiscounted setting which generalizes the two previous settings to the case where $H \to +\infty$ and $\gamma \to 1$ respectively (see Sec. 2.2). Although REGAL.C can be efficiently implemented in the tabular case, it is difficult to extend it to more scalable approaches like DRL. In contrast, as already mentioned, the exploration bonus approach is simpler to adapt to large scale problems and inspired count based methods in DRL.

SCAL$^+$ is the first algorithm that relies on an exploration bonus to efficiently balance exploration and exploitation in the infinite-horizon undiscounted setting. All the exploration bonuses that were previously introduced in the RL literature explicitly depend on γ or H which are known to the learning agent. In the infinite-horizon undiscounted case, there is no predefined parameter informing the agent about the range of the Q-function. This makes the design of an exploration bonus very challenging. To overcome this limitation, we make the same assumption as in REGAL.C and SCAL i.e., we assume that the agent knows an upper-bound c on the span (i.e., range) of the optimal bias (i.e., value function). The exploration bonus used by SCAL$^+$ is thus $b(s,a) = \widetilde{\Theta}\big(\max\{c, r_{\max}\}/\sqrt{N(s,a)}\big)$. In comparison, other algorithms in the infinite horizon undiscounted setting like UCRLB or SCAL can, to a certain extent, be interpreted as virtually using an exploration bonus of order $\widetilde{\Theta}\big(\max\{\Lambda, r_{\max}\}\sqrt{\Gamma/N(s,a)}\big)$ and $\widetilde{\Theta}\big(\max\{c, r_{\max}\}\sqrt{\Gamma/N(s,a)}\big)$ respectively. This is bigger by a multiplicative factor $\sqrt{\Gamma}$. As a result, to the best of our knowledge, SCAL$^+$ achieves a *"tighter"* optimism than any other existing algorithm in the infinite horizon undiscounted setting and is therefore less prone to *over-exploration*. Surprisingly, the tighter optimism introduced by SCAL$^+$ compared to SCAL and UCRLB is not reflected in the final regret bound with the current statistical analysis ($\sqrt{\Gamma}$ appears in the bound although not being included in the bonus). We isolate and discuss where the term $\sqrt{\Gamma}$ appears in the proof sketch of Sect. 5.7.3. While Azar et al. (2017); Kakade et al. (2018); Jin et al. (2018) managed to remove the $\sqrt{\Gamma}$ term in the finite horizon setting, it remains an open question whether their result can be extended to the infinite horizon case (for example, the two definitions of regret do not match and differ by a linear term). Finally, the analysis of Sec. 3.6 does not apply to SCAL$^+$ because c explicitly appears outside the square-root in the expression of the bonus. Overall, SCAL only achieves a regret of order $\widetilde{\mathcal{O}}(\max\{r_{\max}, c\}\sqrt{\Gamma SAT})$ which is worse than SCAL.

Despite achieving a looser regret bound, SCAL$^+$ achieves a tighter optimism. In Sec. 5.8 we show how to combine the advantages of SCAL and SCAL$^+$ into a single algorithm named SCAL*.

The work presented in this chapter extends the conference paper (Fruit et al., 2018b) and the paper under submission (Qian et al., 2018b).

5.2 Span-constrained exploration–exploitation in RL: REGAL.C and relaxations

5.2.1 The approach of REGAL.C

Our first algorithm SCAL (Sec. 5.5) is a *tractable* variant of REGAL.C. We therefore start by recalling the algorithmic structure of REGAL.C. REGAL.C follows the same steps as UCRL2 (and UCRLB) but instead of solving problem (2.34) at each episode (see Chap. 2 and 3), it tries to find the best *optimistic* model $M \in \mathcal{M}_c$ having constrained *optimal* bias span i.e.,

$$\sup_{M \in \mathcal{M}_c} \left\{ \max_{\pi \in \Pi^{\mathrm{SD}}} g_M^\pi \right\} = \sup_{M \in \mathcal{M}_c} g_M^* \tag{5.1}$$

where the bounded parameter MDP $\mathcal{M}_c$ is the set of plausible MDPs with span of the optimal bias bounded by c i.e.,

$$\mathcal{M}_c := \{M \in \mathcal{M} : sp\,(h_M^*) \leq c\}. \tag{5.2}$$

REGAL.C *discards* any MDP $M \in \mathcal{M}$ whose *optimal* policy has a span larger than c (i.e., such that $sp\,(h_M^*) > c$) and looks for the MDP with highest *optimal* gain g_M^* among all *remaining* MDPs.

Well-posedness. There is no guarantee that all the MDPs in $\mathcal{M}_c$ are weakly communicating and thus have *state-independent* gain.[2] This could make the comparison of policies difficult. Two policies $\pi_1, \pi_2 \in \Pi^{\mathrm{SR}}$ with state-dependent gain cannot necessarily be compared since we might have $g_M^{\pi_1}(s) > g_M^{\pi_2}(s)$ for some state $s \in \mathcal{S}$ while $g_M^{\pi_1}(s') < g_M^{\pi_2}(s')$ for some other state $s' \neq s$. When there is no constraint on the bias, this is not a problem as we can prove that there always exists a policy that *dominates* all others component-wise (Puterman, 1994, Chapter 9).[3] When there exists a constraint on the bias, this may no longer be the case. As a result, unlike in the case of UCRL2 and UCRLB, the supremum (5.1) might not always be *well-defined*[4] and we suspect the problem to be *ill-posed* in general. This intuition comes from Ex. 5.1a (that will be presented in Sec. 5.3) where we show the necessity of *enforcing* a state-independent gain (i.e., as a constraint of the optimization problem). Moreover, even if we ignored all the problems in the formulation of REGAL.C and assumed that (5.1) was *well-posed*, searching the space $\mathcal{M}_c$ seems to be *computationally intractable*. Finally, for any $M \in \mathcal{M}$, there may exist several optimal policies with different bias and some of them may not satisfy the Bellman optimality equation (see Prop. 2.4) and are thus difficult to compute. In the next section, we introduce a *relaxation* of problem 5.1 that is both well-posed and easier to analyse.

[2] For example, the extended MDPs that we have considered so far contain multi-chain MDPs.

[3] If the MDP is weakly-communicating, the optimal gain is even state-independent as shown in Prop. 2.4 which is why (2.34) is well-posed.

[4] Making the problem well-posed would require to fix a "reference" state or a distribution over states.

5.2.2 A first relaxation of REGAL.C

The high-level idea of our relaxation is to replace the constraint on the *set of plausible MDPs* (bounded parameter MDP) by a constraint on the *policy space*. Formally, we modify problem (5.1) as follows:

$$\sup_{M \in \mathcal{M}} \left\{ \sup_{\pi \in \Pi_c(M)} g_M^\pi \right\} \tag{5.3}$$

where the policy space $\Pi_c(M)$ is defined as

$$\Pi_c(M) := \left\{ \pi \in \Pi^{SR} : sp\left(h_M^\pi\right) \leq c \quad \text{and} \quad sp\left(g_M^\pi\right) = 0 \right\}. \tag{5.4}$$

By convention, we set $\max_{\pi \in \Pi_c(M)} \{g_M^\pi\}$ to $-\infty$ when $\Pi_c(M) = \emptyset$. The condition $sp\left(g_M^\pi\right) = 0$ makes sure that the policy space only contains policies with *state-independent gain*. As a result, two policies can always be compared by comparing their gains (and so problem (5.3) is well-posed). Note that we do not restric attention to deterministic stationary policies, but consider also *randomized* policies. It will quickly become clear that considering randomized policies makes the problem easier to solve and analyze.

Equivalent extended formulation. One of the advantages of (5.3) over (5.1) is that it can be reformulated as finding a *gain-maximizing* policy of an *extended MDP*. Just as solving (2.34) is equivalent to solving (2.35) (see Chap. 2), problem (5.3) is *equivalent* to solving the following optimization problem:

$$\sup_{\pi^+ \in \Pi_c(\mathcal{M}^+)} g_{\mathcal{M}^+}^\pi \tag{5.5}$$

where $\mathcal{M}^+$ is the extended MDP associated with the bounded parameter MDP $\mathcal{M}$. Unlike (5.1), for *every* MDP in $\mathcal{M}$ (not just those in $\mathcal{M}_c$), (5.3) considers *all* (stationary) policies with *constant gain* satisfying the *span constraint* (not just the deterministic optimal policies).

Existence of the maximum and relaxation. Since $(M, \pi) \mapsto g_M^\pi$ and $(M, \pi) \mapsto sp\left(h_M^\pi\right)$ are in general non-continuous functions, the argmax in (5.1) and (5.3) may not exist (i.e., the maximum may not be reached). Despite this technical difficulty, we can show that (5.3) is always a *relaxation* of (5.1) in terms of supremum value (provided we enforce the additional

constraint of state-independent gain in (5.1)).

Proposition 5.1

Define $\overline{\mathcal{M}}_c := \mathcal{M}_c \cap \{M \in \mathcal{M} : \ sp\,(g_M^*) = 0\}$ *the restriction of* $\mathcal{M}_c$ *to MDPs that have state-independent optimal gain. Then*

$$\sup_{M \in \overline{\mathcal{M}}_c} \left\{ \max_{\pi \in \Pi^{SD}} g_M^\pi \right\} = \sup_{M \in \overline{\mathcal{M}}_c} \{g_M^*\} \leq \sup_{M \in \mathcal{M}} \left\{ \sup_{\pi \in \Pi_c(M)} g_M^\pi \right\}.$$

Proof. Let $M \in \overline{\mathcal{M}}_c$ and denote by π^* an optimal policy of M, with g_M^* and h_M^* the associated gain an bias. By definition $sp\,(g_M^*) = 0$ and $sp\,(h_M^*) \leq c$ and so $\pi^* \in \Pi_c(M)$. Therefore, $g_M^* \leq \sup_{\pi \in \Pi_c(M)} g_M^\pi$. Since g_M^* is the optimal gain of M (maximum over all policies), we actualy have an equality: $g_M^* = \sup_{\pi \in \Pi_c(M)} g_M^\pi$. Since this is true for all $M \in \overline{\mathcal{M}}_c$, we have

$$\sup_{M \in \overline{\mathcal{M}}_c} \{g_M^*\} = \sup_{M \in \overline{\mathcal{M}}_c} \left\{ \sup_{\pi \in \Pi_c(M)} g_M^\pi \right\} \leq \sup_{M \in \mathcal{M}} \left\{ \sup_{\pi \in \Pi_c(M)} g_M^\pi \right\}$$

where the inequality follows from the fact that $\overline{\mathcal{M}}_c \subseteq \mathcal{M}$. ■

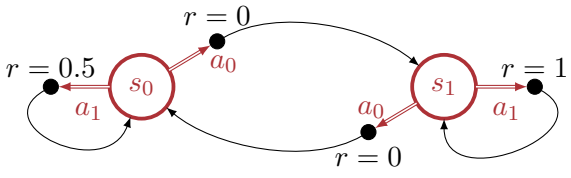
Due to Prop. 5.1, if the solution of (5.1) is *optimistic* i.e., bigger than the optimal gain g^* of the true unknown MDP, so is the solution of (5.3). As a result, any algorithm solving (5.3) should intuitively enjoy the same regret guarantees as REGAL.C (which solves (5.1)). In the following we further characterize problem (5.3), introduce a *truncated* value iteration algorithm to solve it (called SCOPT), and finally integrate it into a UCRL2-like scheme to recover REGAL.C regret guarantees.

5.3 The Optimization Problem

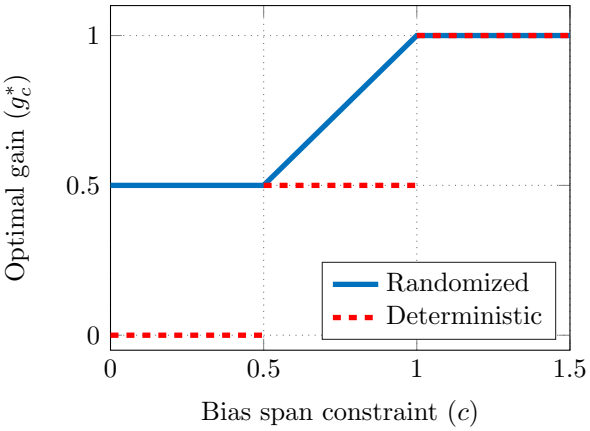
In the previous section, we showed that our new optimization problem (Eq. 5.3) can be equivalently formulated as a *span-constrained gain-maximization* problem on the extended MDP (Eq. 5.5). In this section we analyze some properties of the following optimization problem (of which (5.5) is an instance),

$$\sup_{\pi \in \Pi_c(M)} g_M^\pi := g_c^*(M) \tag{5.6}$$

where M is any MDP (with discrete or compact action space) such that $\Pi_c(M) \neq \emptyset$. Problem (5.6) aims at finding a policy that maximizes the gain g_M^π within the set of randomized policies with constant gain (i.e., $sp\,(g_M^\pi) = 0$) and bias span smaller than c (i.e., $sp\,(h_M^\pi) \leq c$). Since $g_M^\pi \in [0, r_{\max}]$ (i.e., g_M^π is bounded), the supremum always exists and we denote it by $g_c^*(M)$. The set of maximizers is denoted by $\Pi_c^*(M) \subseteq \Pi_c(M)$, with elements $\pi_c^*(M)$ (if $\Pi_c^*(M)$ is non-empty). In order to give some intuition about the solution(s) of problem (5.6), we introduce the following illustrative MDP.



(a)



(b)

Figure 5.1: Toy example illustrating the properties of optimization problem (5.6). Fig. 5.1a: The MDP is communicating and only has deterministic transitions and rewards for all actions (2 actions per state). Fig. 5.1b: Maximum gain achievable g_c^* (y-axis) as a function of the span constraint c (x-axis) with all (randomized) stationary policies (blue line) and only deterministic policies (dashed red line). Only policies with state-independent gain are considered (i.e., the policy playing a_1 in both states is ignored).

Example Consider the two-state MDP depicted in Fig. 5.1a. Since there are only two actions a_0 and a_1 in both states, for any stationary policy $\pi = (d)^\infty \in \Pi^{\mathrm{SR}}$, the associated decision rule $d \in D^{\mathrm{MR}}$ can be parametrized by two quantities: x (the probability to play a_0 in s_0) and y (the probability to play a_0 in s_1). With this parametrization:

$$P_d = \begin{bmatrix} 1-x & x \\ y & 1-y \end{bmatrix}, \; r_d = \begin{bmatrix} \frac{1-x}{2} \\ 1-y \end{bmatrix}.$$

We can compute the gain $g = [g_1, g_2]$ and the bias $h = [h_1, h_2]$ by solving the linear system (2.11). For any $x > 0$ or $y > 0$, we obtain

$$g_1 = g_2 = \frac{1}{2} + \frac{x(1-3y)}{2(x+y)}; \quad h_2 - h_1 = \frac{1}{2} + \frac{1-3y}{2(x+y)},$$

while for $x = 0$, $y = 0$, we have $g_1 = 1/2$ and $g_2 = 1$, with $h_2 = h_1 = 0$. Note that $0 \leq sp\,(h^\pi) \leq 1$ for any $\pi \in \Pi^{\mathrm{SR}}$. By considering different values for x and y, this example allows us to analyze the properties of optimization problem (5.6). For example, on Fig. 5.1b we show how the solution of (5.6) varies with the span constraint c. We also show the evolution when the policy space is restricted to deterministic policies. This curves can be easily deduced from the above formulas for g_1/g_2 and $h_2 - h_1$. Note that Fig. 5.1b ignores the case $x = y = 0$ since it corresponds to the only policy with state-dependent gain.

Randomized policies. When the bias span is unconstrained, there always exist an optimal stationary *deterministic* policy (see Sec. 2.2). In contrast, the following lemma shows that there may not exist any deterministic policy solution to (5.6) even if a randomized solution

exists.

Lemma 5.1

There exists an MDP M and a scalar $c \geq 0$, such that $\Pi_c^*(M) \neq \emptyset$ and $\Pi_c^*(M) \cap \Pi^{SD}(M) = \emptyset$ (i.e., the solution of (5.6) is not a deterministic policy).

Proof. Consider Ex. 5.1a with constraint $1/2 < c < 1$ (see Fig. 5.1b for a graphical representation). The only deterministic policy π_D with constant gain and bias span smaller than c corresponds to $x = 0$ and $y = 1$, which leads to $g^{\pi_D} = 1/2$ and $sp\left(h^{\pi_D}\right) = 1/2$. On the other hand, the randomized policy π_R corresponding to $x = 1$ and $y = (1-c)/(1+c)$, satisfies $sp\left(h^{\pi_R}\right) = c$ and $g^{\pi_R} = c > g^{\pi_D}$, thus proving the statement. ■

Constant gain. The following lemma shows that if we consider non-constant-gain policies, the supremum in (5.6) may not be well defined, as no *dominating* policy exists. A policy $\pi \in \Pi^{\mathrm{SR}}$ is *dominating* if for any policy $\pi' \in \Pi^{\mathrm{SR}}$, $g^{\pi}(s) \geq g^{\pi'}(s)$ in all states $s \in \mathcal{S}$.

Lemma 5.2

There exists an MDP M and a scalar $c \geq 0$, such that there exists no dominating policy π in Π^{SR} with constrained bias span (i.e., $sp\left(h^{\pi}\right) \leq c$).

Proof. Consider Ex. 5.1a with constraint $1/2 < c < 1$ (see Fig. 5.1b for a graphical representation). As shown in the proof of Lem. 5.1, the optimal stationary policy π_R with constant gain satisfies $g_c^* = [c, c]$. On the other hand, the only policy π with non-constant gain is $x = 0$, $y = 0$, which has $sp\left(h^{\pi}\right) = 0 < c$ and $g^{\pi}(s_0) = 1/2 < c = g_c^*$ and $g^{\pi}(s_1) = 1 > c = g_c^*$, thus proving the statement. ■

Lem. 5.2 shows that when the search space is *not* restricted to policies with state-independent gain, problem (5.6) is *not well-posed*. We suspect that the same problem arises with REGAL.C (see (5.1)) although it is much more difficult to derive a counter-example in that case ($\mathcal{M}_c$ is a more complex mathematical object).

Existence of a maximizer. Whether problem (5.6) always admits a maximizer (i.e., whether $\Pi_c^*(M) \neq \emptyset$) when the search space is not empty (i.e., when $\Pi_c(M) \neq \emptyset$) is left as an open question. This question may not be easy to answer since in general, $\pi \mapsto g^{\pi}$ is not a continuous map and Π_c is not a closed set (and therefore classical results of topology do not apply). For instance in Ex. 5.1a, although the maximum is attained, the point $x = 0$, $y = 0$ does not belong to Π_c (i.e., Π_c is not closed) and g^{π} is not continuous at this point. Notice that in the particular case where the MDP is *unichain* (see Def. 2.2), Π_c is compact, $\pi \mapsto g^{\pi}$ is continuous, and we can prove the following lemma:

Lemma 5.3

If M is unichain then $\Pi_c^*(M) \neq \emptyset$.

Proof. The proof can be found in (Fruit et al., 2018b, Appendix A.1). ■

The goal of this section was to better understand problem (5.5) (equivalent to (5.3)) by analyzing the more general problem (5.6). We saw that this problem is not as easy as its unconstrained counterpart (2.35). In the next sections, we will show how to construct an extended MDP so that (5.5) admits a maximizer (e.g., the extended MDP will be unichain so that Lem. 5.3 holds) and the problem can be efficiently solved.

5.4 Planning with SCOPT

In this section, we introduce SCOPT and derive *sufficient conditions* for its convergence to the solution of (5.6). In Fruit et al. (2018b, Appendix B) we show examples where convergence to the solution of (5.6) does not hold when these conditions are not satisfied, implying that these conditions are also *necessary* in some sense. In the next section, we will show that these conditions always hold when SCOPT is carefully integrated into UCRLB.

5.4.1 Span-constrained value and policy operators

SCOPT is a version of (relative) value iteration (Puterman, 1994; Bertsekas, 1995), where the optimal Bellman operator is modified ("truncated") to return value functions with span bounded by c, and the stopping condition is tailored to return a *constrained greedy* policy with near-optimal gain.

Topology of the span "truncation" operator. Let $\mathcal{B}_c := \{v : sp(v) \leq c\}$ be the *"semi-ball"* of span constrained value functions (we recall that $sp(\cdot)$ is a semi-norm).

Definition 5.1

For any vector $v \in \mathbb{R}^S$ and any $c \geq 0$, the span-truncation *operator $\Gamma_c : \mathbb{R}^S \to \mathcal{B}_c$ is defined as: $\Gamma_c v(s) := \min\{v(s), \min_x v(x) + c\}$ for all $v \in \mathbb{R}^S$ and $s \in \mathcal{S}$.*

The following lemma shows that Γ_c can be seen as a *projection* operator (in span semi-norm) on the semi-ball $\mathcal{B}_c$.

Lemma 5.4

For any vector $v \in \mathbb{R}^S$ and $c \geq 0$, $\Gamma_c v$ is a projection of v on the semi-ball V_c in span semi-norm i.e.,

$$\Gamma_c v \in \arg\min_{z \in \mathcal{B}_c} sp(z - v).$$

Proof. See App. C. ■

Note that the projection is not uniquely defined: for any $\lambda \in \mathbb{R}$, $\Gamma_c v + \lambda e$ is also the projection of v on the semi-ball $\mathcal{B}_c$ (because $sp(e) = 0$). We provide a *geometric illustration*

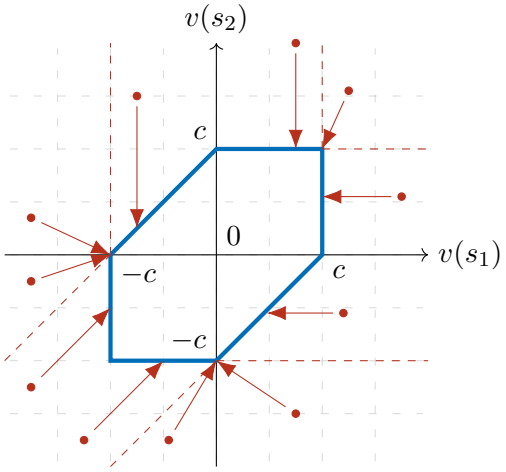


Figure 5.2: Geometric representation of projection Γ_c in the 3-dimensional case ($S = 3$).

of Γ_c in the three-dimensional case ($S = 3$) on Fig. 5.2. For simplicity we represent Γ_c in the normed *quotient space* induced by the semi-norm $sp(\cdot)$ on $\mathbb{R}^3$. In the quotient space, $sp(\cdot)$ is an actual norm and $\mathcal{B}_c$ is an actual ball of radius c for that norm. Since the *null space* of $sp(\cdot)$ is the set of vectors of the form λe with $\lambda \in \mathbb{R}$, it is immediate to see that the quotient space is in bijection with $\mathbb{R}^2 \times \{0\}$ (one coordinate is set to 0 and the others are free variables). In Fig. 5.2 the tird dimension $v(s_3)$ is set to 0 while $v(s_1)$ and $v(s_2)$ are represented on the x and y axis respectively. The ball $\mathcal{B}_c$ is represented by a bue line and the red arrows correspond to the projection Γ_c on $\mathcal{B}_c$. We can divide $\mathbb{R}^2$ in different areas (separated by dashed red lines on the figure) where projecting a point located outside the ball onto the ball has a different effect. By definition of Γ_c, every point inside the ball is an invariant of Γ_c.

Like L, Γ_c satisfies 3 important properties that are key to apply the tools of Chap. 3 while enforcing the constraint on the bias: *monotonicity*, *non-expansiveness* and *"linearity"*.

Lemma 5.5 (Analogue of Prop. 2.5)

Let v and u be any two vectors in $\mathbb{R}^S$, then:

(a) Γ_c *is* monotone*:* $v \geq u \implies \Gamma_c v \geq \Gamma_c u$.

(b) Γ_c *is* non-expansive *both in span semi-norm and ℓ_∞-norm:*

$$sp(\Gamma_c v - \Gamma_c u) \leq sp(v - u) \quad \text{and} \quad \|\Gamma_c v - \Gamma_c u\|_\infty \leq \|v - u\|_\infty.$$

(c) Γ_c *is* linear[5]*:* $\forall \lambda \in \mathbb{R}, \quad \Gamma_c(v + \lambda e) = \Gamma_c v + \lambda e$.

Proof. The proof can be found in (Fruit et al., 2018b, Lemma 15, Appendix D.2). ■

[5]We recall that the word "linear" is an abuse of terminology and does not refer to the same property as in linear algebra (see Prop. 2.5).

Span truncated greedy operator. We now introduce a *constrained* (truncated) version of the optimal Bellman operator by composing L with the span truncation (projection) Γ_c.

Definition 5.2

Given $c \geq 0$, we define operator $T_c : \mathbb{R}^S \to \mathbb{R}^S$ as: $T_c v := \Gamma_c(Lv)$, for all $v \in \mathbb{R}^S$.

In other words, operator T_c applies a *span truncation* to the one-step application of L, that is, for any state $s \in \mathcal{S}$, $T_c v(s) = \min\{Lv(s), \min_x Lv(x) + c\}$, which guarantees that $sp(T_c v) \leq c$ (by definition). A first major observation is that unlike L, operator T_c is not always *associated with a decision rule* d s.t. $T_c v = L_d v$.

Definition 5.3

We say that T_c is feasible *at $v \in \mathbb{R}^S$ and $s \in \mathcal{S}$ if there exists a Markov decision rule $d \in D^{MR}$ such that $T_c v(s) = L_d v(s)$. When T_c is feasible at v and all states $s \in \mathcal{S}$ (i.e., when there exists a Markov decision rule $d \in D^{MR}$ such that $T_c v = L_d v$ componentwise) we say that T_c is* globally feasible *at v.*

In the following lemma, we identify sufficient and necessary conditions for (global) *feasibility* of T_c.

Lemma 5.6

Operator T_c is feasible at $v \in \mathbb{R}^S$ and $s \in \mathcal{S}$ if and only if

$$\min_{a \in \mathcal{A}_s} \{r(s,a) + p(\cdot|s,a)^T v\} \leq \min_{s'} \{Lv(s')\} + c. \tag{5.7}$$

Furthermore, let

$$D(c,v) := \left\{ d \in D^{MR} \mid sp(L_d v) \leq c \right\} \tag{5.8}$$

be the set of randomized decision rules d whose associated operator L_d returns a span-constrained value function when applied to v. $T_c v$ is globally feasible if and only if $D(c,v) \neq \emptyset$, in which case we have

$$T_c v = \max_{d \in D(c,v)} L_d v. \tag{5.9}$$

Proof. The proof can be found in (Fruit et al., 2018b, Appendix D.1). ■

Lem. 5.6 shows that it is suffient to have $sp(L_d v) \leq c$ for at least one decision rule $d \in D^{\mathrm{MR}}$ in order to guarantee that $T_c v = L_\delta v$ for some $\delta \in D^{\mathrm{MR}}$ (potentially different than d). This result is a priori not so obvious although it is not difficult to prove. The last part of this lemma shows that when T_c is globally feasible at v (i.e., when $D(c,v) \neq \emptyset$), $T_c v$ is the *componentwise maximal* value function of the form $L_d v$ with decision rule $d \in D^{\mathrm{MR}}$ satisfying $sp(L_d v) \leq c$. Surprisingly, even in the presence of a constraint on the one-step value span, such a *componentwise* maximum still exists. This is not as obvious as in the case of the greedy operator L since the constraint on the span creates a *correlation* between states (while all

Algorithm 8 Span truncated greedy operator (T_c, G_c)

Input: MDP M (with optimal Bellman operator L), span constraint c, vector $v \in \mathbb{R}^S$
Output: Span constrained vector $w \in \mathbb{R}^S$, Decision rule $d_c^v \in D^{\text{MR}}$

1: Compute $u \leftarrow Lv$ and $d^+ \in \arg\max_{d \in D^{\text{MD}}} \{L_d v\}$ ▷ Break ties arbitrarily
2: Set $u_{\min} \leftarrow \min_{s \in \mathcal{S}} \{u(s)\}$
3: **for** $s \in \mathcal{S}$ **do** ▷ This loop can be parallelized
4: **if** $u(s) > u_{\min} + c$ **then**
5: $w(s) \leftarrow u_{\min} + c$ ▷ See Def. 5.2
6: $m \leftarrow \min_{a \in \mathcal{A}_s} \{r(s,a) + p(\cdot|s,a)^\intercal v\}$
7: $a^- \in \arg\min_{a \in \mathcal{A}_s} \{r(s,a) + p(\cdot|s,a)^\intercal v\}$ ▷ Break ties arbitrarily
8: $d_c^v(a^-|s) \leftarrow \min\left\{(u(s) - u_{\min} - c)/(u(s) - m), 1\right\}$
9: $d_c^v(d^+(s)|s) \leftarrow \max\left\{(u_{\min} + c - m)/(u(s) - m), 0\right\}$
10: $d_c^v(a|s) \leftarrow 0$ for all $a \neq a^-, d^+(s)$
11: **else**
12: $w(s) \leftarrow u(s)$ ▷ See Def. 5.2
13: $d_c^v(d^+(s)|s) \leftarrow 1$ ▷ Greedy action
14: $d_c^v(a|s) \leftarrow 0$ for all $a \neq d^+(s)$
15: **end if**
16: **end for**

states are independent in the case of operator L). As a consequence, whenever $D(c, v) \neq \emptyset$, optimization problem (5.9) can be seen as the solution of the following LP-problem:

$$\max_{d \in D(c,v)} \{(L_d v)^\intercal e\} \tag{5.10}$$

where $d \mapsto L_d v$ is a linear map and the set $D(c, v)$ can be expressed as a set of $S \times (S-1)$ linear constraints on $L_d v$:

$$L_d v(s) - L_d v(s') \leq c, \quad \forall s \neq s'.$$

It goes without saying that it is *computationally more efficient* to calculate $T_c v$ using Def. 5.2 than solving the LP (5.10). Moreover, to compute the decision rule $d_c^v \in D(c, v)$ achieving the maximum value $T_c v$ in (5.9), there is also a much more efficient algorithm than using a generic LP solver on (5.10). Alg. 8 describes how to *simulataneously* (and efficiently) compute $T_c v$ and the associated policy d_c^v when $D(c, v) \neq \emptyset$. In the states $s \in \mathcal{S}$ where the span constraint c is not violated, $d_c^v(\cdot|s)$ just plays the *greedy* action with probability 1 (associated to the optimal Bellman operator L). In the states $s \in \mathcal{S}$ where the constraint is violated, $d_c^v(\cdot|s)$ assigns non-zero probability mass to the greedy action as well as the *"anti-greedy"* action (i.e., the action achieving the minimum value instead of the maximum, see line 7 of Alg. 8). The probability mass is tuned so as to ensure that the expectated value is exactly equal to $\min\{Lv(s)\} + c$, therefore matching the value of $T_c v(s)$. More precisely, using the notation in Alg. 8, whenever $D(c, v) \neq \emptyset$ and $u(s) > u_{\min} + c = \min\{Lv(s)\} + c$, we always have (as a consequence of Eq. (5.7) in Lem. 5.6):

$$d_c^v(a^-|s) = \min\left\{\frac{u(s)-u_{\min}-c}{u(s)-m}, 1\right\} = \frac{u(s)-u_{\min}-c}{u(s)-m}$$
$$\text{and}\quad d_c^v(d^+(s)|s) = \max\left\{\frac{u_{\min}+c-m}{u(s)-m}, 0\right\} = \frac{u_{\min}+c-m}{u(s)-m}$$

and therefore:

$$\left(\frac{\cancel{u(s)}-u_{\min}-c}{u(s)-m}\right)\cdot m + \left(\frac{u_{\min}+c-\cancel{m}}{u(s)-m}\right)\cdot u(s) = u_{\min}+c = w(s) = T_c v(s).$$

On the other hand, whenever $D(c,v) \neq \emptyset$, there exists at least one state $s \in \mathcal{S}$ such that $u(s) \geq m > u_{\min}+c$ (as a consequence of Eq. (5.7) in Lem. 5.6). In this case, $d_c^v(\cdot|s)$ just plays the "anti-greedy" action with probability 1 and $T_c v \neq L_{d_c^v} v$ but there exists no decision rule satisfying the equality in any case (Lem. 5.6). However, it is immediate to verify that $d_c^v \in \arg\min_{d\in D^{\text{MR}}}\{|T_c v(s) - L_d v(s)|\}$ for all states $s \in \mathcal{S}$ and so in some sense, d_c^v is the decision rule that is the *"closest"* to $T_c v$.

Definition 5.4

We define the operator $G_c : \mathbb{R}^S \to D^{MR}$ by $G_c v := d_c^v$ for all $v \in \mathbb{R}^S$, where d_c^v is the decision rule output by Alg. 8 (with c and v as inputs).[6]

We conclude this paragraph with three useful properties satisfied by operator T_c (analogue of Lem. 5.5).

Lemma 5.7

Let v and u be any two vectors in $\mathbb{R}^S$, then:

(a) T_c is monotone*: $v \geq u \implies T_c v \geq T_c u$.*

(b) T_c is non-expansive *both in span semi-norm and ℓ_∞-norm:*

$$sp\left(T_c v - T_c u\right) \leq sp\left(v-u\right) \quad \text{and} \quad \|T_c v - T_c u\|_\infty \leq \|v-u\|_\infty.$$

(c) T_c is linear*: $\forall \lambda \in \mathbb{R},\ \ T_c\left(v+\lambda e\right) = T_c v + \lambda e$.*

Proof. Both L and Γ_c satisfy *(a)*, *(b)* and *(c)* and since $T_c = \Gamma_c L$ (Def. 5.2), the result follows by composition of operators. ∎

Span truncated value iteration. We are now ready to introduce ScOpt (Alg. 9). Given a vector $v_0 \in \mathbb{R}^S$ and a reference state $\overline{s}$, ScOpt implements relative value iteration where L is replaced by T_c, i.e., $v_{n+1} = T_c v_n - T_c v_n(\overline{s})e$ for some arbitrary reference state $\overline{s} \in \mathcal{S}$. Notice that the term $(T_c v_n)(\overline{s})e$ subtracted at any iteration n prevents v_n from increasing linearly with n and thus avoids *numerical instability*. However, the subtraction can be *dropped* without affecting the convergence properties of ScOpt (see Alg. 3 and the discussion in Sec. 2.1.3). If the stopping condition is met at iteration n, ScOpt returns a policy $\pi_n = (d_n)^\infty$ where $d_n = G_c v_n$ (among other things).

[6] When there are multiple greedy and anti-greedy actions, Alg. 8 break ties arbitrarily.

Algorithm 9 Span-Constrained Optimization (ScOpt)

Input: Operators $T_c : \mathbb{R}^S \mapsto \mathbb{R}^S$ and $G_c : \mathbb{R}^S \mapsto \Pi^{\mathrm{SR}}$, accuracy $\varepsilon \in]0, +\infty[$, arbitrary reference state $\bar{s} \in \mathcal{S}$, initial vector $v_0 \in \mathbb{R}^S$, contractive factor $\gamma \in [0, 1[$
Output: Gain $g \in [0, r_{\max}]$, bias $h \in \mathbb{R}^S$, stationary policy $\pi \in \Pi^{\mathrm{SR}}$

Initialize $n = 0$
$v_1 := T_c v_0$
while $sp\left(v_{n+1} - v_n\right) + \frac{2\gamma^n}{1-\gamma} sp\left(v_1 - v_0\right) > \varepsilon$ **do** ▷ Loop until termination
    Increment $n \leftarrow n + 1$
    Shift $v_n \leftarrow v_n - v_n(\bar{s})e$ ▷ Avoids numerical instability ($v_n \not\to +\infty$)
    Compute $(v_{n+1},\ d_n) := (T_c v_n,\ G_c v_n)$ ▷ Alg. 8
end while
Set $g := \frac{1}{2}\Big(\max\{v_{n+1} - v_n\} + \min\{v_{n+1} - v_n\}\Big)$, $h := v_n$ and $\pi := (d_n)^\infty$

5.4.2 Convergence and Optimality Guarantees

In order to derive convergence and optimality guarantees for ScOpt we need to analyze the properties of operator T_c. We start by proving that T_c preserves the one-step *span contraction* property of L. Note that in general L is not a contractive operator (in span semi-norm). In the special case where the MDP is *unichain* and *aperiodic*, L is a J-stage contraction with $S \geq J \geq 1$ (Puterman, 1994, Theorem 8.5.2). In Asm. 5.1 we assume that $J = 1$.

Assumption 5.1

The optimal Bellman operator L is a 1-step γ-span-contraction, i.e., there exists a $\gamma < 1$ such that for any vectors $u, v \in \mathbb{R}^S$, $sp\left(Lu - Lv\right) \leq \gamma sp\left(u - v\right)$.

Lemma 5.8

Under Asm. 5.1, T_c is a γ-span contraction.

Proof. Since Γ_c is non-expansive (property *(b)* in Lem. 5.5) and L is γ-contractive, the result follows by composition. ■

As a consequence of Lem. 5.8 and the *Banach fixed point theorem*, T_c admits a unique fixed point in the *quotient space* induced by the span semi-norm on $\mathbb{R}^S$. In $\mathbb{R}^S$, the fixed point equation has the same form as the Bellman optimality equation satisfied by L (see Prop. 2.4), with an associated gain (unique) and bias (unique up to a constant shift). Moreover, ScOpt converges to the fixed point of this equation and we also show that the associated "gain" is an upper-bound on the solution of (5.6) (due to the monotonicity property of T_c, see property

(a) of Lem. 5.7). We formally state these results in Lem. 5.9.

Lemma 5.9

Under Asm. 5.1, the following properties hold:

1. *Optimality equation and uniqueness: There exists a solution* $(g^+, h^+) \in \mathbb{R} \times \mathbb{R}^S$ *to the optimality equation*

$$T_c h^+ = h^+ + g^+ e. \tag{5.11}$$

If $(g, h) \in \mathbb{R} \times \mathbb{R}^S$ *is another solution of (5.11), then* $g = g^+$ *and there exists* $\lambda \in \mathbb{R}$ *s.t.* $h = h^+ + \lambda e$.

2. *Convergence: For any initial vector* $v_0 \in \mathbb{R}^S$, *the sequence* (v_n) *generated by* SCOPT *converges to a solution vector* h^+ *of the optimality equation (5.11), and*

$$\lim_{n \to +\infty} T_c^{n+1} v_0 - T_c^n v_0 = g^+ e.$$

3. *Dominance: If there exists a scalar* g *and a vector* $h \in \mathbb{R}^S$ *such that* $T_c h \geq h + ge$ *then* $g^+ \geq g$. *As a consequence, the gain* g^+ *is an upper-bound on the supremum of (5.6), i.e.,* $g^+ \geq g_c^*$.

Proof. The formal proof can be found in (Fruit et al., 2018b, Appendix D.3). ■

Point 3 of Lem. 5.9 is the analogue of Prop.3.3 stated in Sec. 3.2. Prop.3.3 was a key step in the proof of *optimism* for UCRLB. Lem. 5.9 will play a similar role for SCAL. A direct consequence of point 2 of Lem. 5.9 (convergence) is that SCOPT always stops after a finite number of iterations. Nonetheless, T_c may not always be globally feasible at h^+ (Fruit et al., 2018b, Appendix B) and thus there may not exist a policy associated to optimality equation (5.11). Furthermore, even when there is one, Lem. 5.9 provides no guarantee on the performance of the policy returned by SCOPT after a finite number of iterations. To overcome these limitations, we introduce an additional assumption, which leads to stronger performance guarantees for SCOPT.

Assumption 5.2

Operator T_c *is globally feasible at any vector* $v \in \mathbb{R}^S$ *such that* $sp(v) \leq c$.

Theorem 5.1

Assume Asm. 5.1 and 5.2 hold and let γ *denote the contractive factor of* T_c *(Asm. 5.1). For any* $v_0 \in \mathbb{R}^S$ *such that* $sp(v_0) \leq c$, *any* $\overline{s} \in \mathcal{S}$ *and any* $\varepsilon > 0$, *the policy* π_n *output by* SCOPT$(v_0, \overline{s}, \gamma, \varepsilon)$ *is such that* $\|g^+ e - g^{\pi_n}\|_\infty \leq \varepsilon$. *Furthermore, if in addition the policy* $\pi^+ = (G_c h^+)^\infty$ *is unichain,* g^+ *is the solution to optimization problem (5.6) i.e.,* $g^+ = g_c^*$ *and* $\pi^+ \in \Pi_c^*$.

Proof. The proof can be found in (Fruit et al., 2018b, Appendix D.4). ■

The first part of the theorem shows that with the stopping condition used in Alg. 9 (line 3), SCOPT returns an *ε-optimal* policy π_n.

The second part is more subtle. Although it may seem counter-intuitive at first, even though $sp\left(h^+\right) = sp\left(T_c h^+\right) \leq c$ (by definition of T_c), in general when the policy $\pi^+ = (G_c h^+)^\infty$ associated to h^+ is *not unichain*, we might have $sp\left(h^+\right) < sp\left(h^{\pi^+}\right)$. This is because h^{π^+} is not necessarily the unique solution (up to constant shift) to the Bellman evaluation equation associated to π^+ and so it is possible that $sp\left(h^+\right) \neq sp\left(h^{\pi^+}\right)$. Consequently, we cannot guarantee that g^+ is the solution of (5.6) (the constraint $sp\left(h^{\pi^+}\right) \leq c$ should be satisfied). On the other hand, Corollary 8.2.7. of Puterman (1994) ensures that if π^+ is unichain then $sp\left(h^+\right) = sp\left(h^{\pi^+}\right)$, hence $g^+ = g^{\pi^+}$.

Notice that no matter whether π^+ is unichain or not, we cannot guarantee that π_n satisfies the span constraint, i.e., $sp\left(h^{\pi_n}\right)$ may be *arbitrary larger* than c. Nonetheless, the proof of UCRLB only requires to bound the span of a vector h solution to an (approximate) Bellman equation $L_d h \backsimeq h + g$ with $g \geq g^*$ (optimism), no matter whether h matches the definition of bias (Eq. 6.3) for policy $\pi = d^\infty$. Similarly, in the next section we show that the condition $sp\left(h^{\pi_n}\right) \leq c$ is not needed and Thm. 5.1 is sufficient to derive regret bounds when SCOPT is integrated into UCRL2.

5.5 Learning with SCAL

In this section we introduce SCAL, an optimistic online RL algorithm that employs SCOPT to compute policies that efficiently balance exploration and exploitation. We prove that the assumptions stated in Sec. 5.4.2 hold when SCOPT is integrated into the optimistic framework. Finally, we show that SCAL enjoys the same regret guarantees as REGAL.C, while being the first implementable and efficient algorithm to solve bias-span constrained exploration-exploitation.

5.5.1 Learning algorithm

For any extended MDP $\mathcal{M}$ (see Sec. 2.1.5), based on Def. 5.2 we define $\mathcal{T}_c$ as the *span truncation* of the optimal Bellman operator $\mathcal{L}$ of $\mathcal{M}$. In the rest of this chapter, we will refer to this operator as the *"span-truncated Bellman operator"*. In particular, we denote by $\mathcal{L}_k$ and $\mathcal{T}_c^k$ the operators associated to $\mathcal{M}_k$. Given the structure of problem (5.3), one might consider applying SCOPT to the extended MDP $\mathcal{M}_k$ (using $\mathcal{T}_c^k$). Unfortunately, in general $\mathcal{L}_k$ does not satisfy Asm. 5.1 and 5.2 and thus $\mathcal{T}_c^k$ may not enjoy the properties of Lem. 5.9 and Thm. 5.1. To overcome this problem, we slightly *modify* $\mathcal{M}_k$ as described in Def. 5.5.

Definition 5.5

Let $\mathcal{M}$ be an extended MDP defined by the confidence intervals $B_r(s,a) = [r(s,a)^-, r(s,a)^+]$ and $B_p(s,a,s') = [p(s'|s,a)^-, p(s'|s,a)^+]$ for all state-action pairs (s,a). Let $1 \geq \eta > 0$ and $\overline{s} \in \mathcal{S}$ an arbitrary "reference" state. We define the "modified" *MDP $\widetilde{\mathcal{M}}$ associated to $\mathcal{M}$ by*

$$\forall (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}, \qquad \widetilde{B}_r(s,a) := \left[0, r(s,a)^+\right], \tag{5.12}$$

$$\widetilde{B}_p(s,a,s') := \begin{cases} B_p(s,a,s') & \text{if } \; s' \neq \overline{s}, \\ B_p(s,a,\overline{s}) \cap [\eta, 1] & \text{otherwise}, \end{cases} \tag{5.13}$$

where we assume that η is small enough so that:

$$B_p(s,a,\overline{s}) \cap [\eta, 1] \neq \emptyset, \quad \text{and} \quad \sum_{s' \in \mathcal{S}} p(s'|s,a)^- \leq 1 \leq \sum_{s' \in \mathcal{S}} p(s'|s,a)^+$$

We denote by $\widetilde{\mathcal{L}}$ the optimal Bellman operator of $\widetilde{\mathcal{M}}$ and by $\widetilde{\mathcal{T}}_c$ the span truncation of $\widetilde{\mathcal{L}}$ (see Def. 5.2).

We will now justify the two *transformations* introduced in Def. 5.5: the *"perturbation"* of the transition probabilities (5.13) as well as the *"augmentation"* of the rewards (5.12)[7].

By slightly *perturbing* the confidence intervals B_p of the transition probabilities, we enforce that the "*attractive*" state $\overline{s}$ is reached with non-zero probability from any state-action pair (s,a). A direct implication is that the *ergodic coefficient* of $\widetilde{\mathcal{M}}$ defined as

$$\gamma := 1 - \min_{\substack{s,x \in \mathcal{S}, \\ a,b \in \mathcal{A} \\ p,q \in \widetilde{B}_p}} \left\{ \sum_{y \in \mathcal{S}} \underbrace{\min\{p(y|s,a), q(y|x,b)\}}_{\geq \eta \text{ if } y = \overline{s}} \right\}$$

is smaller than $1-\eta < 1$, so that $\widetilde{\mathcal{L}}$ is γ-contractive (Puterman, 1994, Thm. 6.6.6). Therefore, Asm. 5.1 holds. Moreover, for any policy $\pi \in \Pi^{\mathrm{SR}}(\widetilde{\mathcal{M}})$, the state $\overline{s}$ necessarily belongs to all *recurrent classes* of π implying that π is unichain. Thus, $\widetilde{\mathcal{M}}$ is a unichain MDP. As we will later show, the η-perturbation of B_p only introduces a *small bias* ηc in the *optimism*. Given that c is known and $\eta > 0$ can be tuned, the magnitude of this bias can be *controlled*.

Let's now ignore the η-perturbation of B_p and focus on the augmentation of B_r. By *augmenting* (without perturbing) the confidence intervals B_r of the rewards, we ensure two useful properties. First of all, the maximal reward $r(s,a)^+$ of $\widetilde{B}_r(s,a)$ is unchanged and so for any vector $v \in \mathbb{R}^S$, $\widetilde{\mathcal{L}} v = \mathcal{L} v$ and thus $\widetilde{\mathcal{T}}_c v = \mathcal{T}_c v$ (by definition of $\mathcal{T}_c$). Secondly, let $d \in D^{\mathrm{MD}}(\widetilde{\mathcal{M}})$ be any (Markov deterministic) decision rule such that $\forall s \in \mathcal{S}$, $\widetilde{r}(s, d(s)) = 0$ (such a decision rule always exists given the definition of $\widetilde{B}_r(s,a)$ in Eq. 5.12). We denote by $\widetilde{\mathcal{L}}_d$ the Bellman evaluation operator of decision rule d in the extended MDP $\widetilde{\mathcal{M}}$ (see Eq. 2.4:

[7] It is immediate to see that $\widetilde{B}_r(s,a) \subseteq B_r(s,a)$, hence the name "augmentation".

$\widetilde{\mathcal{L}}_d v := \widetilde{r}_d + \widetilde{P}_d v$ for all $v \in \mathbb{R}^S$). Since the reward associated to d is 0 in all states, we have $sp\left(\widetilde{\mathcal{L}}_d v\right) = sp\left(\widetilde{P}_d v\right) \leq sp\,(v)$ (the last inequality is a direct consequence of Proposition 6.6.1 of Puterman, 1994). Therefore, if $sp\,(v) \leq c$ then $sp\left(\widetilde{\mathcal{L}}_d v\right) \leq c$ meaning that $d \in \widetilde{D}(c, v) \neq \emptyset$ (where $\widetilde{D}(c, v) \neq \emptyset$ is defined in Lem. 5.6). By Lem. 5.6, $\widetilde{D}(c, v) \neq \emptyset$ implies that $\widetilde{\mathcal{T}_c}$ is globally feasible at v. To summarize, for all $v \in \mathbb{R}^S$ satisfying $sp\,(v) \leq c$, $\widetilde{\mathcal{T}_c}$ is globally feasible at v. This matches the statement of Asm. 5.2.

When combining both the perturbation of B_p and the augmentation of B_r, both Asm. 5.1 and 5.2 hold and we obtain Thm. 5.2 (see Fruit et al., 2018b, Theorem 11).

Theorem 5.2

Let $\mathcal{M}$ be an extended MDP and $\widetilde{\mathcal{M}}$ its "modified" counterpart with perturbation $\eta \geq 0$ (see Def. 5.5). Then

1. *$\widetilde{\mathcal{L}}$ is a γ-span contraction with $\gamma \leq 1-\eta < 1$ (i.e., Asm. 5.1 holds) and thus Lem. 5.9 applies to $\widetilde{\mathcal{T}_c}$. We denote by (g^+, h^+) a solution to equation (5.11) for $\widetilde{\mathcal{T}_c}$.*
2. *$\widetilde{\mathcal{T}_c}$ is globally feasible at any $v \in \mathbb{R}^S$ satisfying $sp\,(v) \leq c$ (i.e., Asm. 5.2 holds) and $\widetilde{\mathcal{M}}$ is unichain implying that $\pi^+ = (G_c h^+)^\infty$ is unichain.*
   Thus Thm. 5.1 applies to $\widetilde{\mathcal{T}_c}$.

Proof. The proof can be found in (Fruit et al., 2018b, Appendix E). ■

SCAL is a variant of UCRLB that applies ScOpt (see Alg. 9) instead of EVI on the extended MDP $\widetilde{\mathcal{M}}_k$ obtained by modifying $\mathcal{M}_k$ (see Def. 5.5) in each episode k in order to solve the optimization problem[8]

$$\max_{M \in \widetilde{\mathcal{M}}_k,\ \pi \in \Pi_c(M)} g_M^\pi := g_k^+ \tag{5.14}$$

where the maximum always exists (Thm. 5.2 applies to $\widetilde{\mathcal{M}}_k$). The maximizing policy is denoted π_k^+. The intervals $\widetilde{B}_p^k$ of $\widetilde{\mathcal{M}}_k$ are constructed using parameter[9] $\eta_k = r_{\max}/(c \cdot t_k)$ and an arbitrary attractive state $\overline{s} \in \mathcal{S}$. ScOpt is then run with an initial value function $v_0 = 0$, the same reference state $\overline{s}$ used for the construction of $\widetilde{B}_p^k$, contraction factor $\gamma_k = 1 - \eta_k$, and accuracy $\varepsilon_k = r_{\max}/t_k$. ScOpt finally returns a policy which is executed until the end of episode k.

More precisely, SCAL implements Alg. 5 (UCRLB) with the difference that $\mathcal{M}_k$ should be replaced by $\widetilde{\mathcal{M}}_k$ in line 5 (see Def. 5.5 for how to compute $\widetilde{\mathcal{M}}_k$ based on $\mathcal{M}_k$). Also, line 9 (Eq. 3.5) should be replaced by:

$$(g_k, h_k, \pi_k) := \text{ScOpt}\left(\widetilde{\mathcal{T}_c^k}, \widetilde{\mathcal{G}_c^k}, \frac{r_{\max}}{t_k}, s_1, 0, \gamma_k\right). \tag{5.15}$$

The rest of Alg. 5 is unchanged. Note that in theory, the aperiodicity transformation

[8] This optimization problem is a specific instance of (5.3) in Sec. 5.2 with $\mathcal{M} \leftarrow \widetilde{\mathcal{M}}_k$.

[9] Notice that given that $\beta_{p,k}^{sas'} \geq \eta_k$ for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, the assumptions of Def. 5.5 hold.

is useless in SCOPT because the η-perturbation of B_p^k already ensures *aperiodicity* of $\widetilde{\mathcal{M}}_k$. In our experiments, η is set to 0 since SCOPT still converges (see Sec. 5.6). In that case, it may be useful to integrate the aperiodicity transformation into SCOPT. The aperiodicity transformation affects both $\widetilde{\mathcal{L}}_k$ and the truncation Γ_c since the constraint c should be replaced by $c/(1-\alpha)$ as a consequence of the following theorem.

Theorem 5.3

Let M be an MDP and M_α the MDP obtained after aperiodicity transformation of parameter α. For any (stationary) policy $\pi \in \Pi^{SR}$: $h^\pi_{M_\alpha} = 1/(1-\alpha)h^\pi_M$ where $h^\pi_{M_\alpha}$ and h^π_M are the bias associated to policy π in M and M_α respectively. In particular, $h^*_{M_\alpha} = 1/(1-\alpha)h^*_M$ and so $sp\left(h^*_{M_\alpha}\right) = 1/(1-\alpha)sp\left(h^*_M\right)$.

Proof. See App. C.2. ■

5.5.2 Analysis of SCAL

Gain optimism

Thm. 5.2 only guarantees gain-optimism (i.e., $g_k^+ \geq g^*$) when $M \in \widetilde{\mathcal{M}}_k$. Unfortunately, although $M \in \mathcal{M}_k$ with high probability by construction (see Thm. 3.1), this may no longer be true for $\widetilde{\mathcal{M}}_k$ due to the η_k-perturbation of B_p^k. Since the "inclusion argument" seem to fail here, we will use the new proof technique introduced in Sec. 3.2.1 that relies on the "dominance property" of $\mathcal{L}_k$ (we will need to use the dominance property of $\widetilde{\mathcal{T}_c^k}$ instead). As discussed in Sec. 3.2, a direct consequence of Thm. 3.1 is that with probability at least $1-\frac{\delta}{3}$:

$$\forall k \geq 1, \quad \mathcal{L}_k h^* \geq Lh^* = h^* + g^* e.$$

where we recall that g^* and h^* respectively denote the optimal gain and bias of the true (unknown) MDP M. In Chap. 3 we argued that this simple inequality and the "dominance property" of Prop. 3.3 are sufficient to show that UCRLB is gain-optimistic. We proceed similarly for SCAL.

By assumption $sp\left(Lh^*\right) = sp\left(h^*\right) \leq c$ implying that $\Gamma_c(Lh^*) = Lh^*$ by definition of Γ_c (see Sec. 5.4.1). Using the monotonicity property of Γ_c (property *(a)* in Lem. 5.5) we deduce that with probability at least $1-\frac{\delta}{3}$:

$$\forall k \geq 1, \quad \mathcal{T}_c^k h^* = \Gamma_c(\mathcal{L}_k h^*) \geq \Gamma_c(Lh^*) = Lh^* = h^* + g^* e \tag{5.16}$$

The idea is to now use point 3 of Lem. 5.9 ("dominance property") in order to prove *optimism*. The problem is that SCOPT uses $\widetilde{\mathcal{T}_c^k}$ instead of $\mathcal{T}_c^k$ to compute policy π_k. The following lemma

shows that the two operators give similar results up to a small bias of order $\eta_k \cdot c$.

Lemma 5.10

Let $\mathcal{M}$ be an extended MDP and $\widetilde{\mathcal{M}}$ its "modified" counterpart with perturbation $\eta \geq 0$ (see Def. 5.5). Denote by $\mathcal{T}_c$ and $\widetilde{\mathcal{T}}_c$ the span-truncated Bellman operators of $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ respectively (see Def. 5.2). For any vector $h \in \mathbb{R}^S$:

$$\left\| \mathcal{T}_c h - \widetilde{\mathcal{T}}_c h \right\|_\infty \leq \eta \cdot sp\,(h) \tag{5.17}$$

Proof. See (Fruit et al., 2018b, Lemma 19, Appendix E). ■

When the transition probabilities are perturbed by η, the application of $\widetilde{\mathcal{T}}_c$ on h results in a perturbation of η amplified by $sp\,(h)$ i.e., $\eta \cdot sp\,(h)$.

As a direct consequence of Lem. 5.10 and Eq. 5.16 and the assumption that $sp\,(h^*) \leq c$, with probability at least $1 - \frac{\delta}{3}$:

$$\forall k \geq 1, \quad \widetilde{\mathcal{T}}_c^k h^* \geq h^* + (g^* - \eta_k \cdot c)\, e \quad \text{and so} \quad g_k^+ \geq g^* - \eta_k \cdot c \;= g^* - \frac{r_{\max}}{t_k} \tag{5.18}$$

where the second inequality is a direct application of the dominance property proved in Lem. 5.9. SCAL is therefore approximately gain-optimistic. As shown in Chap. 3, the term $r_{\max}/t_k$ only has a negligible impact on the regret (negligible logarithmic term).

Bound on the range of the optimistic bias

Due to Thm. 5.2, (g_k, h_k) (see Eq. 5.15) satisfies an approximate Bellman equation (similar to (3.22) for UCRLB) i.e.,

$$\left\| \widetilde{\mathcal{T}}_c^k h_k - h_k - g_k e \right\|_\infty \leq \frac{r_{\max}}{t_k}. \tag{5.19}$$

Thm. 5.2 also shows that $\widetilde{\mathcal{T}}_c^k$ is globally feasible at h_k implying that $\widetilde{\mathcal{T}}_c^k h_k = \widetilde{\mathcal{L}}_k^{d_k} h_k$ with $\pi_k = (d_k)^\infty$. Finally, $sp\,(h_k) \leq c$ since either $h_k = v_0 = 0$ or there exists $v \in \mathbb{R}^S$ such that $h_k = \widetilde{\mathcal{T}}_c^k v$ (by design of ScOpt).

Regret guarantees

We are now ready to prove two regret bounds for SCAL (as we did for UCRLB).

Theorem 5.4

There exists a numerical constant $\beta > 0$ such that for any weakly communicating MDP satisfying $sp(h^*) \leq c$, with probability at least $1-\delta$, it holds that for all initial state distributions $\mu_1 \in \Delta_S$, and for any time horizon $T > 1$, the regret of SCAL *is bounded as*

$$\begin{aligned}\Delta(\text{SCAL}, T) \leq\ & \beta \cdot \max\{r_{\max}, c\} \sqrt{\left(\sum_{s,a} \Gamma(s,a)\right) T \ln\left(\frac{T}{\delta}\right)} \\ & + \beta \cdot \max\{r_{\max}, c\} S^2 A \ln\left(\frac{T}{\delta}\right) \ln(T)\,.\end{aligned} \tag{5.20}$$

Proof. The proof is identical to the proof of Thm. 3.4 for UCRLB. The only difference is that we bound $sp(h_k)$ by c instead of Λ and a factor 2 appear when using the optimism property since $g_k \geq g^* - 2r_{\max}/t_k$ (η_k-perturbation combined with ε_k-approximation). ■

Theorem 5.5

There exists a numerical constant $\beta > 0$ such that for any weakly communicating communicating MDP satisfying $sp(h^*) \leq c$, with probability at least $1-\delta$, it holds that for all initial state distributions $\mu_1 \in \Delta_S$ and for all time horizons $T > 1$, the regret of SCAL *is bounded as*

$$\begin{aligned}\Delta(\text{UCRLB}, T) \leq\ & \beta \cdot \max\{r_{\max}, \sqrt{r_{\max} c}\} \sqrt{\left(\sum_{s,a} \Gamma(s,a)\right) T \ln\left(\frac{T}{\delta}\right) \ln(T)} \\ & + \beta \cdot \max\left\{r_{\max}, \frac{c^2}{r_{\max}}\right\} S^2 A \ln\left(\frac{T}{\delta}\right) \ln(T)\,.\end{aligned} \tag{5.21}$$

Proof. The proof is identical to the proof of Thm. 3.5 for UCRLB with the same two (minor) differences mentioned in the proof of Thm. 5.4. ■

The previous bound shows that when $c \leq \Lambda$, SCAL scales linearly with c, while UCRLB scales linearly with Λ (all other terms being equal). Notice that the gap between $sp(h^*)$ and Λ can be arbitrarily large, and thus the improvement can be significant in many MDPs. As an extreme case, in weakly communicating MDPs the travel-budget can be infinite, leading UCRLB to suffer linear regret (see Chap. 4), while SCAL is still able to achieve sub-linear regret without requiring the algorithmic modifications presented in Chap. 4 (TUCRL). SCAL is able to *learn* in any weakly-communicating MDP like TUCRL and unlike UCRLB (which is only able to learn in a communicating MDP). However, we conjecture that SCAL (unlike TUCRL) does not suffer from the limitations mentioned in Sec. 4.5 of Chap. 4 i.e., while the regret of TUCRL will always grow as $\sqrt{T}$ when the true MDP is *not communicating*, the regret of SCAL eventually grows *logarithmically with T*. SCAL is able to exploit additional

prior knowledge about $sp(h^*)$ that TUCRL does not have. Since TUCRL is solving a more difficult problem, it is reasonable to expect the algorithm to perform worse than SCAL (at least asymptotically). More precisely, we make this conjecture for two reasons. The first reason is that it seems straightforward to extend the proof of Jaksch et al. (Theorem 4 2010) to SCAL (and UCRLB). We recall that this theorem shows that the regret of UCRL2 eventually grows logarithmically with T (for T big enough) and not as $\sqrt{T}$. We keep the formal proof of this conjecture for future work. The second reason is that the experiments presented in the next section tend to validate the conjecture.

When $c > \Lambda$, due to the η_k-perturbation pf B_p^k, it seems not trivial to relate the span of h_k with Λ (unlike in the case of UCRLB, see Sec. 3.3). Nevertheless, we can *slightly modify* SCAL to address this issue: at the beginning of any episode k, we run both ScOpt (with the same inputs) and EVI (as in UCRLB) in parallel and pick the policy associated to the optimistic bias with smallest span. With this modification, SCAL enjoys the *best of both worlds*, i.e., the regret scales with $\min\{c, \Lambda\}$ *instead of* c.

When c is *wrongly chosen* ($c < sp(h^*)$), SCAL learns a *span-constrained* optimal policy with an associated gain g_c^* (solution to (5.6)) that can potentially be arbitrary smaller than g^*. In this scenario, the regret is bounded as

$$\widetilde{\mathcal{O}}\left(\sqrt{r_{\max}\min\{c,\Lambda\}\left(\sum_{s,a}\Gamma(s,a)\right)T\ln\left(\frac{T}{\delta}\right)\ln(T)}\right)+(g^*-g_c^*)\cdot T$$

For a given horizon T, there is clearly a trade-off in the choice of c: a big value minimizes the linear term $(g^* - g_c^*) \cdot T$ but increases the $\sqrt{T}$-term, and conversely. The best way to choose c depends on the amount of prior knowledge about the true MDP.

To conclude this section, we emphasize that the benefit of SCAL over UCRL2 comes at a *negligible additional computational cost* (EVI and ScOpt have comparable time and space complexities).

5.6 Numerical Experiments

In this section, we numerically validate our theoretical findings. In particular, we show that the regret of UCRLB indeed scales with the travel-budget, while SCAL achieves much smaller regret that only depends on the span. This result is even more extreme in the case of non-communicating MDPs, where $\Lambda = +\infty$.

5.6.1 Toy MDP

Consider the simple but descriptive three-state domain shown in Fig. 4.5 (Chap. 4) where instead of being deterministic, all rewards are Bernouilli random variables (with the same means). This small change slightly increases the complexity of the problem. The optimal policy π^* is such that $\pi^*(s_2) = a_1$ with gain $g^* = \frac{2}{3}$ and bias $h^* = \left[\frac{-2-\delta}{3(1-\delta)}, \frac{-1}{1-\delta}, 0\right]$. If δ is

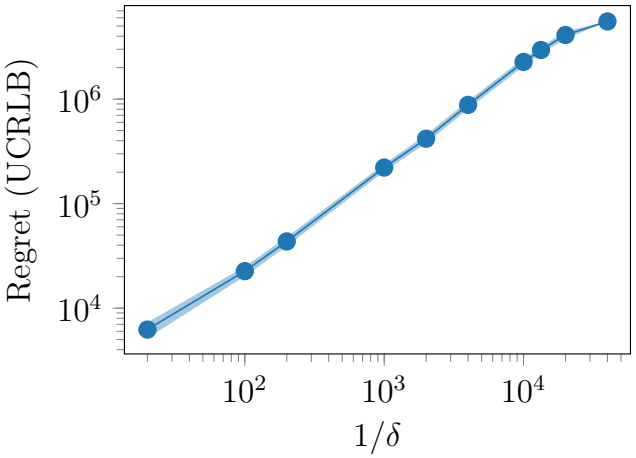
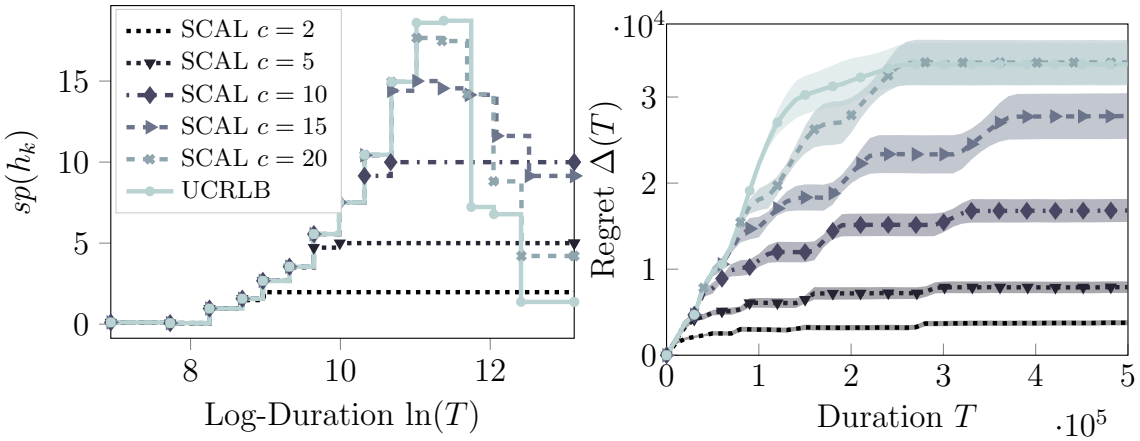


Figure 5.3: Cumulative regret incurred by UCRLB after $T = 2.5 \cdot 10^7$ steps as a function of $1/\delta \propto \Lambda$ (averaged over 20 runs).

small, $sp(h^*) = \frac{1}{1-\delta} \approx 1$, while $\Lambda \propto \frac{1}{\delta}$. Fig. 5.3 shows that, as predicted by theory, the regret of UCRLB (for a fixed horizon T) grows with $\frac{1}{\delta} \approx \Lambda$. The optimal bias span however is roughly equal to 1. Therefore, we expect SCAL to clearly outperform UCRLB on this example. In all the experiments, we noticed that perturbing the extended MDP was not necessary to ensure convergence of ScOpt and so we set $\eta_k = 0$. We also set $\gamma_k = 0$ to speed-up the execution of ScOpt (see stopping condition in Alg. 9).

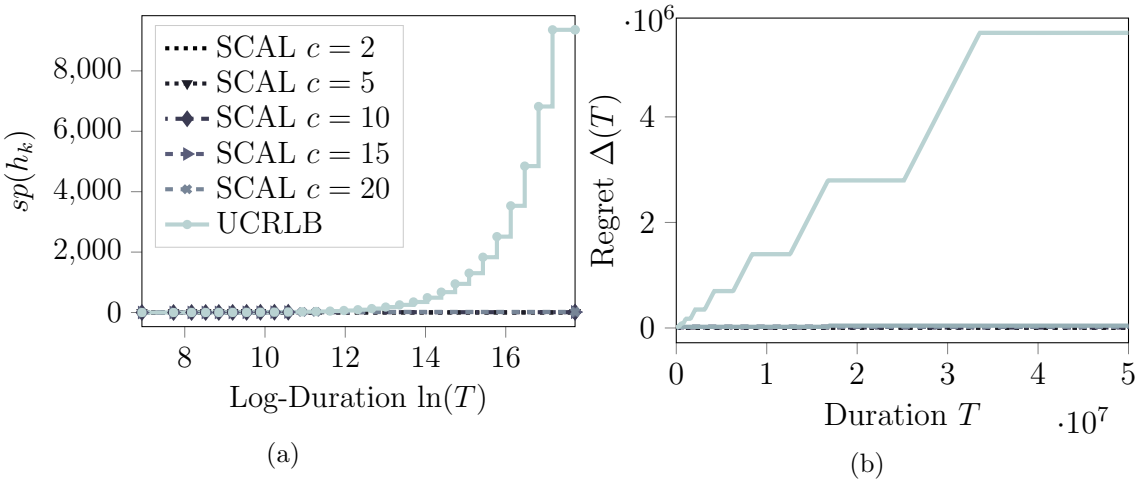
Communicating MDPs. We first set $\delta = 0.005 > 0$, giving a communicating MDP (Fig. 5.4). With such a small δ, visiting state s_1 is rather unlikely. Nonetheless, UCRLB keeps trying to visit s_1 (i.e., play a_0 in s_2) until it collects enough samples to understand that s_1 is actually a *bad* state (before that, UCRLB *"optimistically"* assumes that s_1 is a *highly rewarding* state). Therefore, UCRLB plays a_0 in s_2 for a long time and suffers large regret. This problem is particularly challenging for any learning algorithm solely employing *optimism* like UCRLB (cf. (Ortner, 2008) for a more detailed discussion on the intrinsic limitations of optimism in RL). In contrast, SCAL is able to mitigate this issue when an appropriate constraint c is used. More precisely, whenever s_1 is believed to be the most rewarding state, the value function (bias) is maximal in s_1 and ScOpt applies a "truncation" in that state and "mixes" deterministic actions. In other words, SCAL leverages on the prior knowledge of the optimal bias span to understand that s_1 cannot be as good as predicted (from optimism). The exploration of the MDP is greatly affected as SCAL quickly discovers that action a_0 in s_2 is suboptimal. Therefore, SCAL is always performing better than UCRL (Fig. 5.4b) and the smaller c, the better the regret. Surprisingly the *actual* policy played by SCAL in this particular MDP is always deterministic. ScOpt mixes actions in s_1 where only one *true* action is available but the mixing happens in the *extended* MDP $\widetilde{\mathcal{M}}_k$ where the action set is compact. The policy that ScOpt outputs is thus *stochastic* in the *extended* MDP but *deterministic* in the *true* MDP.

Infinite travel-budget. By selecting $\delta = 0$ (Fig. 5.5) the diameter becomes infinite ($D = +\infty$) but the MDP is still *weakly* communicating (with transient state s_1). UCRLB is not



(a) (b)

Figure 5.4: Results in the three-states domain with $\delta = 0.005$. We report the span of the optimistic bias (Fig. 5.4a) and the cumulative regret (Fig. 5.4b) as a function of T.



(a) (b)

Figure 5.5: Results in the three-states domain with $\delta = 0$. We report the span of the optimistic bias (Fig. 5.5a) and the cumulative regret (Fig. 5.5b) as a function of T.

able to handle this setting and suffers linear regret. On the contrary, SCAL is able to quickly recover the optimal policy (see Fig. 5.5). Note that unlike with TUCRL, the regret of SCAL seems to achieve a logarithmic "plateau" even in the non-communicating case. This may seem paradoxical but actually Thm. 4.2 does not apply in the case where a bound on the optimal bias span is known since the MDPs with sufficiently small ε in Fig. 4.8 (used to prove Thm. 4.2) do not satisfy $sp(h^*) \leq c$. We conjecture that a logarithmic regret bound similar to Thm. 2.37 can be derived for SCAL, SCAL$^+$ and SCAL*, with D replaced by $c/r_{\max}$. This simple example shows the dramatic impact of prior knowledge on the exploration-exploitation performance.

5.6.2 Knight Quest

We now consider a second environment that takes inspiration from classical arcade games. The goal is to rescue a prisoner in the shortest time without being killed by the dragon. To achieve this task, the knight needs to collect gold, buy a key and deliver the prisoner. A

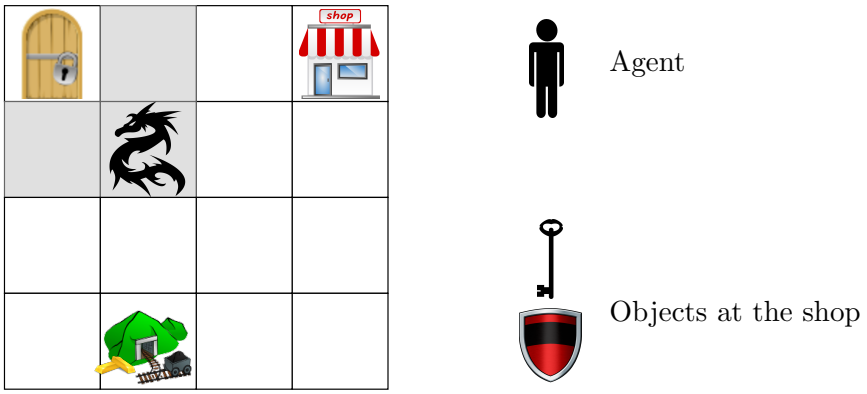


Figure 5.6: Representation of the Knight Quest 4×4 map. The grey shadowed cells are the locations where the dragon can move.

representation of the environment is provided in Fig. 5.6. The elements of the game are: a knight, a prisoner, a dragon patrolling around the prisoner, a gold mine and, a shop, a key and a shield.

Shop, Prisoner and Gold Mine. These elements are special states of the environment. The shop is the place where the knight can buy objects. Every time the knight is killed by the dragon or delivers the prisoner, it restarts from the shop. The prisoner is located behind the locked door in the terminal state. The knight can collect gold at the gold mine.

Dragon. The dragon is the enemy and it is randomly moving around the prisoner's location. Let's denote with $d \in \{0, 1, 2\}$ the position of the dragon such that: $d = 0$ is the bottom left grey cell, $d = 1$ is the bottom right grey cell and $d = 2$ the top grey cell. The transition probabilities of the dragon are:

$$p(\cdot|0) = [0.4, 0, 0.6]^\intercal; \quad p(\cdot|1) = [0, 0.4, 0.6]^\intercal; \quad p(\cdot|2) = [0.4, 0.2, 0.4]^\intercal.$$

The dragon kills the knight when they are both at the same position and the knight does not have the shield.

Knight. The knight is the only player of the game. He or she moves in the environment using the four cardinal actions (i.e.,, *right*, *down*, *left* and *up*) plus an action to keep the current position (*stay*). We refer to these 5 actions as *movement actions*. Additionally, the knight can collect the gold (action *CG*), buy a key (action *BK*) or buy a shield (action *BS*).

State representation, actions and reward. A state of the game is represented by the following elements:

- Knight position: coordinates of the grid (row, col), $row, col \in 0, 1, 2, 3$;
- Gold level: the amount of gold owned by the knight, $g \in \{0, 1\}$;

- Dragon position: $d \in \{0, 1, 2\}$;
- Object identifier: object(s) carried by the knight, $o = \{0, 1, 2, 3\}$ where $0 \Leftrightarrow$ nothing, $1 \Leftrightarrow$ key, $2 \Leftrightarrow$ armour and $3 \Leftrightarrow$ key and armour.

We can now can explain the effects of actions, i.e., how the next state is generated. The movement actions have the trivial effect of changing the knight position. The action CG changes the state only when the knight is at the mine. In this case the level of gold is incremented by one, formally, $g \leftarrow \min\{1, g + 1\}$. Actions BK and BA alter the state only when executed in the shop with gold-level equal to 1. All the actions are deterministic when the knight does not carry the shield. When the knight carries the shield, he or she cannot be killed by the dragon (i.e., knight and dragon can occupy the same cell). However, due to the weight of the armour, the knight's gait is unsteady and other tasks are more challenging i.e.,

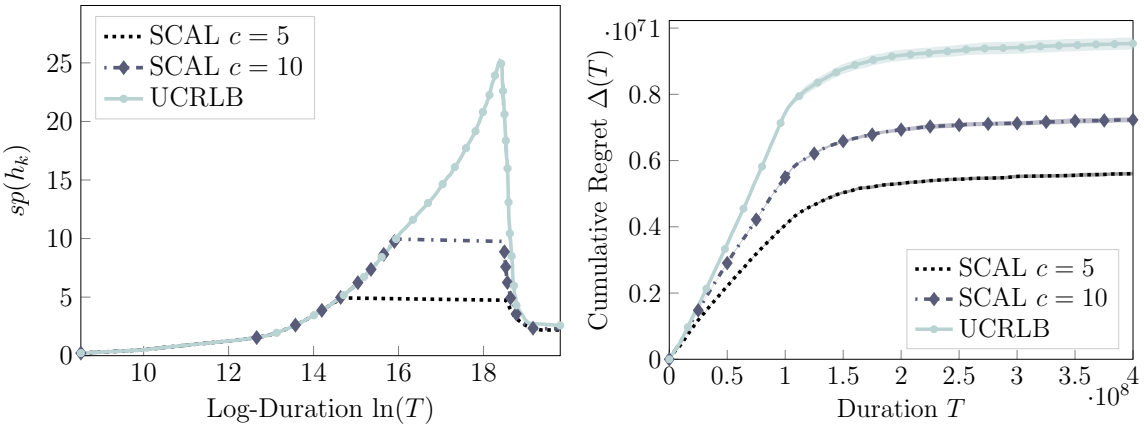
- the cardinal actions result in a normal (correct) transition with probability 0.5, otherwise the current position is kept,
- CG fails with probability 0.99, i.e., with probability 0.01 the gold level is incremented,
- actions BK and BS are not modified.

The basic reward signal is -1 at each time step. The knight also receives a reward of -10 when he or she executes CG, BK or BA outside the designed location (i.e., mine and shop). Finally, he or she obtains a reward of 20 when reaching the prisoner with the key and -20 when killed by the dragon. For the experiments, we rescaled the reward to lie in $[0, 1]$.

Features of the game. The state and action space size are $S = 360$ and $A = 8$, while the travel-budget of the MDP is $\Lambda \approx 130$. The associated shortest path starts from the shop with the shield and no gold, and eventually delivers the prisoner with one unit of gold and the key. In contrast, the optimal strategy simply consists in collecting gold, buying the key and rescueing the prisoner (there is no need to buy the shield as the dragon can be bypassed). We have: $g^* \approx 0.5$, $sp\,(h^*) \approx 3.28$.

This game is challenging since the worst shortest path (achieving the travel-budget) is "orthogonal" to the optimal policy (achieving optimal gain). Many common real-world RL tasks appear to share this property: the agent can face several choices (actions) and most of them are useless. The span constraint c cansomehow be interpreted as a prior on the level of difficulty of the game.

Results. We run UCRLB and SCAL over an horizon $T = 4 \cdot 10^8$, with different priors c. As in the toy example, SCAL is run with the augmented reward but no perturbation of the transition matrix ($\eta_k = 0$), and γ_k is set to 0. Results are reported in Fig. 5.7. We can notice that SCAL is able to outperform UCRL2 by a big margin. This is because unlike UCRLB, SCAL can leverage the knowledge of c to better direct the exploration.



(a) (b)

Figure 5.7: Behaviour of UCRLB and SCAL in the knight quest game. Figures show the span of the optimistic bias (Fig. 5.7a) and the cumulative regret (Fig. 5.7b) as a function of T. Results are averaged over 15 runs and 95% confidence intervals of the mean are shown for the regret.

5.7 SCAL$^+$: SCAL with exploration bonus

In this section, we introduce SCAL$^+$, an online RL algorithm that leverages an *exploration bonus* to achieve *near-optimal* regret guarantees. Similar to SCAL, SCAL$^+$ takes as input an upper-bound c on the optimal bias span (i.e., $sp(h^*) \leq c$) to constrain the planning problem solved over time. The crucial difference with SCAL is that SCAL$^+$ does not require planning with an *extended* Bellman operator, but it directly computes the optimal policy of the *estimated* Bellman operator, where the reward is increased by an *exploration bonus*. As proved in Sec. 5.7.2 the bonus is carefully tuned so as to guarantee optimism and small regret at the same time (Thm. 5.6).

5.7.1 The algorithm

The pseudo-code of SCAL$^+$ is reported in Alg. 10. Similarly to SCAL and UCRLB, SCAL$^+$ proceeds in episodes (indexed by k). At the beginning of each episode k, SCAL$^+$ constructs an *estimated* MDP $M_k = (\mathcal{S}, \mathcal{A} \times \{0,1\}, p_k, r_k)$ (line 5 of Alg. 10). Unlike the extended MDP used in SCAL, M_k has a *finite* action space. The maximum likelihood estimator would be the natural choice to define the transition probabilities and rewards of M_k i.e., $p_k \leftarrow \widehat{p}_k$ and $r_k \leftarrow \widehat{r}_k$. Unfortunately, this choice does not guarantee that the optimal gain g_k^* of M_k is *bigger or equal* than the optimal gain of the true unknown MDP g^*. To ensure gain-optimism (see Lem. 5.12), we increase the reward by an exploration bonus b_k (Eq. 5.22) i.e., we define $r_k \leftarrow \widehat{r}_k + b_k$ (Eq. 5.25). Intuitively, the exploration bonus is *large* for *poorly visited state-action pairs*, while it *decreases* as the *number of visits increases*. A crucial aspect in the formulation of b_k is that it scales with the bound on the bias span $c \geq sp(h^*)$. In fact, the exploration bonus is tailored to guarantee the dominance property $L_k h^* \geq L h^*$ holds with high probability, where L_k is the optimal Bellman operator of M_k. Therefore b_k is not

Algorithm 10 SCAL$^+$ (SCAL with exploration bonus)

Input: Confidence $\delta \in]0,1[$, maximal reward $r_{\max}$, set of states $\mathcal{S}$, set of actions $\mathcal{A}$, positive scalar $c \geq 0$

Set initial time $t := 1$, observe initial state s_1 and initialize for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$:
- counters $N_1(s,a,s') := 0$ and $N_1(s,a) := 0$,
- empirical averages $\widehat{p}_1(s'|s,a) := 0$ and $\widehat{r}_1(s,a) := 0$,

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

Set the starting time of the episode $t_k := t$ and initialize for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$: episode counters $\nu_k(s,a,s') := 0$ and $\nu_k(s,a) := 0$, and cumulative rewards $R_k(s,a) := 0$.

For all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, compute *exploration bonus*:

$$b_k(s,a) := \; c \cdot \min\left\{\beta_k^{sa} + \frac{1}{N_k(s,a)+1}, 2\right\} + r_{\max} \cdot \min\left\{\beta_k^{sa}, 1\right\} \tag{5.22}$$

$$\text{with} \quad \beta_k^{sa} := \sqrt{\frac{1}{N_k^+(s,a)} \ln\left(\frac{20SAN_k^+(s,a)}{\delta}\right)} \tag{5.23}$$

Set $M_k := \{\mathcal{S}, \mathcal{A} \times \{0,1\}, r_k, p_k\}$ to be the *"augmented"* and *"perturbed" estimated* MDP defined by

$$p_k(s'|s,a_i) := \frac{N_k(s,a)\widehat{p}_k(s'|s,a)}{N_k(s,a)+1} + \frac{\mathbb{1}(s'=s_1)}{N_k(s,a)+1}, \tag{5.24}$$

$$r_k(s,a_i) := (\widehat{r}_k(s,a) + b_k(s,a)) \cdot \mathbb{1}(i=1) \tag{5.25}$$

for all $s \in \mathcal{S}$, $a_i = (a,i) \in \mathcal{A} \times \{0,1\}$.

Compute policy π_k using ScOpt (see Alg. 9):

$$(g_k, h_k, \pi_k) := \textsc{ScOpt}\left(L_k, G_k, \frac{r_{\max}}{t_k}, s_1, 0, \frac{1}{t_k+1}\right) \tag{5.26}$$

Sample action $a_t \sim \pi_k(\cdot|s_t)$.

while True do ▷ Execute policy π_k until the end of episode k

Execute action a_t, obtain reward r_t, and observe next state s_{t+1}.

Increment episode counters:

$\nu_k(s_t,a_t,s_{t+1}) \leftarrow \nu_k(s_t,a_t,s_{t+1}) + 1$ and $\nu_k(s_t,a_t) \leftarrow \nu_k(s_t,a_t) + 1$

Increment cumulative reward $R_k(s_t,a_t) \leftarrow R_k(s_t,a_t) + r_t$

if $\nu_k(s_t,a_t) \geq N_k^+(s_t,a_t)$ **then** ▷ Stopping condition of episode k

Increment time $t \leftarrow t+1$ and **Break**

else

Increment time $t \leftarrow t+1$ and set action $a_t \sim \pi_k(\cdot|s_t)$.

end if

end while

Update counters, empirical averages and sample variances for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$:

$$N_{k+1}(s,a,s') := N_k(s,a,s') + \nu_k(s,a,s') \text{ and } N_{k+1}(s,a) := N_k(s,a) + \nu_k(s,a) \tag{5.27}$$

$$\widehat{p}_{k+1}(s'|s,a) := \frac{N_k(s,a)}{N_{k+1}^+(s,a)} \cdot \widehat{p}_k(s'|s,a) + \frac{\nu_k(s,a,s')}{N_{k+1}^+(s,a)} \tag{5.28}$$

$$\widehat{r}_{k+1}(s,a) := \frac{N_k(s,a)}{N_{k+1}^+(s,a)} \cdot \widehat{r}_k(s,a) + \frac{R_k(s,a)}{N_{k+1}^+(s,a)} \tag{5.29}$$

end for

just designed as an upper-confidence bound on the reward[10], but it is designed to take into consideration how estimation errors on both p and r may *propagate* to the bias function through application of the Bellman operator. As the constant c provides *prior knowledge* about the span of the optimal bias vector, the exploration bonus is obtained by considering that *"local"* estimation errors may be *amplified up to* a factor c.

The planning problem. To further exploit the prior knowledge $sp\,(h^*) \leq c$ we would like to solve the optimization problem

$$g_c^*(M_k) := \sup_{\pi \in \Pi_c(M_k)} \{g_{M_k}^\pi\}, \tag{5.30}$$

which is an instance of problem (5.6). We recall that SCAL also requires solving an instance of (5.6) but on an extended MDP (see problem (5.5)). We extensively studied (5.6) in Sec. 5.4 and derived ScOpt to solve the problem. While in general ScOpt may *fail* to converge or may return a value function whose associated greedy policy is not a solution to the original optimization problem (Fruit et al., 2018b, Appendix B), we provided a series of *sufficient conditions* on the MDP for which *convergence* and *optimality* properties are recovered (see Sec. 5.4.2). We follow the same approach as in Sec. 5.5.1 and design M_k so as to *enforce* these sufficient conditions (as we did with the extended MDP of SCAL).

Instead of defining $p_k \leftarrow \widehat{p}_k$, we slightly *perturb* the transition probability to ensure that the ergodic coefficient γ_k of M_k is strictly less than 1 (see Asm. 5.1). More precisely, we set $p_k \leftarrow \left(1 - \frac{1}{N_k+1}\right)\widehat{p}_k + \frac{1}{N_k+1} e_{s_1}$ (see Eq. 5.24), where e_{s_1} is the vector with zero values everywhere except at the s_1-th coordinate (s_1 is the initial state at the beginning of the learning process). Note that p_k is a *biased* but *asymptotically consistent* estimator of p. While in the extended MDP of SCAL, the perturbations of transition probabilities were the same in all state-action pairs (s, a), here the perturbation depends on $N_k(s, a)$. In this case we cannot directly apply Lem. 5.10 to show that optimism is preserved up to an η-accuracy. However, we can *adjust* the exploration bonus in order to *compensate* for this small bias by adding a term of order c/N_k (see Eq. 5.22). This will only have a minor impact on the final regret (logarithmic term). Finally, since $t_k \geq N_k$, we have $\gamma_k \leq 1 - \frac{1}{t_k+1} < 1$ and so we can give this value as input to ScOpt (see (5.26)).

We also *augment* the rewards by duplicating every action (the action space of M_k is $\mathcal{A} \times \{0, 1\}$). For every $a_i = (a, i) \in \mathcal{A} \times \{0, 1\}$, the reward $r_k(s, a_i)$ is $\widehat{r}_k(s, a) + b_k(s, a)$ for $i = 1$, and 0 for $i = 0$, while the transition probability is unchanged (same for both a_0 and a_1). By construction, there always exists a policy achieving 0 reward in every state in M_k (any policy taking action a_0). Such a policy has zero gain and bias and so according to Lem. 5.6, $\Pi_c(M_k) \neq \emptyset$.

Following similar steps as in Sec. 5.5, we can prove that M_k satisfies all sufficient conditions

[10] In that case, setting $b_k(s, a) = r_{\max}\beta_k^{sa}$ (see Eq. 5.23) would be enough.

for ScOpt to converge and return an approximate solution to (5.30).

Lemma 5.11

The MDP M_k satisfies the following properties:

1. *the optimal Bellman operator L_k is a γ_k-span-contraction with $\gamma_k \leq 1 - \frac{1}{t_k+1} < 1$,*
2. *all policies are unichain,*
3. *the operator $T_c^k := \Gamma_c L_k$ is globally feasible at any vector $v \in \mathbb{R}^S$ such that $sp(v) \leq c$.*

Therefore, Thm. 5.1 holds. In particular, ScOpt *converges and returns a policy π_k (approximately) solving (5.30).*

Proof. See (Qian et al., 2018b, Proposition 2). ■

The policy π_k returned by ScOpt is obtained by projecting the policy $\widetilde{\pi}_k$ obtained in the *augmented* set $\mathcal{A} \times \{0, 1\}$ and it can be "projected" on $\mathcal{A}$ as $\pi_k(s, a) \leftarrow \widetilde{\pi}_k(s, a_1) + \widetilde{\pi}_k(s, a_2)$. The associated greedy operator is denoted G_k.

Comparison to SCAL. While SCAL$^+$ runs (relative) value iteration directly on the MDP M_k, which has a similar structure as the original MDP (finite action space), SCAL runs *extended* value iteration on an *extended* MDP, whose (uncountable) action space is augmented to take into consideration the confidence intervals on rewards and transition probabilities. As a result, at each iteration of ScOpt, SCAL applies the optimal Bellman operator of the extended MDP to the current value vector. This requires to solve SA different *linear programs* to find the optimistic transition probabilities. Using LProba (see Alg. 7), this can be done in at most $\mathcal{O}(S \ln(S) + S^2 A) = \mathcal{O}(S^2 A)$ computations by first sorting the value vector and then applying LProba (which requires $\mathcal{O}(S)$ computations) to all (s, a) pairs. Overall, every iteration of ScOpt requires $\mathcal{O}(S^2 A)$ computations in SCAL. In comparison, in SCAL$^+$, every iteration of ScOpt can also be done in $\mathcal{O}(S^2 A)$ computations. Therefore, even though SCAL$^+$ requires fewer computations at every iteration of ScOpt, the order of magnitude is the same $\mathcal{O}(S^2 A)$. Nevertheless, SCAL$^+$ is conceptually simpler and has a simpler algorithmic structure, which makes it potentially more flexible and easier to generalize to more complex tasks.

5.7.2 Optimistic Exploration Bonus

We now formally show that $g_c^*(M_k)$ (see Eq. 5.30) is upper-bounding g^*. As explained in the previous section, the exploration bonus was tailored to enforce this property. We denote by L_k (resp. T_c^k) the (resp. truncated) Bellman operator of M_k.

Lemma 5.12

With probability at least $1 - \frac{\delta}{5}$, for all $k \geq 1$, $L_k h^* \geq L h^*$ and therefore by monotonicity of Γ_c, $T_c^k h^* \geq L h^*$. If in addition, $sp(h^*) \leq c$, then $g_c^*(M_k) \geq g^*$ as a consequence of property 3. of Lem. 5.9 (dominance of operator T_c).

Tightness of optimism. Although this might not be straightforward from the statement of Lem. 5.12, SCAL$^+$ achieves a *"tighter"* optimism (i.e., is less prone to *over-exploration*) than SCAL. More precisely, $T_c^k h^*$ upper-bounds $T_c h^* = Lh^*$ by a term approximately scaling as $\widetilde{\Theta}\left(\max\{r_{\max}, c\}/\sqrt{N_k(s,a)}\right)$ (corresponding to the exploration bonus). In contrast, the truncated Bellman operator used by SCAL applied to h^* i.e., $\widetilde{\mathcal{T}}_c^k h^*$, is bigger than $T_c h^* = Lh^*$ by approximately $\widetilde{\Theta}\left(\max\{r_{\max}, c\}\sqrt{\Gamma(s,a)/N_k(s,a)}\right)$. The optimism in SCAL$^+$ is therefore tighter by a multiplicative factor $\sqrt{\Gamma}$. Unfortunately, the tighter degree of optimism is not sufficient to *remove* the $\sqrt{\Gamma}$ in the *final regret bound*. In the next section (see proof sketch of Thm. 5.6), we will explain why the $\sqrt{\Gamma}$ cannot be removed with the current analysis.

5.7.3 Regret Analysis of SCAL$^+$

We now prove a regret bound similar to SCAL (Thm. 5.4).

Theorem 5.6

There exists a numerical constant $\beta > 0$ such that for any weakly communicating MDP satisfying $sp(h^*) \leq c$, with probability at least $1-\delta$, it holds that for all initial state distributions $\mu_1 \in \Delta_S$, and for any time horizon $T > 1$, the regret of SCAL$^+$ *is bounded as*

$$\begin{aligned}\Delta(\text{SCAL}^+, T) \leq\ & \beta \cdot \max\{r_{\max}, c\} \sqrt{\left(\sum_{s,a} \Gamma(s,a)\right) T \ln\left(\frac{T}{\delta}\right)} \\ & + \beta \cdot \max\{r_{\max}, c\} S^2 A \ln\left(\frac{T}{\delta}\right) \ln(T)\end{aligned} \tag{5.31}$$

Proof. The detailed proof can be found in (Qian et al., 2018b, Theorem 6, Appendix B). In the following, all inequalities should be interpreted up to minor approximations and in high probability. Let $\nu_k(s,a)$ be the number of visits in (s,a) during episode k and k_T be the total number of episodes before time T. Using Lem. 5.12, we have:

$$\Delta(\text{SCAL}^+, T) \lesssim \sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a) \Big(g_k - \sum_a r(s,a)\pi_k(s,a) \Big) \tag{5.32}$$

where g_k, h_k and π_k are respectively the gain, bias and policy returned by SCOPT (see Eq. 5.26). SCOPT ensures that: $g_k + h_k(s) \simeq \sum_a \pi_k(s,a)\left(r_k(s,a) + p_k(\cdot|s,a)^\intercal h_k\right)$. By plugging this inequality into (5.32) we obtain two terms: $\widehat{r}_k(s,a) - r(s,a) + b_k(s,a)$ and $(\widehat{p}_k(\cdot|s,a) - e_s)^\intercal h_k$ (where e_s is the unit vector with all zeros except at the s-th coordinate). We can then add and subtract the true probability $(p_k(\cdot|s,a) - p(\cdot|s,a))^\intercal h_k + (p(\cdot|s,a) - e_s)^\intercal h_k$. Since $sp(h_k) \leq c$, the second term is of order $\widetilde{O}(c\sqrt{T} + cSA)$ when summed over $\mathcal{S}$, $\mathcal{A}$ and episodes k (martingale difference sequence bounded with Azuma's inequality). On the other hand, the term $(p_k(\cdot|s,a) - p(\cdot|s,a))^\intercal h_k$ represents the error of using p_k in place of p in SCOPT. It is

the dominant term in the regret bound. Since h_k depends on p_k, we cannot apply Hoeffding-Azuma inequality as done in the proof of Lem. 5.12 to prove gain-optimism. Instead, we use Hölder's inequality and bound separately $\|p_k(\cdot|s,a) - p(\cdot|s,a))\|_1 \lesssim \sqrt{\Gamma(s,a)}\beta_k^{sa}$ (see Eq. 5.23) and $sp\,(h_k) \leq c$. This eventually introduce a $\sqrt{\Gamma}$ factor in the final regret bound. It is worth pointing out that Γ only appears due to *statistical fluctuations* that we *cannot control*, and not from the optimism (i.e., exploration bonus) that is *explicitly encoded* in the algorithm. For the reward we have $|r_k(s,a) - r(s,a)| \leq r_{\max}\beta_k^{sa}$. As a consequence, we can approximately write that:

$$\Delta(\mathrm{SCAL}^+, T) \lesssim \sum_{k=1}^{m} \sum_{s,a} \nu_k(s,a)\pi_k(s,a) \Bigg(\underbrace{b_k(s,a)}_{\leq d_k(s,a)} + \underbrace{\left(c\sqrt{\Gamma(s,a)} + r_{\max}\right)\beta_k^{sa} + \frac{c}{(N_k(s,a)+1)}}_{:= d_k(s,a)} \Bigg)$$

The remaining terms can be bounded as in SCAL (and UCRLB). ■

Γ-dependency. Since the optimism in SCAL$^+$ is tighter than in SCAL by a $\sqrt{\Gamma}$-factor, one might have expected to get a regret bound scaling as $c\sqrt{SAT}$ instead of $c\sqrt{S\Gamma AT}$ (as pointed out in Sec.5.7.2), thus matching the lower bound of Jaksch et al. (2010) as for the dependency in S. Unfortunately, such a bound seems difficult to achieve with SCAL$^+$ (and even SCAL) for the reason explained in the proof sketch (correlation between h_k and p_k). We refer to the discussion in Sec. 3.7 for more details on closing the gap between lower and upper bounds. The analysis of SCAL$^+$ suggests that the $\sqrt{\Gamma}$-factor arises due to *unavoidable statistical fluctuations* (and not to gain-optimism). We leave as an open question whether the *current analysis* of SCAL$^+$ could be *refined* or whether a *bigger lower bound* should be derived. It is also possible that a $c\sqrt{SAT}$ regret bound can only be achieved with a different algorithm.

c-dependency. The regret bound of SCAL$^+$ does not scale with $\min\{\Lambda, c\}$ like SCAL (when SCAL is modified as explained in Sec. 5.5.2). The difference resides in the fact SCAL builds an extended MDP with *Bellman shortest path operator* (see Sec. 3.3) upper-bounding the Bellman shortest path operator of the true unknown MDP. In this case, the fact that $\Lambda_k \leq \Lambda$ (i.e., the "optimistic" travel-budget is bigger than the true travel-budget) is a consequence of Thm. 3.5. Unfortunately, it is not clear how to apply Thm. 3.5 to M_k. In this MDP, the reward is no longer bounded by $r_{\max}$ and the MDP is *not communicating* (unlike the extended MDP $\mathcal{M}_k$) implying that the assumptions of Thm. 3.5 no longer hold. We leave as an open question whether this analysis can be refined.

Finally, it also seems difficult to prove a regret bound analogue to (5.31) for SCAL$^+$ i.e., scaling with $\sqrt{c}$ instead of c (see Thm. 5.5 for SCAL). This is because the *exploration bonus* itself scales *linearly* with c and explicitly appears in the regret bound when introducing the (approximate) Bellman optimality equation of M_k in the equations. We can no longer make appear a sum of variances like in Sec. 3.6.

5.8 SCAL*: SCAL with tighter optimism

In the previous section, we showed that SCAL$^+$ is less prone to over-exploration than SCAL due to a *tighter degree of optimism*. Although this improvement was not reflected in the *final regret bound* due to the presence of higher order terms, one should expect to observe it *empirically*. Unfortunately, it seems that SCAL$^+$ does not achieve the *optimal dependency* in c and Λ. It is therefore challenging to compare SCAL and SCAL$^+$ in general (even empirically) as the $\sqrt{\Gamma}$-advantage in optimism could be alleviated by the worsening in the c-dependency.

In this section, we present SCAL*, a variant of SCAL that achieves the *best of both algorithms* by leveraging insights from SCAL$^+$ to further constrain the confidence intervals used to construct the extended truncated Bellman operator. Moreover, the *computational complexity* of SCAL* is comparable to the one of SCAL (if not better).

5.8.1 Combining the confidence sets of SCAL with the exploration bonus of SCAL$^+$

Intuition

As we recalled in Sec. 5.5.2, the confidence sets used to build the extended MDP $\mathcal{M}_k$ of UCRLB (see Eq. 3.3 and 3.4 in Alg. 5) ensure that with high probability, the dominance property $\mathcal{L}_k h^* \geq h^* + g^* e$ holds for all k. The dominance property is a sufficient condition to guarantee gain-optimism and derive regret guarantees (see Sec. 3.2). By Hoeffding's inequality, we also know that for all all pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$ and with high probability, the inequality $p_k(\cdot|s,a)^\intercal h^* \leq \widehat{p}_k(\cdot|s,a)^\intercal h^* + sp\,(h^*)\,\beta_k^{sa}$ holds for all k, where β_k^{sa} is defined in Eq. 5.23. When $sp\,(h^*) \leq c$ with c known, these inequalities are used to define the exploration bonus $b_k(s,a)$ of SCAL$^+$, and it is also tempting to try to refine the definition of $\mathcal{L}_k$ by adding the constraints $p_k(\cdot|s,a)^\intercal h_k \leq \widehat{p}_k(\cdot|s,a)^\intercal h_k + c\beta_k^{sa}$. The main difficulty is that these constraints involve both p_k and h_k. One idea could be to enforce the constraint $p_k(\cdot|s,a)^\intercal v_n \leq \widehat{p}_k(\cdot|s,a)^\intercal v_n + \beta_k^{sa}$ at every iteration $n \geq 0$ of EVI. For a fixed v_n, the constraint is linear in p_k and so with this additional constraint, the optimization problem $\max_{p \in B_p^k(s,a)} \{p^\intercal v\}$ is still a linear program. Unfortunately, the operator associated to this refined confidence set is no longer an (extended) Bellman operator. This is because the confidence set now depends on the specific vector v and can no longer be mapped to an extended action space (see Sec. 2.1.5). Nevertheless, in the rest of this section we show that after applying the transformations already introduced for SCAL$^+$ (e.g., η-perturbation of the transition probabilities), all the useful properties of Bellman operators that we have been exploiting in this thesis still hold (e.g., convergence of value iteration, dominance, etc.).

Refined operator

We now formally define the new operator discussed in the previous paragraph:

$$\forall v \in \mathbb{R}^S,\ \forall s \in \mathcal{S},\quad \mathfrak{L}_k v(s) := \max_{a \in \mathcal{A}_s} \left\{ \max_{r \in B_r^k(s,a)} \{r\} + \max_{p \in B_p^k(s,a) \cap \Theta_p^k(s,a,v)} \{p^\intercal v\} \right\} \tag{5.33}$$

where $\Theta_p^k(s,a,v) := \{p \in \Delta_S :\ p(\cdot|s,a)^\intercal v \leq \widehat{p}_k(\cdot|s,a)^\intercal v + c\beta_k^{sa}\}$. The only difference with the extended Bellman operator $\mathcal{L}_k$ (2.17) is that the initial confidence set $B_p^k(s,a)$ is *intersected* with $\Theta_p^k(s,a,v)$. Since the set $\Theta_p^k(s,a,v)$ depends on v, it is clear that $\mathfrak{L}_k$ is *not* a Bellman operator (the "extended action space" now depends on v). Fortunately, $\mathfrak{L}_k$ share a lot of properties with $\mathcal{L}_k$ as shown in the following lemmas.

Lemma 5.13 (Analogue to Lem. 2.5 and 5.7)

Let v and u be any two vectors in $\mathbb{R}^S$, then:

(a) $\mathfrak{L}_k$ *is* monotone*:* $v \geq u \implies \mathfrak{L}_k v \geq \mathfrak{L}_k u$.

(b) $\mathfrak{L}_k$ *is* non-expansive *both in span semi-norm and ℓ_∞-norm:*

$$sp\left(\mathfrak{L}_k v - \mathfrak{L}_k u\right) \leq sp\left(v - u\right) \quad \textit{and} \quad \|\mathfrak{L}_k v - \mathfrak{L}_k u\|_\infty \leq \|v - u\|_\infty.$$

(c) $\mathfrak{L}_k$ *is* linear*:* $\forall \lambda \in \mathbb{R},\quad \mathfrak{L}_k(v + \lambda e) = \mathfrak{L}_k v + \lambda e$.

Proof. See App. C.3. ■

Similarly to Sec. 5.5.1, we define $\widetilde{\mathfrak{L}}_k$ by *replacing* $B_r^k(s,a)$ and $B_p^k(s,a)$ by respectively $\widetilde{B}_r^k(s,a)$ and $\widetilde{B}_p^k(s,a)$ in Eq. 5.33 (see Def. 5.5), with the choice $\eta_k = r_{\max}/(c \cdot t_k)$ (as in Sec. 5.5) so that $\widetilde{B}_p^k(s,a) \neq \emptyset$. To define $\widetilde{\mathfrak{L}}_k$, we also *substitute* $\Theta_p^k(s,a,v)$ by $\widetilde{\Theta}_p^k(s,a,v)$ defined by:

$$\widetilde{\Theta}_p^k(s,a,v) := \{p \in \Delta_S :\ p(\cdot|s,a)^\intercal v \leq \widetilde{p}_k(\cdot|s,a)^\intercal v + c\beta_k^{sa}\}$$

where $\widetilde{p}_k(\cdot|s,a)$ is any *ℓ_1-projection* of $\widehat{p}_k(\cdot|s,a)$ onto $\widetilde{B}_p^k(s,a)$ (convex set). Since by definition $\widetilde{p}_k(\cdot|s,a) \in \widetilde{B}_p^k(s,a)$, the intersection $\widetilde{B}_p^k(s,a) \cap \Theta_p^k(s,a,v)$ is *never empty* and $\widetilde{\mathfrak{L}}_k$ is well-defined. The projection satisfies $\|\widetilde{p}_k(\cdot|s,a) - \widehat{p}_k(\cdot|s,a)\|_1 = 2 \cdot \max\{0, \eta_k - \widehat{p}_k(\overline{s}|s,a)\} \leq 2\eta_k$ where $\overline{s}$ is the reference state used to construct $\widetilde{B}_p^k(s,a)$ (see Def. 5.5 and the η_k-perturbation). In particular, it always holds that $\widetilde{p}_k(\overline{s}|s,a) - \widehat{p}_k(\overline{s}|s,a) = \max\{0, \eta_k - \widehat{p}_k(\overline{s}|s,a)\}$. To summarize, $\widetilde{\mathfrak{L}}_k$ is formally defined by:

$$\forall v \in \mathbb{R}^S,\ \forall s \in \mathcal{S},\quad \widetilde{\mathfrak{L}}_k v(s) := \max_{a \in \mathcal{A}_s} \left\{ \max_{r \in \widetilde{B}_r^k(s,a)} \{r\} + \max_{p \in \widetilde{B}_p^k(s,a) \cap \widetilde{\Theta}_p^k(s,a,v)} \{p^\intercal v\} \right\}. \tag{5.34}$$

$\widetilde{\mathfrak{L}}_k$ also satisfies Lem. 5.13 (the proof is similar, see App. C). Moreover, unlike $\mathfrak{L}_k$, $\widetilde{\mathfrak{L}}_k$ is always contractive (by construction) while being not too different from $\mathfrak{L}_k$ as shown in the following lemma.

Lemma 5.14 (See Def. 5.5 and Lem. 5.10)

The operator $\widetilde{\mathfrak{L}}_k$ is a 1-step γ_k-span-contraction with $\gamma_k \leq 1 - \eta_k < 1$, and for any vector $h \in \mathbb{R}^S$, $\|\mathfrak{L}_k h - \widetilde{\mathfrak{L}}_k h\|_\infty \leq \eta_k \cdot sp\left(h\right)$.

Proof. See App. C.4. ■

Finally, we define the associated "truncated operators" by composing $\widetilde{\mathfrak{L}}_k$ (resp. $\mathfrak{L}_k$) with the span truncation Γ_c defined in Def. 5.1: $\widetilde{\mathfrak{T}}_c^k := \Gamma_c \widetilde{\mathfrak{L}}_k$ (resp. $\mathfrak{T}_c^k := \Gamma_c \mathfrak{L}_k$). Due to Lem. 5.5, $\widetilde{\mathfrak{T}}_c^k$ also satisfies Lem. 5.13 and 5.14 (by composition). We can then deduce the following corollary.

Corollary 5.1 (See Lem. 5.9)

The following properties hold for $\widetilde{\mathfrak{T}}_c^k$:

1. *Optimality equation and uniqueness: There exists a solution $(\mathfrak{g}_k^+, \mathfrak{h}_k^+) \in \mathbb{R} \times \mathbb{R}^S$ to the optimality equation*

$$\widetilde{\mathfrak{T}}_c^k \mathfrak{h}_k^+ = \mathfrak{h}_k^+ + \mathfrak{g}_k^+ e. \tag{5.35}$$

   If $(g, h) \in \mathbb{R} \times \mathbb{R}^S$ is another solution of (5.35), then $g = \mathfrak{g}_k^+$ and there exists $\lambda \in \mathbb{R}$ s.t. $h = \mathfrak{h}_k^+ + \lambda e$.
2. *Convergence: For any initial vector $v_0 \in \mathbb{R}^S$, the sequence (v_n) generated by* ScOpt *(with operator $\widetilde{\mathfrak{L}}_k$ instead of L) converges to a solution vector $\mathfrak{h}_k^+$ of the optimality equation (5.35), and*

$$\lim_{n \to +\infty} \left(\widetilde{\mathfrak{T}}_c^k\right)^{n+1} v_0 - \left(\widetilde{\mathfrak{T}}_c^k\right)^n v_0 = \mathfrak{g}_k^+ e.$$

3. *(Approximate) Dominance: If $sp(h^*) \leq c$ and $\mathfrak{L}_k h^* \geq L h^*$ then $\mathfrak{g}_k^+ \geq g^* - \eta_k \cdot c$.*

5.8.2 Implementation and performance

Algorithm

The pseudo-code of SCAL* is similar to SCAL except that ScOpt is called with the refined operator $\widetilde{\mathfrak{L}}_k$ instead of $\widetilde{\mathcal{L}}_k$. In SCAL, line 9 of Alg. 5 (Eq. 3.5) was replaced by Eq. 5.15. In SCAL* this equation becomes:

$$(g_k, h_k, \pi_k) := \textsc{ScOpt}\left(\widetilde{\mathfrak{L}}_k, \widetilde{\mathfrak{G}}_k, \frac{r_{\max}}{t_k}, s_1, 0, \gamma_k\right). \tag{5.36}$$

We also introduce RLProba (Alg. 11), a slight modification of LProba that can solve the *refined* optimization problem $\max_{p \in \widetilde{B}_p^k(s,a) \cap \widetilde{\Theta}_p^k(s,a,v)} \{p^\intercal v\}$. Compared to LProba, RLProba takes an additional input ζ. The scalar w output by RLProba is identical to the scalar output by LProba if smaller than ζ, otherwise it is set equal to ζ (line 9 of Alg. 11). Since the value of w is increased at every iteration i (denoted w_i in Alg. 11, see e.g., line 4), it is possible to reduce the number of iterations of RLProba by checking whether the value is bigger than ζ and terminating the algorithm accordingly (line 2). Therefore, the computational

Algorithm 11 *Refined* Linear Programming for probability maximization (RLProba)

Input: A vector $v \in \mathbb{R}^S$ sorted in decreasing order $v(1) \geq v(2) \geq \cdots \geq v(S)$, S closed intervals $([a_i, b_i])_{1 \leq i \leq S}$ s.t. $1 \geq b_i \geq a_i \geq 0$ and $\sum_{i=1}^S a_i \leq 1 \leq \sum_{i=1}^S b_i$, *a scalar* $\zeta \in \mathbb{R}$
Output: A scalar w

Set $w_0 := \sum_{i=1}^S a_i \times v(i)$, $\Delta_0 := 1 - \sum_{i=1}^S a_i$ and $i := 1$ ▷ Initialization
while $\Delta_{i-1} > 0$ and $w_{i-1} < \zeta$ **do** ▷ Main loop
  Set $\delta_i := \min\{\Delta_{i-1},\ b_i - a_i\}$
  Update $w_i \leftarrow w_{i-1} + \delta_i \times v(i)$ ▷ Assign allowed weights to highest values of v first
  Update $\Delta_i \leftarrow \Delta_{i-1} - \delta_i$
  Increment $i \leftarrow i + 1$
end while
if $w_{i-1} > \zeta$ **then**
  Set $w := \zeta$
else
  Set $w := w_{i-1}$
end if

complexity of RLProba is comparable to the one of LProba, and sometimes even smaller. The correctness of Alg. 11 is a direct consequence of the proof of Lem. 5.13 in App. C.

In practice, at every iteration $n \geq 0$ of ScOpt, and for all state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$, RLProba is called with $\zeta = \widetilde{p}_k(\cdot|s,a)^\intercal v_n + c\beta_k^{sa}$. The outputs of RLProba are then used to compute $\widetilde{\mathfrak{L}}_k v_n$ and then $\widetilde{\mathfrak{T}}_c^k v_n$ (see Alg. 8).

Regret guarantees

By construction, SCAL* satisfies exactly the same regret guarantees as SCAL (Thm. 5.4 and 5.5) but the *degree of optimism* is now potentially *tighter* due to the *restriction* $p_k(\cdot|s,a) \in \widetilde{\Theta}_p^k(s,a,h_k)$ for all state-action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$ and all episodes k. As discussed in Sec. 5.7, this restriction does not allow to *refine* the final regret bound with current proof techniques.

5.9 Conclusion

In this chapter we introduced SCAL, a UCRL2-like algorithm that is able to efficiently balance exploration and exploitation in any *weakly communicating* MDP for which a finite bound c on the optimal bias span $sp(h^*)$ is known. While UCRLB exclusively relies on *optimism* and uses EVI to compute the exploratory policy, SCAL leverages the knowledge of c through the use of ScOpt, a new planning algorithm specifically designed to handle constraints on the bias span. We showed both theoretically and empirically that SCAL achieves smaller regret than UCRL2, with a negligible additional computational cost. Although SCAL was inspired by Regal.C, it is the only *implementable* approach so far. Therefore, this paper answers the long-standing open question of whether it is actually possible to design an *algorithm* that does not scale with the diameter (or the travel-budget) in the worst case. SCAL also paves the way for implementable algorithms able to learn in an MDP with *continuous* state

space (Qian et al., 2018b). Indeed, existing algorithms achieving regret guarantees in this framework (Ortner and Ryabko, 2013; Lakshmanan et al., 2015) all rely on REGAL.C.

Inspired by SCAL we derived SCAL$^+$, the first analysis of exploration bonus in infinite-horizon undiscounted problems. We showed that SCAL$^+$ achieves the tightest level of optimism for OFU algorithms by achieving the optimal dependence in the bonus w.r.t. the state dimensionality (it cannot further reduced while preserving theoretical guarantees given the lower-bound of Prop. 2.12). Unfortunately, this tighter optimism does not imply a tighter bound.

We combined the advantages of both SCAL and SCAL$^+$ into a single algorithm: SCAL*.

For all the algorithms presented in this chapter (SCAL, SCAL$^+$ and SCAL*), it is an open question whether the assumption that c is known can be relaxed. We conjecture that the knowledge of $sp\left(h^*\right)$ is necessary to improve the regret upper-bound of UCRLB (i.e., replace the travel-budget by the optimal bias span), even though we leave this question for future work.

In Chap. 4, we showed that when the MDP is not communicaing, the regret of any "efficient" learning algorithm cannot grow logarithmically with time. However, Thm. 4.2 does not apply in the case where a bound on the optimal bias span is known since the MDPs with small ε in Ex. 4.5 (used to prove Thm. 4.2) do not satisfy $sp\left(h^*\right) \leq c$. We conjecture that a logarithmic regret bound similar to Thm. 2.37 can be derived for SCAL, SCAL$^+$ and SCAL*, with D replaced by $c/r_{\max}$.

6 Hierarchical exploration–exploitations with options

6.1 Introduction

Tractable learning of how to make good decisions in complex domains over many time steps almost definitely requires some form of hierarchical reasoning. One powerful and popular framework for incorporating temporally-extended actions and hierarchical structures in the context of reinforcement learning is the *options* framework (Sutton et al., 1999). An important feature of this framework is that MDP planning and learning algorithms can be easily extended to accommodate options, thus obtaining algorithms such as option value iteration and Q-learning (Sutton et al., 1999), LSTD (Sorg and Singh, 2010), and actor-critic (Bacon and Precup, 2015). Temporally extended actions are particularly appealing for high dimensional problems that naturally decompose into a hierarchy of subtasks. Creating and leveraging options has been the subject of many papers over the last two decades (see e.g., McGovern and Barto (2001); Menache et al. (2002); Şimşek and Barto (2004); Castro and Precup (2012); Levy and Shimkin (2011); Sairamesh and Ravindran (2012); Mann et al. (2014)) and it has been of particular interest recently in combination with deep reinforcement learning, with a number of impressive empirical successes. For instance, Tessler et al. (2016) recently obtained promising results by combining options and deep learning for lifelong learning in the challenging domain of Minecraft.

Intuitively (and empirically) temporal abstraction can help speed up learning (reduce the amount of experience needed to learn a good policy) by shaping the actions selected towards more promising sequences of actions (Stolle and Precup, 2002), and it can reduce planning computation through reducing the need to evaluate over all possible actions (see e.g., Mann and Mannor (2014)). A large body of the literature has focused on how to automatically construct options that are beneficial to the learning process within a single task or across similar tasks. An alternative approach is to design an initial set of options and optimize it during the learning process itself (see e.g., interrupting options (Mann et al., 2014) and options with exceptions (Sairamesh and Ravindran, 2012)).

Despite the empirical evidence of the effectiveness of most of these methods, it is well

known that options may as well worsen the performance w.r.t. learning with "primitive" actions (Jong et al., 2008). Intuitively, limiting action selection only to temporally-extended options might hamper the exploration of the environment by restricting the policy space. Moreover, most of the proposed methods are heuristic in nature and the theoretical understanding of the actual impact of options on the learning performance is still fairly limited. Notable exceptions are the recent results of Mann and Mannor (2014) and Brunskill and Li (2014). Nonetheless, Mann and Mannor (2014) rather focus on a batch setting and they derive a sample complexity analysis of approximate value iteration with options. Brunskill and Li (2014) derived sample complexity bounds for an RMax-like exploration-exploitation algorithm for semi-Markov decision processes (SMDPs). While MDPs with options can be mapped to SMDPs, we will later show that their analysis cannot be immediately translated into the PAC-MDP sample complexity of learning in an MDP with options, which makes it harder to evaluate their potential benefit. Therefore, we argue that in addition to the exciting work being done in heuristic and algorithmic approaches that leverage and/or dynamically discover options, it is important to build a formal understanding of how and when options may help or hurt reinforcement learning performance, and that such insights may also help inform empirically motivated options-RL research. In this chapter, we consider the case where a fixed set of options is provided and we study their impact on the learning performance w.r.t. learning without options. In particular, we derive the first regret analyses of learning with options.

Relying on the fact that using options in an MDP induces a semi-Markov decision process (SMDP), we first introduce a variant of UCRLB for SMDPs and we upper and lower-bound its regret. While this result is of independent interest for learning in SMDPs, its most interesting aspect is that it can be translated into a regret bound for learning with options in MDPs and it provides a first understanding on the sufficient conditions for a set of options to reduce the regret w.r.t. learning with primitive actions. The resulting analysis explicitly shows how options can be beneficial whenever the navigability among the states in the original MDP is not compromised (i.e., the MDP travel-budget is not significantly increased), the level of temporal abstraction is high (i.e., options have long durations, thus reducing the number of decision steps), and the optimal policy with options performs as well as the optimal policy using primitive actions. While this result makes explicit the impact of options on the learning performance, the proposed algorithm (SUCRL in short) needs prior knowledge on the parameters of the distributions of cumulative rewards and durations of each option to construct confidence intervals and compute optimistic solutions. In the second part of this chaper, we remove the limitations of having prior knowledge on options by introducing a "prior knowledge-free" version of SUCRL named FSUCRL. We derive regret bounds for FSUCRL that clarify the regret bound of SUCRL. Finally, we provide illustrative emperiments where the empirical results support the theoretical findings. We also empirically compare FSUCRL to SUCRL and UCRLB (i.e., learning without options).

The work presented in this chapter extends the conference papers (Fruit and Lazaric, 2017) and (Fruit et al., 2017).

6.2 The option framework

6.2.1 Formal definition of options

We start this section with the formal definition of an option.

Definition 6.1 (Sutton et al. (1999))

A (Markov) option *is a 3-tuple* $o = \{\mathcal{I}_o, \beta_o, \pi_o\}$ *where*

- $\mathcal{I}_o \subseteq \mathcal{S}$ *is the set of states where the option can be* initiated,
- $\beta_o : \mathcal{S} \to [0, 1]$ *is the probability distribution that the option* ends *in a given state,*
- $\pi_o \in \Pi^{SR}$ *is the* policy *followed until the option ends.*

An agent can decide to *play* option o in any state belonging to $\mathcal{I}_o$. Once option o has been initiated, policy π_o is executed until the *termination condition* of the option is triggered. During the execution of option o, a *Bernouilli* random variable with success probability $\beta_o(s)$ is sampled *independently* from the past history every time a new state $s \in \mathcal{S}$ is *visited*. The execution of the option *ends* if and only if the outcome of the Bernouilli is a *success*. It is worth pointing out that any *"primitive"* action $a \in \mathcal{A}_s$ available in state $s \in \mathcal{S}$ can be interpreted as an option with an arbitrary initial state space $\mathcal{I}_a \ni s$, a stopping distribution $\beta_a(s) = 1$ for all $s \in \mathcal{S}$, and any arbitrary policy $\pi_a \in \Pi^{\text{SR}}$ satisfying $\pi_a(s) = a$. The converse is of course not true: all options are not primitive actions since an option can *last* for more than just 1 time step (unlike a primitive action). Since the only restriction is that all 3 components of an option should satisfy the *Markov property*[1], Def. 6.1 provides a very rich and flexible definition of *temporally extended actions*. It is possible to extend Def. 6.1 by relaxing the Markov constraint, although it is unclear whether such a level of generality can be of any interest given the *Markov structure* of the *underlying* MDP.

In this chapter, we assume that the original action space $\mathcal{A}$ of the MDP is *replaced* by a set of options $\mathcal{O}$ *given* (i.e., known) to the learning agent, and possibly containing primitive actions. This new framework is therefore a generalization of the MDP framework considered so far (and introduced at the beginning of the thesis, see Sec. 2.1). Given a set of (Markov) options $\mathcal{O}$ satisfying Def. 6.1, we denote by $\mathcal{O}_s$ the set of options available in state $s \in \mathcal{S}$ i.e., $\mathcal{O}_s := \{o \in \mathcal{O} : \ s \in \mathcal{I}_o\}$. In the previous chapters, we have always considered state-action *pairs* $(s, a) \in \mathcal{S} \times \mathcal{A}$ rather than isolated actions. Similarly, in this chapter the state-option pairs $(s, o) \in \mathcal{S} \times \mathcal{O}$ will be the fundamental *bricks* of the *decision problem* at hand. In the rest of this chapter, we will slightly abuse notation and denote by $\mathcal{S} \times \mathcal{O}$ the set of *"admissible"* state-option pairs i.e., the set $\{(s, o) : \ o \in \mathcal{O}, \ s \in \mathcal{I}_o\}$.

As shown in the seminal work of Sutton et al. (1999), one possible way to describe the decision process *induced* by a set of options $\mathcal{O}$ onto an MDP M is through the notion of *Semi-Markov Decision Process* (SMDP). We will make this statement formal later and start by briefly presenting the concept of SMDP in the next section.

[1] The starting state, the terminal condition and the policy all depend exclusivey on the current state.

6.2.2 Semi-Markov Decision Processes

Definition

A Semi-Markov Decision Process (SMDP) M is a 5-tuple[2] $\langle \mathcal{S}, \mathcal{A}, r, p, \tau \rangle$. As in the definition of an MDP (see Sec. 2.1.1), $\mathcal{S}$ and $\mathcal{A}$ denote respectively the state and action space of the SMDP, and r and p the expected rewards and transition probabilities. The last term τ in the definition of an SMDP refers to *holding times*. After playing action a in state s, the agent waits for an *expected duration* $\tau(s,a) > 0$ before observing the next state s' with probability $p(s'|s,a)$ and receiving the expected reward $r(s,a)$. We make the same assumptions on $\mathcal{S}$ and $\mathcal{A}$ as in Sec. 2.1 i.e., $\mathcal{S}$ is assumed to be *finite* while $\mathcal{A}$ is either *finite* or *compact* depending on the context. When $\mathcal{A}$ is a compact set, we also assume that for all $s, s' \in \mathcal{S}$, the maps $a \mapsto r(s,a)$, $a \mapsto \tau(s,a)$ and $a \mapsto p(s'|s,a)$ are *continuous* functions of a. A *major difference* with Sec. 2.1 is that we assume that all *sampled* (as opposed to expected) rewards and holding times are *positive* but *not necessarily bounded*, although we will also study this specific case in detail. The reason is that this assumption is too *restrictive* to model options, as will be clear in the next section. Nevertheless, we always assume that the *expected value* $\tau(s,a)$ are *uniformly bounded* on $\mathcal{S} \times \mathcal{A}$ i.e., $\tau_{\max} := \sup_{s,a} \tau(s,a) < +\infty$, and we also assume that there exists $r_{\max} > 0$ such that $r(s,a) \leq r_{\max} \tau(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. As a consequence, $r(s,a)$ is also uniformly bounded on $\mathcal{S} \times \mathcal{A}$. Finally, we assume that there exists $\tau_{\min} > 0$ such that for all state-action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\tau(s,a) \geq \tau_{\min}$.

Since the *classification* of MDPs presented in Def. 2.2 (see Sec. 2.1.1) only depends on p (i.e., on transition probabilities), we can also apply it to SMDPs. SMDPs can thus be classified according to their chain structure just like MDPs: *ergodic*, *unichain*, *communicating*, *weakly communicating* or *multi-chain*.

Note that an MDP can be interpreted as a particular case of an SMDP where $\tau(s,a) = 1$ for all state-action pairs. SMDPs are therefore a generalization of MDPs with a *temporal component*.

Gain optimality

Like in the MDP case (see Sec. 2.2), in the undiscounted setting the goal is to maximize the long-term average reward which is now expressed as an *average over elapsed time* (and not just over time steps as in Eq. 2.8):

$$\sup_{\pi \in \Pi} \left\{ \liminf_{n \to +\infty} \frac{\mathbb{E}^{\pi} \left[\sum_{i=1}^{n} r_i \middle| s_1 \sim \mu_1 \right]}{\mathbb{E}^{\pi} \left[\sum_{i=1}^{n} \tau_i \middle| s_1 \sim \mu_1 \right]} \right\}. \tag{6.1}$$

If $\tau(s,a) = 1$ for all state-action pairs, then (6.1) is equivalent to (2.8). Similarly, for any stationary randomized policy $\pi \in \Pi^{\mathrm{SR}}$, the gain of π is defined as

[2] In comparison, an MDP is usually described as a 4-tuple, see Sec. 2.1.1.

$$g^\pi(s) := \lim_{n\to+\infty} \frac{\mathbb{E}^\pi\left[\sum_{i=1}^n r_i \middle| s_1 = s\right]}{\mathbb{E}^\pi\left[\sum_{i=1}^n \tau_i \middle| s_1 = s\right]} \tag{6.2}$$

where the limit always exists. On the other hand, the bias is defined as

$$h^\pi(s) := \underset{n\to+\infty}{C\text{-}\lim} \mathbb{E}^\pi\left[\sum_{i=1}^n \left(r_i - \tau_i \cdot g^\pi(s_i)\right) \middle| s_1 = s\right], \tag{6.3}$$

where the Cesaro-limit always exists. For any randomized Markov decision rule $d \in D^{\mathrm{MD}}$, we denote by $\tau_d \in \mathbb{R}^S$ the vector of holding times i.e., $\tau_d(s) = \tau(s, d(s))$ for all $s \in \mathcal{S}$. The following proposition is the generalization of Prop. 2.4 to SMDPs (see (Schweitzer, 1985, Theorem 1) for the proof).

Proposition 6.1

Let M be a weakly communicating *MDP and denote by $\Pi^* \subseteq \Pi^{SD}$ the set of maximizers of (6.1) in Π^{SD}. If any of the following two assumptions hold:*

1. *the action space A is* finite,
2. $\Pi^* \neq \emptyset$ *and* $\sup_{\pi\in\Pi^*} sp\left(h^\pi\right) < +\infty$,

then there exists a solution $(g^*, h^*) \in \mathbb{R} \times \mathbb{R}^S$ to the fixed point equation:

$$h^* = \max_{d\in D^{MD}} \{r_d - \tau_d \cdot g^* + P_d h^*\}.$$

Moreover, for any such solution (g^*, h^*) and for all $s \in \mathcal{S}$,

$$g^* = \max_{\pi\in\Pi} \left\{ \liminf_{n\to+\infty} \frac{\mathbb{E}^\pi\left[\sum_{i=1}^n r_i \middle| s_1 = s\right]}{\mathbb{E}^\pi\left[\sum_{i=1}^n \tau_i \middle| s_1 = s\right]} \right\}.$$

Finally, any stationary greedy policy $\pi^* = (d^*)^\infty$ satisfying $d^* \in \arg\max_{d\in D^{MR}} \{r_d + P_d h^*\}$ is optimal i.e., $\pi^* \in \Pi^*$.

We recall that unlike g^*, h^* is not unique (see Sec. 2.2).

A natural next step is to derive an algorithm to compute an optimal policy. To that end, we first introduce a transformation called *uniformization*.

Uniformization of an SMDP

We call *"uniformization"* the transformation of an SMDP $M = \langle \mathcal{S}, \mathcal{A}, r, p, \tau \rangle$ into an MDP $M_{\mathrm{eq}} = \langle \mathcal{S}, \mathcal{A}, r_{\mathrm{eq}}, p_{\mathrm{eq}} \rangle$ with identical state and action spaces, and such that $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$:

$$
\begin{aligned}
r_{\mathrm{eq}}(s,a) &:= \frac{r(s,a)}{\tau(s,a)} \\
p_{\mathrm{eq}}(\cdot|s,a) &:= \frac{\alpha}{\tau(s,a)}\left(p(\cdot|s,a) - e_s\right) + e_s
\end{aligned}
\tag{6.4}
$$

where $\alpha < \tau_{\min}$. The assumption $\tau(s,a) \geq \tau_{\min}$ ensures that $p_{\mathrm{eq}}(\cdot|s,a)$ is a well-defined transition probability. Furthermore, since $p_{\mathrm{eq}}(s|s,a) > 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, the Markov Chain induced by any Markov randomized decision rule $d \in D^{\mathrm{MR}}$ is *aperiodic*. In the following, we denote by L_{eq} the optimal Bellman operator of M_{eq}.

We first notice that the transformation *preserves the chain structure* e.g., if M is weakly communicating/unichain/etc., so is M_{eq}. This is immediate to see since the chain structure of M only depends on which transition probabilities $p(s'|s,a)$ with $s' \neq s$ are equal to 0, and $p(s'|s,a) = 0 \iff p_{\mathrm{eq}}(s'|s,a) = 0$ whenever $s' \neq s$.

In the case of a compact action space $\mathcal{A}$, $a \mapsto r_{\mathrm{eq}}(s,a)$ and $a \mapsto p_{\mathrm{eq}}(\cdot|s,a)$ are *continuous* mappings since $a \mapsto r(s,a)$, $a \mapsto \tau(s,a)$ and $a \mapsto p(s'|s,a)$ are assumed to be continuous (see above). Moreover, the condition $r(s,a) \leq r_{\max}\tau(s,a)$ implies that $r_{\mathrm{eq}}(s,a) \in [0, r_{\max}]$. As a result, if SMDP M satisfies the assumptions stated earlier, M_{eq} satisfies the assumptions of all the MDPs studied so far in this thesis (see Sec. 2.1.1).

Uniformization allows to analyze an SMDP as if it was an MDP (Puterman, 1994, Section 11.4.3). We illustrate this claim with the following lemma.

Proposition 6.2 (Proposition 11.4.5 of Puterman (1994))

If there exists $(g^*_{eq}, h^*_{eq}) \in \mathbb{R} \times \mathbb{R}^S$ *solution to optimal Bellman equation of* M_{eq} *i.e.,*

$$h^*_{eq} + g^*_{eq} e = L_{eq} h^*_{eq},$$

then $(g^*_{eq}, \alpha h^*_{eq})$ *is solution to the optimal Bellman equation of* M *i.e.,*

$$\alpha h^*_{eq} = \max_{d \in D^{MD}} \left\{ r_d - \tau_d \cdot g^*_{eq} + P_d \left(\alpha h^*_{eq} \right) \right\}.$$

Instead of looking for a solution to the optimality equation of SMDP M, we can search for a solution to the optimality equation of MDP M_{eq} using the tools of Sec. 2.2. Rather than checking whether M satisfies the assumptions of Prop. 6.1, we can verify whether M_{eq} satisfies the assumptions of Prop. 2.4 (existence of a solution to the MDP Bellman optimality equation). Whenever Prop. 2.4 holds for M_{eq}, it is clear that Prop. 2.6 holds as well (i.e., value iteration converges) since all stationary determinisitic decision rule in M_{eq} are aperiodic (see above).

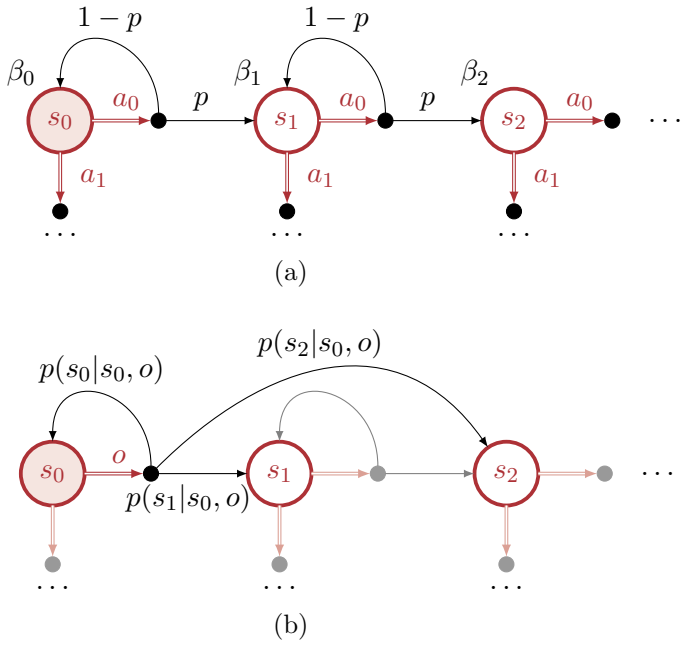


Figure 6.1: MDP with a state-option (s_0, o) executing a_0 in all states with termination probabilities $\beta_o(s_0) = \beta_0$, $\beta_o(s_1) = \beta_1$ and $\beta_o(s_2) = 1$ (Fig. 6.1a), and dynamics of the SMDP associated to this state-option (Fig. 6.1b).

6.2.3 Markov options as absorbing Markov Chains

Markov Chain of an option. Any option defined on an MDP can be described by a Markov Reward Process (MRP) i.e., a Markov Chain (MC) together with a reward function. The state space of the MC contains all states that are *reachable* by the option and all *terminal states* are *absorbing* states of the MC (see Fig. 6.1 and 6.2). More formally, for any state-option pair $(s, o) \in \mathcal{S} \times \mathcal{O}$ the set of *inner states* $\mathcal{S}_{s,o}$ includes the *initial state* s and all states x with $\beta_o(x) < 1$ that are reachable by executing π_o starting from s (e.g., $\mathcal{S}_{s,o} = \{s_0, s_1\}$ in Fig. 6.1), while the set of *absorbing states* $\mathcal{S}^{\text{abs}}_{s,o}$ includes all states with $\beta_o(x) > 0$ (e.g., $\mathcal{S}^{\text{abs}}_{s,o} = \{s_0, s_1, s_2\}$ in Fig. 6.2). We denote by $S_{s,o}$ (resp. $S^{\text{abs}}_{s,o}$) the cardinality of $\mathcal{S}_{s,o}$ (resp. $\mathcal{S}^{\text{abs}}_{s,o}$). The MC associated to (s, o) is characterized by a transition matrix $P_{s,o}$ of dimension $(S_{s,o} + S^{\text{abs}}_{s,o}) \times (S_{s,o} + S^{\text{abs}}_{s,o})$ with *canonical form*

$$P_{s,o} := \begin{bmatrix} Q_{s,o} & V_{s,o} \\ 0 & I \end{bmatrix} \text{ where } \begin{cases} Q_{s,o}(x,y) := (1-\beta_o(y)) \cdot \sum_a p(y|x,a)\pi_o(a|x), \ \forall x, y \in \mathcal{S}_{s,o} \\ V_{s,o}(x,y) := \beta_o(y) \cdot \sum_a p(y|x,a)\pi_o(a|x), \ \forall (x,y) \in \mathcal{S}_{s,o} \times \mathcal{S}^{\text{abs}}_{s,o} \end{cases}.$$

$Q_{s,o}$ is the transition matrix between *inner states* (dimension $S_{s,o} \times S_{s,o}$), $V_{s,o}$ is the transition matrix from *inner states* to *absorbing states* (dimension $S_{s,o} \times S^{\text{abs}}_{s,o}$), and I is the identity matrix (dimension $S^{\text{abs}}_{s,o} \times S^{\text{abs}}_{s,o}$). Note that some states x may belong both to $\mathcal{S}_{s,o}$ and $\mathcal{S}^{\text{abs}}_{s,o}$ if $1 > \beta_o(x) > 0$ (i.e., $\mathcal{S}_{s,o} \cap \mathcal{S}^{\text{abs}}_{s,o} \neq \emptyset$), and therefore $S_{s,o} + S^{\text{abs}}_{s,o}$ is not always smaller than S (even though it is always upper-bounded by 2S). We also denote by $r_{s,o} := (r(x, \pi_o(x)))_{x \in \mathcal{S}_{s,o}}$ the vector of rewards associated to state-option pair $(s, o) \in \mathcal{S} \times \mathcal{O}$.

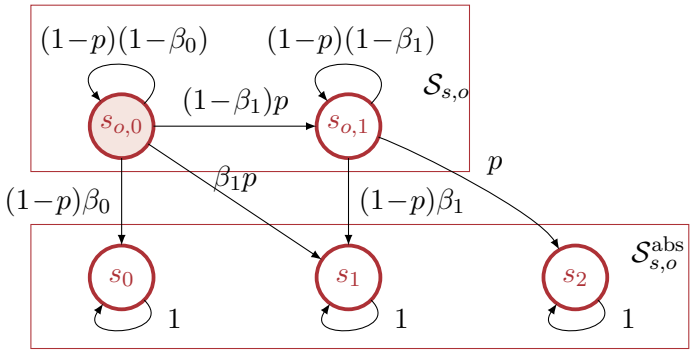


Figure 6.2: Absorbing MC associated to state-option (s_0, o) of Fig. 6.1.

Absorbing property. Nothing in Def. 6.1 guarantees that a state-option pair $(s, o) \in \mathcal{S} \times \mathcal{O}$ will ever end once initiated. This problem will occur if for example $\beta_o(x) = 0$ for all states $x \in \mathcal{S}$, or if $\beta_o(x) > 0$ only for some states $x \in \mathcal{S}$ that are reached with probability 0 under policy π_o. Mathematically, this means that $P_{s,o}$ is not an *absorbing Markov Chain* i.e., absorbing states are reached in *finite time* with probability *strictly less* than 1. A *never-ending* option will be problematic if π_o is very suboptimal compared to other options: once this "pathological" option has started, no other option can ever be played (it is a sort of "deadlock"). For this reason, we make the following assumption.

Assumption 6.1

All options terminate in finite time with probability 1, or equivalently, $P_{s,o}$ is an absorbing *Markov Chain for all state-option pair $(s, o) \in \mathcal{S} \times \mathcal{O}$.*

The MC $P_{s,o}$ is absorbing if and only if $Q_{s,o}$ is *strictly substochastic* i.e., $Q_{s,o}e \leq e$ with the inequality *strict* in at least one coordinate. If $\beta_o(x) > 0$ for all states $x \in \mathcal{S}$, Asm. 6.1 always holds by definition of $Q_{s,o}$. It is thus not necessary to know the dynamics of the MDP to enforce this property, even though having some prior knowledge is usually useful to define a well-behaved option.

Characterization of an absorbing MC. When $Q_{s,o}$ is strictly substochastic, $I - Q_{s,o}$ is always invertible since the *spectral radius* of $Q_{s,o}$ $-\rho(Q_{s,o})-$ is stricly smaller than 1. In the theory of absorbing MCs (Grinstead and Snell, 2003, Section 11.2), the *fundamental matrix* associated to $P_{s,o}$ is defined as

$$N_{s,o} := (I - Q_{s,o})^{-1} \tag{6.5}$$

i.e., $N_{s,o}(j|i)$ (i-th row and j-th column) is the expected number of times inner state $j \in \mathcal{S}_{s,o}$ is visited when starting from inner state $i \in \mathcal{S}_{s,o}$. The *absorbing transition matrix*

$$B_{s,o} := N_{s,o} V_{s,o} \tag{6.6}$$

contains the probability of terminating in an absorbing state $j \in \mathcal{S}_{s,o}^{\text{abs}}$ when starting from an inner state $i \in \mathcal{S}_{s,o}$. The i-th entry of the vector

$$\tau_{s,o} := N_{s,o} e \tag{6.7}$$

corresponds to the expected *number of steps before absorption* when starting from inner state $i \in \mathcal{S}_{s,o}$. For example, $\tau_{s,o}(s)$ is the expected duration of state-option (s,o) while $B_{s,o}(j|s)$ is the probability that it ends in state $j \in \mathcal{S}_{s,o}^{\text{abs}}$. The set of possible *terminal* states of (s,o) is:

$$\mathcal{S}_{s,o}^{\text{term}} := \{j \in \mathcal{S}_{s,o}^{\text{abs}} : \ B_{s,o}(j|s) > 0\}. \tag{6.8}$$

Finally, we denote by $r_{s,o} := \left(\sum_a r(x,a)\pi_o(x|a)\right)_{x \in \mathcal{S}_{s,o}}$ the reward vector associated to (s,o). The i-th entry of the vector

$$R_{s,o} := N_{s,o} r_{s,o} \tag{6.9}$$

is the *expected cumulative reward* collected before absorption when starting from inner state $i \in \mathcal{S}_{s,o}$. In particular, $R_{s,o}(s)$ is the expected cumulative reward of state-option (s,o).

6.2.4 MDP with options as an SMDP

Availability of options. When we consider an arbitrary set of options $\mathcal{O}$, it is possible that some options terminate in states where no other option is *available*. In this case, the decision process is somehow *ill-posed*. To avoid this situation, we make an additional assumption.

Assumption 6.2

For any state-option pair $(s,o) \in \mathcal{S} \times \mathcal{O}$, $x \in \mathcal{S}_{s,o}^{term} \implies \mathcal{O}_x \neq \emptyset$.

Asm. 6.2 is not really restrictive since we can always use primitive actions as default options. Even under Asm. 6.2, it is not a problem that $\mathcal{O}_s = \emptyset$ for some $s \in \mathcal{S}$ as long as state s is not a terminal state for any other state-option pair. Given an initial distribution over states $\mu_1 \in \mathcal{P}(S)$, we recursively define the set of reachable states at *the level of options* $\mathcal{S}_{\mathcal{O}}^{\mu_1} \subseteq \mathcal{S}$:

$$\mathcal{S}_{\mathcal{O}}^{\mu_1} := \bigcup_{k=1}^{+\infty} \mathcal{S}_k \quad \text{where} \quad \begin{cases} \mathcal{S}_1 := \{s \in \mathcal{S} : \ \mu_1(s) > 0\} \\ \mathcal{S}_{k+1} := \bigcup_{s \in \mathcal{S}_k} \bigcup_{o \in \mathcal{O}_s} \mathcal{S}_{s,o}^{\text{term}} \end{cases}. \tag{6.10}$$

Main results. We are now ready to state the main result of this section which relates an MDP with (Markov) options to an SMDP.

Proposition 6.3 (Sutton et al. 1999)

Let $M = \langle \mathcal{S}, \mathcal{A}, r, p \rangle$ be an MDP with bounded rewards $0 \leq r \leq r_{\max}$, $\mathcal{O}$ a set of (Markov) options satisfying both Asm. 6.1 and 6.2, and $\mu_1 \in \mathcal{P}(\mathcal{S})$ an initial distribution over states. For all states $s, s' \in \mathcal{S}_{\mathcal{O}}$ and options $o \in \mathcal{O}_s$, we define the transition probabilities $b(s'|s,o) := B_{s,o}(s'|s)$, reward $R(s,o) := R_{s,o}(s)$ and holding time $\tau(s,o) := \tau_{s,o}(s)$. The decision process $M_{\mathcal{O}}^{\mu_1} = \{\mathcal{S}_{\mathcal{O}}^{\mu_1}, \mathcal{O}, b, R, \tau\}$ is an SMDP satisfying $\tau \geq 1$ and $0 \leq R \leq r_{\max}\tau$.

In SMDP $M_{\mathcal{O}}^{\mu_1}$, $\tau_{\min} = 1$. In the rest of this chapter, we set $\alpha = 0.9 < \tau_{\min}$ for the uniformization coefficient (this choice is arbitrary). We will often remove the dependency in μ_1 and use the notations $M_{\mathcal{O}}$ and $\mathcal{S}_{\mathcal{O}}$ to denote the SMDP and the state space respectively.

Any stationary policy $\pi_{\mathcal{O}} \in \Pi_{M_{\mathcal{O}}}^{\text{SR}}$ can be interpreted as a policy $\pi \in \Pi_M$ so that at each step, π selects an action available in M based on the policy of the current option being executed. Although, $\pi_{\mathcal{O}}$ is stationary, the primitive actions played by π not only depend on the current state in $\mathcal{S}$, but also on the option being executed, potentially inducing a non-stationary policy. The two reward processes induced by π and $\pi_{\mathcal{O}}$ in respectively M and $M_{\mathcal{O}}$ are strongly related as shown in Cor. 6.1.

Corollary 6.1

Let M be an MDP, $\mathcal{O}$ a set of options satisfying both Asm. 6.1 and 6.2 and $M_{\mathcal{O}}$ the corresponding SMDP (Prop. 6.3). Let $\pi_{\mathcal{O}} \in \Pi_{M_{\mathcal{O}}}^{SR}$ be any stationary policy on $M_{\mathcal{O}}$ and $\pi \in \Pi_M$ the equivalent policy on M (not necessarily stationary). For any state $s \in \mathcal{S}_{\mathcal{O}}$, we have: $g_{M_{\mathcal{O}}}^{\pi_{\mathcal{O}}}(s) = g_M^{\pi}(s)$.

Proof. The proof is straightforward (Fruit et al., 2017, Lemma 2). ■

As a result of Cor. 6.1, it makes sense to compare the performances of policies in $\Pi_{M_{\mathcal{O}}}^{\text{SR}}$ and policies in Π_M.

Distribution of holding times and rewards. We will now extend the result of Prop. 6.3 by analyzing the distribution of τ and R in $M_{\mathcal{O}}$. By construction, for any state-option pair $(s,o) \in \mathcal{S}_{\mathcal{O}} \times \mathcal{O}$, the holding time corresponds to the time before absorption starting in the equivalent absorbing MC (described in Sec. 6.2.3). Such discrete random variables (r.v.) are said to follow a *discrete phase-type distribution* (Nielsen, 2012). The probability mass function can be expressed using powers of $Q_{s,o}$ (Nielsen, 2012, Section 1.3.1)[3]:

$$\forall k \in \mathbb{N}^*, \quad \mathbb{P}(\tau(s,o) = k) = e_s^{\intercal}(Q_{s,o})^{k-1} V_{s,o} e. \tag{6.11}$$

[3] We denote by $\mathbb{N}^*$ the set of strictly positive integers.

Discrete phase-type r.v. are almost surely *finite* but not almost surely *bounded* (for any arbitrarily large k, the probability mass in Eq. 6.11 may be non-zero). This is the reason why we did not assume that the *sampled* holding times of an SMDP are *bounded*, but only that they have a *finite* expectation (see above). In all the learning algorithms we have presented so far in this thesis, we used concentration inequalities on *bounded* r.v. To apply the same approach in the context of options, we need to rely on *more general* inequalities that hold for *unbounded* r.v. We introduce the notions of *sub-exponential* and *sub-Gaussian* random variables.

Definition 6.2 (Wainwright (2015))

A random variable X with mean $\mu < +\infty$ is said to be sub-exponential*, if one of the following equivalent conditions is satisfied:*

1. *(Laplace transform condition) There exists*[4] $(\sigma, d) \in \mathbb{R}^+ \times \mathbb{R}^{+*}$ *such that:*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\sigma^2\lambda^2}{2}} \quad \text{for all } \lambda \in \mathbb{R} \text{ s.t. } |\lambda| < \frac{1}{d}. \tag{6.12}$$

   We use the notation $X \in \text{subExp}(\sigma, d)$.
2. *There exists $c_0 > 0$ such that $\mathbb{E}[e^{\lambda(X-\mu)}] < +\infty$ for all $\lambda \in \mathbb{R}$ s.t. $|\lambda| \leq c_0$.*

Definition 6.3 (Wainwright (2015))

A random variable X with mean $\mu < +\infty$ is said to be sub-Gaussian *if and only if there exists $\sigma \in \mathbb{R}^+$ such that:*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\sigma^2\lambda^2}{2}} \quad \text{for all } \lambda \in \mathbb{R}. \tag{6.13}$$

We use the notation $X \in \text{subGauss}(\sigma)$ to denote a sub-Gaussian r.v. with parameter σ.

By definition, if $X \in \text{subGauss}(\sigma)$ then $X \in \text{subExp}(\sigma, d)$ for any $d > 0$ but the reverse is not true i.e., Def. 6.2 is more general than Def. 6.3. Also, if $X \in \text{subGauss}(\sigma)$ (resp. $X \in \text{subExp}(\sigma, d)$) then $-X \in \text{subGauss}(\sigma)$ (resp. $-X \in \text{subExp}(\sigma, d)$). Finally, if $X \in \text{subExp}(\sigma_1, d_1)$ (resp. $X \in \text{subGauss}(\sigma_1)$), $\sigma_2 \geq \sigma_1$ and $d_2 \geq d_1$ then $X \in \text{subExp}(\sigma_2, d_2)$ (resp. $X \in \text{subGauss}(\sigma_2)$).

It is possible to generalize Hoeffding and Bernstein inequalities to respectively sub-exponential and sub-Gaussian random variables.

[4] We denote by $\mathbb{R}^+$ and $\mathbb{R}^{+*}$ the set of nonnegative and strictly positive reals respectively.

Proposition 6.4 ("Bernstein inequality", Wainwright (2015))

Let $(X_i)_{1\leq i\leq n}$ be a collection of independent sub-Exponential random variables s.t. $\forall i \in \{1,...,n\}$, $X_i \in \text{subExp}(\sigma_i, d_i)$ and $\mathbb{E}[X_i] = \mu_i$. The following concentration inequality holds:

$$\forall t \geq 0,\ \mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right) \leq \begin{cases} e^{-\frac{t^2}{2n\sigma^2}}, & \text{if } 0 \leq t \leq \frac{\sigma^2}{d} \\ e^{-\frac{t}{2d}}, & \text{if } t > \frac{\sigma^2}{d} \end{cases} \tag{6.14}$$

where $\sigma = \sqrt{\frac{\sum_{i=1}^{n}\sigma_i^2}{n}}$ and $d = \max_{1\leq i\leq n}\{d_i\}$.

Proposition 6.5 ("Hoeffding inequality", Wainwright (2015))

Let $(X_i)_{1\leq i\leq n}$ be a collection of independent sub-Gaussian random variables s.t. $\forall i \in \{1,...,n\}$, $X_i \in \text{subGauss}(\sigma_i)$ and $\mathbb{E}[X_i] = \mu_i$. The following concentration inequality holds:

$$\forall t \geq 0,\ \mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right) \leq e^{-\frac{t^2}{2n\sigma^2}} \tag{6.15}$$

where $\sigma = \sqrt{\frac{\sum_{i=1}^{n}\sigma_i^2}{n}}$.

The question that arises is whether the holding times τ and rewards R of $M_{\mathcal{O}}$ satisfy either Def. 6.2 or Def. 6.3 so that we can apply Prop. 6.4 or 6.5. Lem. 6.1 gives a complete answer to this question.

Lemma 6.1

The holding times τ and rewards R of $M_{\mathcal{O}}$ are sub-exponential *random variables. Moreover, the holding time of an option is* sub-Gaussian *if and only if it is almost surely bounded.*

Proof. The full proof can be found in App. D.1. We distinguish between two possible cases: either $\rho(Q_{s,o}) = 0$ (the spectral radius of $Q_{s,o}$ is 0), or $1 > \rho(Q_{s,o}) > 0$. The first case characterizes the absence of *cycles* in the abosrbing MC i.e., all states are visited *at most once* with probability 1. This means that the holding time is bounded by S almost surely and is therefore sub-Gaussian. In the second case, the absorbing MC contains cycles i.e., some states are visited at least twice with non-zero probability. The holding time is then sub-exponential but not sub-Gaussian. ■

Thanks to Lem. 6.1, we know that we can always bound τ and R using Prop. 6.4. We also know that Prop. 6.5 is useless since when τ is sub-Gaussian, it is also bounded and so we can directly apply the inequalities used in previous chapters. Brunskill and Li (2014)

addressed the problem of on-line learning with options under the assumption that τ and R are sub-Gaussian. Lem. 6.1 indicates that this assumption is *restrictive* in general, and very *loose* when it holds (bounded is preferable). Despite its simplicitly and importance, it seems that Lem. 6.1 has never been pointed out before in the literature.

6.3 Learning in Semi-Markov Decision Processes

Inspired by the mapping of Prop. 6.3, we now aim at analyzing the exploration-exploitation trade-off in an MDP with options by first analyzing that same trade-off in a generic SMDP (satisfying the assumptions of the previous section). We start by presenting the learning problem and later derive and analyze a UCRL-like learning algorithm.

6.3.1 The learning problem

To avoid any confusion, we use different notations for *time* and *decision steps*: the (possibly continuous) time elapsed is denoted by t while (discrete) decision steps will be indexed by i. At every decision step i, the learning agent is in state s_i and plays an action $a_i \in \mathcal{A}_{s_i}$. The agent then receives reward r_i and ends up in a new state s_{i+1} after a time period τ_i. For any $n \geq 1$, we denote by $T_n := \sum_{i=1}^{n} \tau_i$ the total time elapsed *before* the $n+1$-th decision step. Symmetrically, for any $t \geq 0$, we denote by $N_t := \sup\{n \in \mathbb{N},\ \sum_{i=1}^{n} \tau_i \leq t\}$ the number of decision steps that occurred *before* time t. T_n and N_t are random variables that depend on the policy being executed. The time variable t can either be an integer or a real scalar depending on the SMDP (e.g., in the SMDP $\mathcal{M}_{\mathcal{O}}$ of the previous section, t is discrete by construction).

We evaluate a learning algorithm acting in an SMDP in terms of cumulative *regret*.

Definition 6.4

For any SMDP M, any initial state distribution $\mu_1 \in \mathcal{P}(\mathcal{S})$, and any number of decision steps $n \geq 1$, let $\{\tau_i\}_{i=1}^{n}$ (resp. $\{r_i\}_{i=1}^{n}$) be the random holding times (reps. rewards) observed along the trajectory generated by a learning algorithm $\mathfrak{A}$. Let g^* be the optimal gain of M (Prop. 6.1). The cumaltive regret of $\mathfrak{A}$ after n decision steps *is defined as*

$$\Delta(M, \mathfrak{A}, \mu_1, n) := \sum_{i=1}^{n} (\tau_i \cdot g^* - r_i) = T_n \cdot g^* - \sum_{i=1}^{n} r_i. \tag{6.16}$$

The regret of $\mathfrak{A}$ after T time steps *is defined as $\Delta(M, \mathfrak{A}, \mu_1, T) := \Delta(M, \mathfrak{A}, \mu_1, N_T)$.*

Intuitively, the regret should measure the difference in cumulative reward obtained by an optimal (possibly non-stationary) policy and the learning algorithm after n decision steps (or T time steps). Def. 6.4 is *consistent* with this requirement although other definitions seem equally (if not more) relevant at first sight e.g., replacing T_n by its *expectation*.[5] In

[5] The total duration T_n after n decision steps is a random variable that depends on the algorithm $\mathfrak{A}$ just

the MDP case the optimal expected value function after T time steps v_T^* (see Sec. 2.1.2) is at most $sp(h^*)$-far from Tg^* and it makes sense to substitute v_T^* by Tg^* in the definition of the regret. In the SMDP case, v_n^* is at most $sp(h^*)$-far from $\mathbb{E}^{\pi_n^*}[T_n] \cdot g^*$ where π_n^* is the optimal (non-stationary) policy after n decision steps (note that $\mathbb{E}^{\pi_n^*}[T_n] \neq \mathbb{E}^{\mathfrak{A}}[T_n]$ in general). However, the distance between v_n^* and $T_n \cdot g^*$ is not bounded since the (random) holding times are potentially *unbounded*. To justify our definition, we first notice that in the specific case where the SMDP is an MDP, $\tau_i = 1$ for all $i \geq 1$ (i.e., actions always terminate in one step) implying that T_n and n coincides, and Def. 6.4 reduces to the standard MDP regret. This is also true if we replace T_n by $\mathbb{E}^{\pi_n^*}[T_n]$ in Eq. 6.16. But in addition to being *consistent* with Def.2.5 when all options are primitive actions, Def. 6.4 also satisfies the *compatibility condition* of Lem. 6.2.

Lemma 6.2

Let M be an MDP, $\mathcal{O}$ a set of options satisfying both Asm. 6.1 and 6.2 and $M_\mathcal{O}$ the corresponding SMDP (Prop. 6.3). For any state distribution $\mu_1 \in \mathcal{P}(\mathcal{S}_\mathcal{O})$, any learning algorithm $\mathfrak{A}$ on $M_\mathcal{O}$, and any number of decision steps n we have

$$\Delta(M, \mathfrak{A}, \mu_1, T_n) = \Delta(M_\mathcal{O}, \mathfrak{A}, \mu_1, n) + T_n \cdot \left(g_M^* - g_{M_\mathcal{O}}^*\right). \tag{6.17}$$

Proof. The proof is straightforward (Fruit and Lazaric, 2017, Lemma 2). ■

Since a learning algorithm is nothing more than a policy, any SMDP-learning algorithm $\mathfrak{A}_\mathcal{O}$ applied to $M_\mathcal{O}$ can be interpreted as a learning algorithm $\mathfrak{A}$ on M so that at each time step t, $\mathfrak{A}$ selects an action available in M based on the policy associated to the option started at decision step N_t (see Sec. 6.2.4). In Lem. 6.2, we used the same notation for both $\mathfrak{A}_\mathcal{O}$ and $\mathfrak{A}$ for simplicity. In view of Eq. 6.17, whenever $g_M^* = g_{M_\mathcal{O}}^*$ the two notions of regret for MDP with options and induced SMDP match. Moreover, as a direct consequence of Cor. 6.1, $g_M^* \geq g_{M_\mathcal{O}}^*$ and the equality holds if and only if there exists a *policy over options* that yields an optimal long-term average reward in M. A trivial example where $g_M^* = g_{M_\mathcal{O}}^*$ is when $\mathcal{A} \subseteq \mathcal{O}$ though in general, the introduction of options usually *constrains* the space of policies that can be expressed in M. The additional term $T_n \cdot \left(g_M^* - g_{M_\mathcal{O}}^*\right)$ in Eq. 6.17 corresponds to an *unavoidable approximation error*. This is similar to *supervised learning* where the true function being learned may not belong to the class considered. In the rest of this chapter, we will be focusing on minimizing the regret $\Delta(M_\mathcal{O}, \mathfrak{A}, \mu_1, n)$ which is the only part that can actually be controlled.

Brunskill and Li (2014) followed a similar approach but in the *discounted setting*: instead of directly analyzing the learning performance of an MDP with options, they analyzed the learning performance of the corresponding SMDP. Because of the discount factor, the criterion they used is not the regret but the *sample complexity* (). They also provide a definition of sample complexity for an SMDP (analogue of Def. 6.4 for the sample complexity). Unfortunately, we show in App. D.2 that unlike what the authors claim, the SMDP sample complexity bound cannot be immediately translated into a sample complexity in the origi-

like $\sum_{i=1}^n r_i$. One idea is to replace T_n by its expectation under algorithm $\mathfrak{A}$ (interpreted as a non-stationary policy) or under an optimal (possibly non-stationary) policy.

nal MDP. No analogue of Lem. 6.2 seem to exist with their definition of sample complexity. Whether the definition can be adjusted to recover the compatibility condition of Lem. 6.2 is beyond the scope of this thesis. However, this incompatibility shows the importance of carefully mapping SMDPs to MDPs with options as we did with Lem. 6.2.

6.3.2 SUCRL: Semi-Markov Upper Confidence RL

We introduce SUCRL (Alg. 12), a UCRL2-like algorithm which is able to learn in any *communicating* SMDP. The algorithm is very similar to UCRLB (Alg. 5) with few notable differences (highlightd in Alg. 12).

SUCRL requires additional inputs like $\tau_{\max}$, $\tau_{\min}$ and the *sub-exponential parameters* of the rewards and holding times. SUCRL can accomodate very tight *state-action dependent* sub-exponential parameters as well as very loose *uniform upper bounds*. The tighter the parameteres, the tighter the confidence bounds (6.18) and (6.19). As shown in Sec. 6.2.4,the rewards and holding times can sometime be *bounded* almost surely in which case we can rather use empirical Bernstein confidence bounds like in UCRLB (the bounds should be known and given as input to SUCRL instead of the sub-exponential parameters).

A key idea of the algorithm is to rely on the *transformation* introduced in Sec. 6.2.2 to deal with an extended MDP $\mathcal{M}_k^{\text{eq}}$ rather than an extended SMDP $\mathcal{M}_k$. This allows to use EVI in order to compute π_k (as in UCRLB). To construct the extended SMDP $\mathcal{M}_k$ (line 5 of Alg. 12), we enforce the additional constraint $r_{\max}\tau_k(s,a) \geq r_k(s,a)$. This guarantees that $\mathcal{M}_k^{\text{eq}}$ has a reward function bounded in $[0, r_{\max}]$ but creates a correlation between τ_k and r_k (while p_k can be computed independently from τ_k and r_k like in UCRLB). We now discuss how to implement the constraint $r_{\max}\tau_k(s,a) \geq r_k(s,a)$. For all $v \in \mathbb{R}^S$ and $s \in \mathcal{S}$, the optimal Bellman operator of $\mathcal{M}_k^{\text{eq}}$ can be written as (see Eq. 6.4)

$$\mathcal{L}_k^{\text{eq}} v(s) := \max_{a\in\mathcal{A}_s} \left\{ \max_{\substack{r\in B_r^k(s,a) \\ \tau\in B_\tau^k(s,a) \\ r\leq r_{\max}\tau}} \left\{ \frac{r}{\tau} + \frac{\alpha}{\tau}\left(\max_{p\in B_p(s,a)} \{p^\intercal v\} - v(s) \right) \right\} \right\} + v(s). \tag{6.24}$$

The maximization over r and τ in (6.24) takes the following form:

$$\max_{\substack{r\in[r^-,r^+] \\ \tau\in[\tau^-,\tau^+] \\ r\leq r_{\max}\tau}} \left\{ \frac{r+c}{\tau} \right\} \tag{6.25}$$

where c is a scalar which can be positive, negative or null.[6] For (6.25) to admit a solution

[6] In Eq. 6.24, $c = \alpha\left(\max_{p\in B_p(s,a)} \{p^\intercal v\} - v(s)\right)$.

Algorithm 12 SUCRL

Input: Confidence $\delta \in]0,1[$, maximal holding time $\tau_{\max}$ and per-time step reward $r_{\max}$, minimal holding time $\tau_{\min}$, set of states $\mathcal{S}$, set of actions $\mathcal{A}$, sub-exponential parameters $\sigma_r(s,a)$, $d_r(s,a)$, $\sigma_\tau(s,a)$ and $d_\tau(s,a)$

Set initial decision step $i := 1$, observe initial state s_1 and initialize for all $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$: counters $N_1(s,a,s') := 0$ and $N_1(s,a) := 0$, empirical averages $\widehat{p}_1(s'|s,a) := 0$, $\widehat{r}_1(s,a) := 0$ and $\widehat{\tau}_k(s,a) := 0$, sample variances $\widehat{\sigma}^2_{p,1}(s'|s,a) := 0$.

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

  Set the starting step of the episode $i_k := i$ and initialize for all $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$: episode counters $\nu_k(s,a,s') := 0$ and $\nu_k(s,a) := 0$, and cumulative rewards $R_k(s,a) := 0$ and holding times $T_k(s,a) := 0$. ▷ Initialization of episode k

  For all $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, compute upper confidence bounds:

$$\beta^{sa}_{r,k} := 2\sigma_r(s,a)\sqrt{\frac{\ln\left(7SAN^+_k(s,a)/\delta\right)}{N^+_k(s,a)}} + 4d_r(s,a)\frac{\ln\left(7SAN^+_k(s,a)/\delta\right)}{N^+_k(s,a)} \tag{6.18}$$

$$\beta^{sa}_{\tau,k} := 2\sigma_\tau(s,a)\sqrt{\frac{\ln\left(7SAN^+_k(s,a)/\delta\right)}{N^+_k(s,a)}} + 4d_\tau(s,a)\frac{\ln\left(7SAN^+_k(s,a)/\delta\right)}{N^+_k(s,a)} \tag{6.19}$$

  Set $\mathcal{M}_k := \{\mathcal{S}, \mathcal{A}, r_k, p_k\}$ to be the extended SMDP defined by the confidence intervals $p_k(s'|s,a) \in B^k_p(s,a,s')$ (see Eq. 3.3),

$$r_k(s,a) \in B^k_r(s,a) := \left[\widehat{r}_k(s,a) - \beta^{sa}_{r,k}, \widehat{r}_k(s,a) + \beta^{sa}_{r,k}\right] \cap [0, r_{\max}\tau_{\max}] \tag{6.20}$$

$$\tau_k(s,a) \in B^k_\tau(s,a) := \left[\widehat{\tau}_k(s,a) - \beta^{sa}_{\tau,k}, \widehat{\tau}_k(s,a) + \beta^{sa}_{\tau,k}\right] \cap [\tau_{\min}, \tau_{\max}] \tag{6.21}$$

  and the additional constraint $\tau_k(s,a) \geq r_k(s,a)/r_{\max}$.

  Compute policy π_k using (extended) value iteration on the extended MDP $\mathcal{M}^{\text{eq}}_k$ obtained by *uniformization* of $\mathcal{M}_k$ (see Sec. 6.2.2)

$$(g_k, h_k, \pi_k) := \text{EVI}\left(\mathcal{L}^{\text{eq}}_k, \mathcal{G}^{\text{eq}}_k, \frac{r_{\max}}{\tau_{\max} i_k}, s_1, 1\right) \tag{6.22}$$

  Sample action $a_i \sim \pi_k(\cdot|s_i)$.

  while True do ▷ Execute policy π_k until the end of episode k

    Execute action a_i, obtain reward r_i, and observe duration τ_i and next state s_{i+1}.

    Increment episode counters:

$$\nu_k(s_i,a_i,s_{i+1}) \leftarrow \nu_k(s_i,a_i,s_{i+1}) + 1 \quad \text{and} \quad \nu_k(s_i,a_i) \leftarrow \nu_k(s_i,a_i) + 1$$

    Increment cumulative reward and *holding time*

$$R_k(s_t,a_t) \leftarrow R_k(s_i,a_i) + r_i \quad \text{and} \quad T_k(s_i,a_i) \leftarrow T_k(s_i,a_i) + \tau_i$$

    if $\nu_k(s_i,a_i) \geq N^+_k(s_i,a_i)$ **then** ▷ Stopping condition of episode k

      Increment time $i \leftarrow i+1$ and **Break**

    else

      Increment time $i \leftarrow i+1$ and sample action $a_i \sim \pi_k(\cdot|s_i)$.

    end if

  end while

  Update counters (see Eq. 3.6), empirical averages and sample variances for all $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$ (see Eq. 3.8 for the rewards and Eq. 3.9 for the transition probabilities)

$$\widehat{\tau}_{k+1}(s,a) := \frac{N_k(s,a)}{N^+_{k+1}(s,a)} \cdot \widehat{\tau}_k(s,a) + \frac{T_k(s,a)}{N^+_{k+1}(s,a)} \tag{6.23}$$

end for

we need to assume that $r^- \leq r_{\max}\tau^+$.[7]

Lemma 6.3

The pair $(r^\star, \tau^\star) \in [r^-, r^+] \times [\tau^-, \tau^+]$ *defined as follows is a solution to* (6.25)*:*

$$r^\star := \begin{cases} \min\{r^+, r_{\max}\tau^+\} & \text{if } c \leq 0 \\ \max\{\min\{r^+, r_{\max}\tau^-\}, r^-\} & \text{if } c > 0 \end{cases} \quad \text{and} \quad \tau^\star := \begin{cases} \max\left\{\tau^-, \frac{r^\star}{r_{\max}}\right\} & \text{if } r^\star + c > 0 \\ \tau^+ & \text{if } r^\star + c \leq 0 \end{cases}.$$

Proof. Since $r^- \leq r_{\max}\tau^+$ by assumption, it is clear that $(r^\star, \tau^\star) \in [r^-, r^+] \times [\tau^-, \tau^+]$. If $r \geq r_{\max}\tau^+$ it is obvious that there exists no $\tau \in [\tau^-, \tau^+]$ such that $r \leq r_{\max}\tau$ and so we can restrict attention to $[r^-, \min\{r^+, r_{\max}\tau^+\}]$. Note that $r^\star$ belongs to this interval by definition. For any *fixed* $r \in [r^-, \min\{r^+, r_{\max}\tau^+\}]$, consider the problem

$$\max_{\substack{\tau \in [\tau^-, \tau^+] \\ r \leq r_{\max}\tau}} \left\{\frac{r+c}{\tau}\right\}. \tag{6.26}$$

If $r + c > 0$, the maximizer $\tau^\star(r)$ of (6.26) is given by $\tau^\star(r) = \max\left\{\tau^-, \frac{r}{r_{\max}}\right\}$ while if $r + c \leq 0$, $\tau^\star(r) = \tau^+$. Since this is true for all $r \in [r^-, \min\{r^+, r_{\max}\tau^+\}]$, if we show that $r^\star$ is an optimal value for r in (6.25) then $\tau^\star = \tau^\star(r^\star)$ is an optimal value for τ.

Consider the function $f_c : r \mapsto \frac{r+c}{\tau^\star(r)}$. By construction, any maximizer of f_c gives an optimal value for r in (6.25). Plugging the expression of $\tau^*(r)$ we obtain:

$$f_c(r) = \begin{cases} \frac{r+c}{\tau^+} \text{ if } r \leq -c \\ \frac{r+c}{\tau^-} \text{ if } -c < r < r_{\max}\tau^- \\ r_{\max}\left(1 + \frac{c}{r}\right) \text{ if } r \geq r_{\max}\tau^- \text{ and } r > -c \end{cases}. \tag{6.27}$$

No matter whether $c > r_{\max}\tau^-$ or $c \leq r_{\max}\tau^-$, the function f_c is *continuous* with $f_c(-c) = 0$ and $f_c(r_{\max}\tau^-) = r_{\max} + \frac{c}{\tau^-}$. If $c \leq 0$, f_c is *increasing* on every separate interval and so f_c is also "globally increasing" (by continuity), implying that the maximum of f_c is reached for $r = \min\{r^+, r_{\max}\tau^+\}$. If $c > 0$, then necessarily $c < r_{\max}\tau^-$ since $\tau^- > 0$, and f_c is increasing for $r \leq r_{\max}\tau^-$ and decreasing for $r \geq r_{\max}\tau^-$. As a consequence, if $r^- \leq r_{\max}\tau^-$ then the maximizer of f_c is $r = \min\{r^+, r_{\max}\tau^-\}$, otherwise it is $r = r^-$. This concludes the proof. ∎

Using Lem. 6.3, it is very easy to compute $\mathcal{L}_k^{\text{eq}} v$ (see Eq. 6.24). Moreover, r_k and τ_k can only take *finitely many* possible values (the set of possible values does not depend on v). Therefore, $\mathcal{M}_k^{\text{eq}}$ can be expressed as a discrete MDP just like the extended MDP used in UCRL2. In addition, $\mathcal{M}_k$ is communicating (same argument as in Sec. 3.1.2) and so is $\mathcal{M}_k^{\text{eq}}$ (the transformation preserves the chain structure) implying *existence* of a solution to the Bellman optimality equation and *convergence* of (extended) value iteration in $\mathcal{M}_k^{\text{eq}}$ (Prop. 2.4 and 2.6). Using Prop. 6.2, any optimality equation in $\mathcal{M}_k$ can be converted into an optimality

[7] This assumption is always satisfied as soon as there exists $r \in B_r^k(s,a)$ and $\tau \in B_\tau^k(s,a)$ such that $r \leq r_{\max}\tau$. With high probability, $B_r^k(s,a)$ and $B_\tau^k(s,a)$ contain the true expectated values $r(s,a)$ and $\tau(s,a)$, and $r(s,a) \leq r_{\max}\tau(s,a)$ by assumption.

equation in $\mathcal{M}_k$ (and so existence in $\mathcal{M}_k$ is guaranteed as well). Finally, the outputs of EVI (Eq. 6.22) satisfy the following inequality (Lem. 2.7):

$$\forall s \in \mathcal{S}, \quad \left| \frac{\sum_{a \in \mathcal{A}_s} \pi_k(a|s) \Big(r_k(s,a) + \alpha\big(p_k(\cdot|s,a)^\intercal h_k - h_k(s)\big)\Big)}{\sum_{b \in \mathcal{A}_s} \pi_k(b|s) \tau_k(s,b)} - g_k \right| \leq \frac{r_{\max}}{\tau_{\max} i_k}.$$

After multiplying both side of the above inequality by the expected holding time and using the fact that $\tau_k(s,a) \leq \tau_{\max}$ for all state-action pairs we obtain an inequality similar to (3.41):

$$\forall s \in \mathcal{S}, \quad \sum_{a \in \mathcal{A}_s} \pi_k(a|s) \Big| \tau(s,a) g_k - r_k(s,a) - \alpha\big(p_k(\cdot|s,a)^\intercal h_k - h_k(s)\big)\Big| \leq \frac{r_{\max}}{i_k}. \tag{6.28}$$

The rest of Alg. 12 is pretty standard in comparison with previous chapters.

6.3.3 Regret guarantees of SUCRL

To simplify the regret analysis we define $\sigma_\tau := \max_{s,a} \{\sigma_\tau(s,a)\}$ and $d_\tau := \max_{s,a} \{d_\tau(s,a)\}$ the maximal sub-exponential parameters given as *inputs* of SUCRL (and we define similarly σ_r and b_r). For any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, the support of $p(\cdot|s,a)$ is still denoted $\Gamma(s,a)$. We also need to extend the concepts of diameter and travel-budget to SMDPs. Unsurprisingly, Def. 6.5 and 6.6 almost match the definitions of Sec. 3.3 with the presence of holding times.

Definition 6.5

If $\mathbb{E}^\pi[\cdot|s_1 = s]$ denotes the expectation under policy π starting from s in SMDP M, the diameter *of M is defined as*

$$D := \max_{s,s'} \min_{\pi \in \Pi^{SD}} \mathbb{E}^\pi \left[\sum_{i=1}^{\nu(s')-1} \tau(s_i, a_i) \middle| s_1 = s \right] \tag{6.29}$$

where $\nu(s') := \inf\{n \geq 1 : \ s_n = s'\}$.

The diameter is defined in terms of actual expected *time* (to reach a state starting from another state) rather than expected number of *decision steps*. Like in the MDP case, $D < +\infty$ if and only if M is communicating. Moreover, $D := \max_s \|h^*_{\mapsto s}\|_\infty$ where $h^*_{\mapsto s}$ is the maximal non-positive fixed point of the Bellman shortest path operator $L_{\mapsto s}$ in *MDP* $M' = \langle \mathcal{S}, \mathcal{A}, r', p \rangle$ where $r' = -r_{\max}\tau \leq 0$ (note that M' is an MDP and not an SMDP).

Definition 6.6

The travel-budget *of SMDP M is defined as*

$$\Lambda := \max_{s,s'} \min_{\pi \in \Pi^{SD}} \mathbb{E}^\pi \left[\sum_{i=1}^{\nu(s')-1} r_{\max}\tau(s_i,a_i) - r(s_i,a_i) \middle| s_1 = s \right]. \tag{6.30}$$

Like in the MDP case, $0 \leq \Lambda \leq r_{\max} D$. Similarly to D, $\Lambda := \max_s \|h^*_{\mapsto s}\|_\infty$ where $h^*_{\mapsto s}$ is

the maximal non-positive fixed point of the Bellman shortest path operator $L_{\mapsto s}$ in *MDP* $M' = \langle \mathcal{S}, \mathcal{A}, r', p \rangle$ where $r' = r - r_{\max}\tau$ (which is negative by assumption).

We now present two regret bounds similar to Thm. 3.4 and 3.5 (the main differences are highlighted).

Theorem 6.1 (Analogue of Thm. 3.4)

There exists a numerical constant $\beta > 0$ such that for any *communicating SMDP M, with probability at least $1-\delta$, it holds that for all initial state distributions $\mu_1 \in \Delta_S$ and for all $n > 1$:*

$$\begin{aligned}\Delta(M, \text{SUCRL}, \mu_1, n) \leq\ & \beta \cdot \left(\max\{r_{\max}, \Lambda\} \sqrt{\left(\sum_{s,a} \Gamma(s,a)\right)} \right. \\ & \left. + (r_{\max}\sigma_\tau + \sigma_r)\sqrt{SA} + r_{\max}\tau_{\max} \right) \sqrt{n \ln\left(\frac{n}{\delta}\right)} \\ & + \beta \cdot \left(\max\{r_{\max}, \Lambda\} S + r_{\max} d_\tau + d_r \right) SA \ln\left(\frac{n}{\delta}\right) \ln(n)\,.\end{aligned} \tag{6.31}$$

Theorem 6.2 (Analogue of Thm. 3.5)

There exists a numerical constant $\beta > 0$ such that for any *communicating SMDP M, with probability at least $1-\delta$, it holds that for all initial state distributions $\mu_1 \in \Delta_S$ and for all $n > 1$:*

$$\begin{aligned}\Delta(M, \text{SUCRL}, \mu_1, n) \leq\ & \beta \cdot \left(\left\{ r_{\max}, \sqrt{r_{\max}\Lambda} \right\} \sqrt{\left(\sum_{s,a} \Gamma(s,a)\right) \ln(n)} \right. \\ & \left. + (r_{\max}\sigma_\tau + \sigma_r)\sqrt{SA} + r_{\max}\tau_{\max} \right) \sqrt{n \ln\left(\frac{n}{\delta}\right)} \\ & + \beta \cdot \left(\max\left\{ r_{\max}, \frac{\Lambda^2}{r_{\max}} \right\} S + r_{\max} d_\tau + d_r \right) SA \ln\left(\frac{n}{\delta}\right) \ln(n)\,.\end{aligned} \tag{6.32}$$

In the special case where options are almost surely *bounded* (see Lem. 6.1) by a *known* upper-bound $t_{\max}$, the main terms in the bounds of Thm. 6.1 and 6.2 remain unchanged but $d_r = d_\tau = 0$ and $\sigma_r = r_{\max}\sigma_\tau = r_{\max}t_{\max}$.

6.3.4 Regret analysis of SUCRL

Gain optimism

According to Prop. 6.2, the optimal gains of the true SMDP M and the MDP M_{eq} obtained by uniformization (Eq. 6.4) are equal i.e., $g^* = g^*_{\text{eq}}$. We now show that with high probability, $g^*_{k,\text{eq}} \geq g^*_{\text{eq}}$ where $g^*_{k,\text{eq}}$ is the optimal gain of $\mathcal{M}^{\text{eq}}_k$. We first derive a slightly looser version of

concentration inequality (6.14).

Corollary 6.2

Let $(X_i)_{1\leq i\leq n}$ be a collection of sub-Exponential random variables satisfying the same assumptions as in Prop. 6.4. For all $n\geq 1$ and $\delta\in]0,1[$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i-\mu_i\right|\geq\sqrt{2\left(\sum_{i=1}^n\sigma_i^2\right)\ln\left(\frac{2}{\delta}\right)}+2d\ln\left(\frac{2}{\delta}\right)\right)\leq\delta \tag{6.33}$$

.

Proof. We recall that $d:=\max_{1\leq i\leq n}\{d_i\}$.
If $\sum_{i=1}^n\sigma_i^2\geq 2d^2\ln\left(\frac{2}{\delta}\right)$ we set $t:=\sqrt{2\left(\sum_{i=1}^n\sigma_i^2\right)\ln\left(\frac{2}{\delta}\right)}\leq\sum_{i=1}^n\sigma_i^2/d$ and so the first inequality in Eq. 6.14 of Prop. 6.4 holds implying that $\mathbb{P}\left(|\sum_{i=1}^n X_i-\mu_i|\geq\sqrt{2\left(\sum_{i=1}^n\sigma_i^2\right)\ln\left(\frac{2}{\delta}\right)}\right)\leq\delta$.
If on the other hand $\sum_{i=1}^n\sigma_i^2<2d^2\ln\left(\frac{2}{\delta}\right)$ we set $t:=2d\ln\left(\frac{2}{\delta}\right)>\sum_{i=1}^n\sigma_i^2/d$ and so the second inequality in Eq. 6.14 of Prop. 6.4 holds implying that $\mathbb{P}\left(|\sum_{i=1}^n X_i-\mu_i|\geq 2d\ln\left(\frac{2}{\delta}\right)\right)\leq\delta$.
In conclusion, Eq. 6.14 holds for all $n\geq 1$. ■

Theorem 6.3 (Analogue of Thm. 3.1)

he probability that there exists $n\geq 1$ and $k\geq 1$ s.t. M_{eq} does not belong to the extended MDP $\mathcal{M}_k^{eq}$ is at most $\frac{\delta}{3}$, that is

$$\mathbb{P}\left(\exists n\geq 1,\exists k\geq 1,\ s.t.\ M_{eq}\notin\mathcal{M}_k^{eq}\right)\leq\frac{\delta}{3}.$$

Proof. The proof is almost identical to the proof of Thm. 6.3 but we have to account for the possibility that $\tau(s,a)\notin B_\tau^k(s,a)$. We use Cor. 6.2 with $\delta\leftarrow\frac{\delta}{20SA(N_k^+(s,a))^2}$ for both r and τ. We notice that $\ln\left(\frac{40SA(N_k^+(s,a))^2}{\delta}\right)\leq 2\ln\left(\frac{7SAN_k^+(s,a)}{\delta}\right)$ and after taking a union bound we obtain:

$$\begin{aligned}\mathbb{P}\left(\exists n\geq 1,\exists k\geq 1,\ \text{s.t. } M_{\text{eq}}\notin\mathcal{M}_k^{\text{eq}}\right)&\leq\sum_{s,a}\sum_{n=1}^{+\infty}\left(\frac{\delta}{20n^2SA}+\frac{\delta}{20n^2SA}+\sum_{s'}\frac{\delta}{10n^2S^2A}\right)=\frac{2\pi^2\delta}{60}\\&\leq\frac{\delta}{3}.\end{aligned}$$

■

As a direct consequence of Thm. 6.3, with probability at least $1-\frac{\delta}{3}$, for all $k\geq 1$, $\mathcal{L}_k^{\text{eq}}h_{\text{eq}}^*\geq L_{\text{eq}}h_{\text{eq}}^*$ and so $g_{k,\text{eq}}^*\geq g_{\text{eq}}^*$ (Prop. 3.3) and moreover $g_k\geq g_{k,\text{eq}}^*-\frac{r_{\max}}{2i_k}$ (Prop. 2.7) implying $g_k\geq g^*-\frac{r_{\max}}{2i_k}$.

Range of the optimistic bias

Under the same high probability event as Thm. 6.3, $\mathcal{L}^{\text{eq}}_{k,\mapsto s} h^{\text{eq}}_{\mapsto s} \geq L^{\text{eq}}_{\mapsto s} h^{\text{eq}}_{\mapsto s} = h^{\text{eq}}_{\mapsto s}$ for all $k \geq 1$ and all $s \in \mathcal{S}$ where $L^{\text{eq}}_{\mapsto s}$ and $\mathcal{L}^{\text{eq}}_{k,\mapsto s}$ are the Belmman shortest path operators of M_{eq} and $\mathcal{M}^{\text{eq}}_k$ with rewards r/τ (see (6.4)) replaced by $r/\tau - r_{\max} \leq 0$, and where $h^{\text{eq}}_{\mapsto s} \leq 0$ is the maximal non-positive fixed point of $L^{\text{eq}}_{\mapsto s}$. Due to Prop. 3.5, $\Lambda^{\text{eq}}_k \leq \Lambda_{\text{eq}}$ and due to Thm. 3.35, $sp\left(h_k\right) \leq \Lambda^{\text{eq}}_k$ implying $sp\left(h_k\right) \leq \Lambda_{\text{eq}}$. It remains to relate Λ_{eq} and Λ.

Theorem 6.4

For all $0 < \alpha < \tau_{\min}$, it holds that $\Lambda_{eq} \leq \Lambda/\alpha$.

Proof. By definition, $\Lambda_{\text{eq}} = \max_s \|h^{\text{eq}}_{\mapsto s}\|_\infty$ and $\Lambda = \max_s \|h^*_{\mapsto s}\|_\infty$ where $h^{\text{eq}}_{\mapsto s}$ is the maximal non-positive fixed point of $L^{\text{eq}}_{\mapsto s}$ and $h^*_{\mapsto s}$ the maximal non-positive fixed point of $L_{\mapsto s}$. As shown in the proof of Prop. 2.8 (Sec. 2.1.4), $L_{\mapsto s}$ and $L^{\text{eq}}_{\mapsto s}$ are the Bellman operators of the modified MDPs $M_{\mapsto s}$ and $M^{\text{eq}}_{\mapsto s}$ respectively, where s is an absorbing state with reward zero (the optimal gains of $M_{\mapsto s}$ and $M^{\text{eq}}_{\mapsto s}$ are zero). Prop. 6.2 implies that $\alpha h^{\text{eq}}_{\mapsto s}$ is a fixed point of $L_{\mapsto s}$ and moreover $\alpha h^{\text{eq}}_{\mapsto s} \leq 0$ since $h^{\text{eq}}_{\mapsto s} \leq 0$ and $\alpha > 0$. Since $h^*_{\mapsto s}$ is the maximal non-positive fixed point of $L_{\mapsto s}$, necessarily $h^*_{\mapsto s} \geq \alpha h^{\text{eq}}_{\mapsto s}$ which concludes the proof. ■

In conclusion, we obtain the same bound as in the proof of UCRLB (see Sec. 3.3) i.e., $sp\left(h_k\right) \leq \Lambda/\alpha$. Thm. 6.4 actually provides a tight bound since the equality holds $\Lambda_{\text{eq}} = \Lambda/\alpha$ (we omit the proof since this result is never needed to bound the regret, the interested reader may refer to Thm. 2.1 for an analogy). Note that in Sec. 3.3, α denotes the *aperiodicity coefficient* while here it corresponds to the *uniformization coefficient*. Both coefficients play a similar role and have *no impact* on the regret analysis (they eventually cancel).

Splitting into episodes

Using MDS concentration inequalities for sub-exponential r.v. (Prop. 6.6 below), we substitute the sampled holding times τ_i and rewards r_i (appearing in the definition of the regret) by their expectations.

Proposition 6.6 (Theorem 2.3. Wainwright (2015))

Let $(X_n, \mathcal{F}_n)_{n\in\mathbb{N}}$ be an MDS such that $\mathbb{E}\left[e^{\lambda X_n}\middle|\mathcal{F}_{n-1}\right] \leq e^{\frac{\sigma_n^2\lambda^2}{2}}$ a.s. for any $|\lambda| < 1/d_n$ and $n \in \mathbb{N}$. For all $n \geq 1$ and $t \geq 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq \sqrt{2\left(\sum_{i=1}^n \sigma_i^2\right)\ln\left(\frac{2}{\delta}\right)} + 2\max_{1\leq i\leq n}\{d_i\}\ln\left(\frac{2}{\delta}\right)\right) \leq \delta. \tag{6.34}$$

.

We leverage Prop. 6.6 to bound the sum of rewards and holding times as in Lem. 3.1.

Lemma 6.4

With probability at least $1-\frac{\delta}{6}$*:*

$$\begin{aligned}\forall n\geq 1,\quad -\sum_{i=1}^{n} r_i &\leq -\sum_{i=1}^{n}\sum_{a\in\mathcal{A}_{s_i}} \pi_{k_i}(a|s_i) r(s_i,a) + 2\sigma_r\sqrt{n\ln\left(\frac{6n}{\delta}\right)} + 4d_r\ln\left(\frac{6n}{\delta}\right)\\ \sum_{i=1}^{n}\tau_i &\leq \sum_{i=1}^{n}\sum_{a\in\mathcal{A}_{s_i}} \pi_{k_i}(a|s_i)\tau(s_i,a) + 2\sigma_\tau\sqrt{n\ln\left(\frac{6n}{\delta}\right)} + 4d_\tau\ln\left(\frac{6n}{\delta}\right)\end{aligned} \tag{6.35}$$

.

Proof. We use a martingale argument and Cor. 6.2. ■

Since the rewards r and holding times τ of M satisfy $0\leq r\leq r_{\max}\tau$ (by assumption), the rewards r_{eq} of M_{eq} satisfy $0\leq r_{\text{eq}}\leq r_{\max}$ and consequently $0\leq g^* = g^*_{\text{eq}}\leq r_{\max}$. We can now decompose the regret of SUCRL as we did with the regret of UCRLB in (3.39):

$$\Delta(\text{SUCRL}, n) \leq \sum_{k=1}^{k_n}\Delta_k + 2\left(r_{\max}\sigma_\tau+\sigma_r\right)\sqrt{n\ln\left(\frac{6n}{\delta}\right)} + 4\left(d_r + r_{\max}d_\tau\right)\ln\left(\frac{6n}{\delta}\right), \tag{6.36}$$

where the per-episode regret is now defined as $\Delta_k := \sum_{s,a}\nu_k(s)\pi_k(s,a)\Big(\tau(s,a)g^* - r(s,a)\Big)$. We then introduce r_k and τ_k:

$$\tau(s,a)g^* - r(s,a) = \tau_k(s,a)g^* - r_k(s,a) + \big(\tau_k(s,a)-\tau(s,a)\big)\underbrace{g^*}_{\leq r_{\max}} + \big(r_k(s,a)-r(s,a)\big)$$

By analogy with Sec. 3.5, we define $\Delta_k^r := \sum_{s,a}\nu_k(s)\pi_k(a|s)\Big(r_k(s,a)-r(s,a)\Big)$ and $\Delta_k^{r1} := \sum_{s,a}\nu_k(s,a)\left(r_k(s,a)-r(s,a)\right)$ (similar to Δ_k^r with $\nu_k(s)\pi_k(s,a)$ replaced by $\nu_k(s,a)$). We proceed similarly to define Δ_k^τ and $\Delta_k^{\tau 1}$.

Lemma 6.5 (Analogue to Lem. 3.4)

With probability at least $1-\frac{\delta}{6}$*:*

$$\begin{aligned}\forall n\geq 1,\quad \sum_{k=1}^{k_n}\Delta_k^r &\leq \sum_{k=1}^{k_n}\Delta_k^{r1} + 4r_{\max}\tau_{\max}\sqrt{n\ln\left(\frac{5n}{\delta}\right)}\\ \sum_{k=1}^{k_n}\Delta_k^\tau &\leq \sum_{k=1}^{k_n}\Delta_k^{\tau 1} + 4\tau_{\max}\sqrt{n\ln\left(\frac{5n}{\delta}\right)}\end{aligned}$$

Proof. We use a martingale argument and Prop. 3.7. ■

We then bound Δ_k^{r1} and $\Delta_k^{\tau 1}$:

$$g^*\Delta_k^{\tau 1} + \Delta_k^{r1} \leq \sum_{s,a}\nu_k(s,a)\left(2(\sigma_r+r_{\max}\sigma_\tau)\sqrt{\frac{\ln\left(7SAn/\delta\right)}{N_k^+(s,a)}} + 4(d_r+r_{\max}d_\tau)\frac{\ln\left(7SAn/\delta\right)}{N_k^+(s,a)}\right)$$

Using Lem. 3.6, we obtain (the inequality should be interpreted up to multiplicative numerical constants):

$$\sum_{k=1}^{k_n} g^* \Delta_k^{\tau 1} + \Delta_k^{r1} \lesssim (\sigma_r + r_{\max}\sigma_\tau)\sqrt{SAn \ln\left(\frac{n}{\delta}\right)} + (d_r + r_{\max} d_\tau) SA \ln\left(\frac{n}{\delta}\right) \ln(n)\,.$$

To bound $\tau_k(s,a)g^* - r_k(s,a)$ we use gain optimism $g_k \geq g^* - \frac{r_{\max}}{2i_k}$ and Eq. 6.28 so that:

$$\sum_{a\in\mathcal{A}_s} \pi_k(a|s)\left(\tau_k(s,a)g^* - r_k(s,a)\right) \leq \alpha \sum_{a\in\mathcal{A}_s} \pi_k(a|s)(p_k(\cdot|s,a)^\intercal h_k - h_k(s)) + \frac{3r_{\max}}{2i_k}. \quad (6.37)$$

We recover the exact same term Δ_k^p as in Eq. 3.42. The same analysis as in Sec. 3.5.3 (Thm. 6.1) and Sec. 3.6 can be carried out (with i and n replacing t and T) leading to the same regret bounds.

6.3.5 Minimax lower bound for SMDPs

We have already seen that UCRLB achieves rather *tight* regret guarantees (in a minimax sense). Due to the similarities in the regret analysis of SUCRL and UCRLB, we can expect the bounds of Thm. 6.1 and 6.2 to be as tight. This is confirmed in the following *lower bound*.

Theorem 6.5

There exists a constant $\beta > 0$ such that for any algorithm $\mathfrak{A}$, any integers $S, A \geq 10$, any reals $t_{\max} \geq 3t_{\min} \geq 3$, $r_{\max} > 0$, $\Lambda > r_{\max} \cdot \max\{20t_{\min}\log_A(S), 12t_{\min}\}$, and for $n \geq \max\{\Lambda/r_{\max}, t_{\max}\}SA$, there is an SMDP M with at most S states, A actions, and travel-budget Λ, with holding times in $[t_{\min}, t_{\max}]$ and rewards in $\left[0, \frac{1}{2}r_{\max}t_{\max}\right]$ satisfying $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}_s$, $r(s,a) \leq r_{\max}\tau(s,a)$, such that for any initial distribution $\mu_1 \in \Delta_S$, the expected regret of $\mathfrak{A}$ after n decision steps is lower-bounded by:

$$\mathbb{E}\left[\Delta(M, \mathfrak{A}, \mu_1, n)\right] \geq \beta \cdot \left(\left(\sqrt{r_{\max}\Lambda} + r_{\max}\sqrt{t_{\max}}\right)\sqrt{SAn}\right).$$

Proof. The proof (Fruit and Lazaric, 2017, Appendix C) is based on (Jaksch et al., 2010, Section 6) but it requires to perturb transition probabilities and rewards at the same time to create a family of SMDPs with different optimal policies that are difficult to discriminate. The contributions of the two perturbations can be made independent. More precisely, the lower bound is obtained by designing SMDPs where learning to distinguish between "good" and "bad" transition probabilities and learning to distinguish between "good" and "bad" rewards are two independent problems, leading to two additive terms $\sqrt{r_{\max}\Lambda}$ and $r_{\max}\sqrt{t_{\max}}$ in the lower bound. ■

This lower bound reveals a *gap* with the upper bound of order $\sqrt{\Gamma}$ on the first term (similar to UCRLB) and $\sqrt{t_{\max}}$ on the second term. While closing this gap remains a challenging open question, it is a problem beyond the scope of this thesis.

Thm. 6.5 may not be very relevant for MDPs with options since the resulting SMDPs only

account for a *strict subset* of all possible SMDPs. The rewards and holding times of such SMDPs are always *correlated* due to the inner Markov structure of options. This is not the case for all SMDPs. Actually, the specific family of SMDPs constructed to prove Thm. 6.5 cannot be mapped to any MDP with options for that reason. Nevertheless, we show that a similar lower bound also holds for SMDPs resulting from MDPs with options.

Theorem 6.6

There exists a constant $\beta > 0$ such that for any algorithm $\mathfrak{A}$, any integers $S, A \geq 10$, any reals $t_{\max} \geq 3t_{\min} \geq 3$, $r_{\max} > 0$, $\Lambda > r_{\max} \cdot \max\{20t_{\min}\log_A(S), 12t_{\min}\}$, and for $n \geq \max\{\Lambda/r_{\max}, t_{\max}\}SA$, there is an SMDP M resulting from an MDP with options with at most S states, A actions, and travel-budget Λ, with holding times in $[t_{\min}, t_{\max}]$ and rewards in $\left[0, \frac{1}{2}r_{\max}t_{\max}\right]$ satisfying $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}_s$, $r(s,a) \leq r_{\max}\tau(s,a)$, such that for any initial distribution $\mu_1 \in \Delta_S$:

$$\mathbb{E}\left[\Delta(M, \mathfrak{A}, \mu_1, n)\right] \geq \beta \cdot \left(\left(\sqrt{r_{\max}\Lambda} + r_{\max}\sqrt{t_{\max} - t_{\min}}\right)\sqrt{SAn}\right).$$

Proof. See (Fruit and Lazaric, 2017, Appendix C). ■

6.3.6 Analyzing the impact of options on the learning process

We are now ready to proceed with the comparison of the bounds on the regret of learning with *options* versus *primitive actions*. To facilitate the comparison, we ignore all logarithmic terms and assume that all options are almost surely bounded by $t_{\max}$. We recall that the regret of UCRLB is of order $\Delta(\text{UCRLB}, T_n) = \widetilde{\mathcal{O}}\left(\sqrt{r_{\max}\Lambda\Gamma SAT_n}\right)$. In contrast, SUCRL achieves $\Delta(\text{SUCRL}, T_n) = \widetilde{\mathcal{O}}\left(\left(\sqrt{r_{\max}\Lambda_{\mathcal{O}}\Gamma_{\mathcal{O}}} + r_{\max}t_{\max}\right)\sqrt{S_{\mathcal{O}}On} + T_n \cdot (g^* - g^*_{\mathcal{O}})\right)$. We first notice that since $\mathcal{S}_{\mathcal{O}} \subseteq \mathcal{S}$ we have that $S_{\mathcal{O}} \leq S$. Furthermore, we introduce the simplifying conditions $g^* = g^*_{\mathcal{O}}$ (i.e., the options do not prevent from learning the optimal policy).

While in general comparing upper bounds is potentially loose, we notice that both upper-bounds are derived using *similar techniques* and thus they would be "similarly" loose and they both have *almost matching* worst-case lower bounds. Let $\mathcal{R}(n)$ denote the *ratio* between the regret upper bounds of SUCRL using options $\mathcal{O}$ and UCRLB. Up to numerical constants we have

$$\mathcal{R}(n) \lesssim \frac{(\sqrt{r_{\max}\Lambda_{\mathcal{O}}\Gamma_{\mathcal{O}}} + r_{\max}t_{\max})\sqrt{S_{\mathcal{O}}On}}{\sqrt{r_{\max}\Lambda\Gamma SAT_n}} = \left(\sqrt{\frac{\Lambda_{\mathcal{O}}\Gamma_{\mathcal{O}}}{\Lambda\Gamma}} + \frac{r_{\max}t_{\max}}{\sqrt{r_{\max}\Lambda\Gamma}}\right)\sqrt{\frac{S_{\mathcal{O}}On}{SAT_n}}. \quad (6.38)$$

$\mathcal{R}(n) \leq 1$ indicates that using options is potentially *beneficial* (compared to using primitive actions).

Eq. 6.38 reveals that options can improve the learning speed by reducing the size of the *support* Γ of the dynamics of the environment, for example when options are designed so as to reach a specific goal (very *"sparse"* transition dynamics). This potential advantage matches the intuition on "good" options often presented in the literature (see e.g., the concept of *"funnel"* actions introduced by Dietterich (2000)). However, Γ is absent from the lower

bounds which raises the question whether reducing the size of the support is an actual source of improvement. On the other hand, both upper-bounds and lower-bounds suggest that designing options which reduce the *travel-budget* Λ will have a positive effect on the learning performance. When $\mathcal{S}_{\mathcal{O}} = \mathcal{S}$ and $\mathcal{A} \subseteq \mathcal{O}$, $\Lambda_{\mathcal{O}} = \Lambda$ which implies that the two quantities are indeed comparable (they both measure an expected number of *time steps*). If $\mathcal{S}_{\mathcal{O}} = \mathcal{S}$, $\Lambda_{\mathcal{O}} \geq \Lambda$ and so the only case when $\Lambda_{\mathcal{O}} < \Lambda$ is when $\mathcal{S}_{\mathcal{O}} \subsetneq \mathcal{S}$. In this case, $S_{\mathcal{O}} \leq S$ and so the regret is all the more reduced. The ratio (6.38) also shows that the number of options should not be excessively high compared to the number of actions to preserve some advantage in using options.

Besides the fact that options can potentially reduce the travel-budget Λ, the support Γ of tansition probabilities, the number of states S or the number of actions A, the *main contribution* of this analysis is to exhibit the ratio $\frac{n}{T_n}$ of number of *decision steps* over number of *time steps*. This ratio formalizes the concept of *temporal abstraction* in RL. When using options, the transition dynamics of the environment need only be estimated at the level of macro-actions (i.e., options) causing the regret to grow with the number of *decision steps* rather than time steps like with *primitive actions*.[8] The *longer* the options, the *lower* the regret although this is mitigated by the presence of the additional term $r_{\max} t_{\max}$ ($r_{\max}\sqrt{t_{\max}}$ in the lower bound) which quantifies the difficulty of estimating the parameters of a macro-action. Since $\liminf_{n \to +\infty} \frac{T_n}{n} \geq \min_{s,a} \tau(s,a)$, then (6.38) gives an (asymptotic) sufficient condition for reducing the regret when using options, that is

$$\left(\sqrt{\frac{\Lambda_{\mathcal{O}} \Gamma_{\mathcal{O}}}{\Lambda \Gamma}} + \frac{r_{\max} t_{\max}}{\sqrt{r_{\max} \Lambda \Gamma}} \right) \sqrt{\frac{S_{\mathcal{O}} O}{SA \min_{s,a} \tau(s,a)}} \leq 1. \tag{6.39}$$

Perhaps not surprisingly, options are not always beneficial and can even *worsen* the learning performance if not carefully chosen. This is a form of *no-free lunch* which reminds the supervised learning setting (we recall that defining a set of options amounts to constrain the *policy space*, which can be seen as the equivalent of the *function class* in supervised learning). Accordingly, only adding options to the set of pimitive actions is often a bad strategy (the policy space is the same in that case). This is confirmed by our analysis since in that case $O \geq A$, $\Gamma_{\mathcal{O}} \geq \Gamma$ $S_{\mathcal{O}} = S$ and $\Lambda_{\mathcal{O}} = \Lambda$.

6.4 Learning in MDPs with Options without prior knowledge

At each episode, SUCRL solves an *"optimistic"* version of the optimality equation of M_{eq} (obtained by uniformization of SMDP $M_{\mathcal{O}}$, see Prop. 6.2) i.e., an optimistic version of equation $h^*_{\text{eq}} + g^*_{\text{eq}} e = L_{\text{eq}} h^*_{\text{eq}}$. Gain optimism is achieved by constructing confidence intervals on $R(s,o)$ and $\tau(s,o)$ using parameters $(\sigma_r(s,o), d_r(s,o))$ and $(\sigma_\tau(s,o), d_\tau(s,o))$ (Eq. 6.18 and 6.19). Without any *prior knowledge* on the distribution of options, such confidence intervals cannot be directly constructed and SUCRL cannot be run. Similarly, confidence intervals need to be

[8] The main term of the regret comes from the uncertainty in the environment dynamics

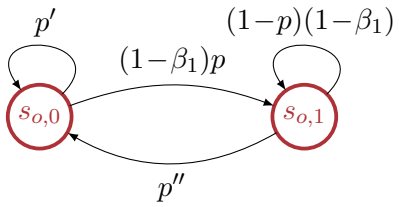


Figure 6.3: Irreducible MC obtained by transforming the absorbing MC of Fig. 6.2 with $p' = (1-\beta_0)(1-p) + \beta_0(1-p) + p\beta_1$ and $p'' = \beta_1(1-p) + p$.

computed for $b(\cdot|s,o)$, but this does not require any prior knowledge on the SMDP since the transition probabilities naturally belong to the simplex over states.

In practice, having access to tight sub-exponential parameters is often a strong requirement and any *incorrect* parametrization (e.g., loose upper-bounds on the true parameters) directly translates into a *poorer* regret performance. Furthermore, even if a hand-designed set of options may come with accurate estimates of their parameters, this would not be possible for automatically generated options, which are of increasing interest to the RL community. Finally, SUCRL views each option as a distinct and atomic *macro-action* (with sub-exponential distribution), thus losing the potential benefit of considering the *inner structure* and the *interaction* between options (*correlated* discrete phase-type distributions with *shared* states and primitive actions, see Sec. 6.2.3), which could be used to significantly improve sample efficiency.

In this section, we combine the semi-Markov decision process view on options and the *intrinsic MDP structure* underlying their execution (see Sec. 6.2.3) to achieve temporal abstraction without relying on sub-exponential parameters that are typically *unknown*. The optimality equation of M_{eq} can be rewritten as:

$$\forall s \in \mathcal{S}_{\mathcal{O}}, \quad g^*_{\text{eq}} = \max_{o \in \mathcal{O}_s} \left\{ \frac{R(s,o)}{\tau(s,o)} + \frac{\alpha}{\tau(s,o)} \left(b(\cdot|s,o)^\intercal h^*_{\text{eq}} - h^*_{\text{eq}}(s) \right) \right\}. \tag{6.40}$$

The term on the right-hand side of Eq. 6.40 is therefore *"homogeneous"* to a gain. We will introduce a *transformation* mapping each state-option pair (absorbing Markov Chain) to an associated *irreducible Markov chain*, where the gain of this Markov chain is the right-hand side term of Eq. 6.40. We will show that optimistic policies can be computed using only the irreducible chains and the SMDP dynamics (i.e., state to state transition probabilities through options). This approach does not need to *explicitly* estimate cumulative rewards and duration of options and their confidence intervals.

6.4.1 From absorbing to irreducible Markov Chains

From Eq. 6.40, we notice that computing the optimal policy only requires computing the ratio $R(s,o)/\tau(s,o) \in [0, r_{\max}]$ and the inverse $1/\tau(s,o) \in [0,1]$. Starting from the absorbing Markov Chain $P_{s,o}$ (Sec. 6.2.3), we can construct an irreducible MC whose stationary distribution is directly related to these terms. We proceed as illustrated in Fig. 6.3: all terminal states are *"merged"* together and their transitions are *"redirected"* to the initial state $s \in \mathcal{S}_{s,o}$. More formally, $v_{s,o} := V_{s,o} e \in \mathbb{R}^{S_{s,o}}$ contains the cumulative probability to transition from an

inner state to *any* terminal state. Then we define $Q'_{s,o} \in \mathbb{R}^{\mathcal{S}_o \times \mathcal{S}_o}$ as equal to $Q_{s,o}$ with $v_{s,o}$ *added to the s-th column of* $Q_{s,o}$. $Q'_{s,o}e = Q_{s,o}e + V_{s,o}e = e$ and $Q'_{s,o} \geq 0$ implying that $Q'_{s,o}$ is a stochastic matrix and the associated MC is necessarily *irreducible* since all states in $\mathcal{S}_{s,o}$ are reachable form s by construction (definition of $\mathcal{S}_{s,o}$), and s is reachable from any state in $\mathcal{S}_{s,o}$ due to the addition of $v_{s,o}$. Therefore, $Q'_{s,o}$ admits a unique *stationary distribution* $\mu_{s,o}$ i.e., a unique solution to the system of equations $\mu_{s,o}^\intercal Q'_{s,o} = \mu_{s,o}^\intercal$ and $\mu_{s,o}^\intercal e = 1$ (Bremaud, 1999, Chapter 3). In order to relate $\mu_{s,o}$ to the optimality equation (6.40), we need an additional assumption on options.

Assumption 6.3

For any state-option pair $(s, o) \in \mathcal{S}_\mathcal{O} \times \mathcal{O}$, the starting state s is also a terminal state i.e., $\beta_o(s) = 1$.

We now analyze the implications of Asm. 6.3. Let $\mathcal{O}$ be a set of options, possibly not satisfying Asm. 6.3, and $\mathcal{O}'$ a slightly different set of options obtained by forcing $\beta_o(s) = 1$ for all state-options pairs $(s, o) \in \mathcal{S}_\mathcal{O} \times \mathcal{O}$. It is straightforward to prove the following equivalence.

Proposition 6.7

Let π be a stationary deterministic *policy over options $\mathcal{O}$. There exists a* stationary deterministic *policy π' over options $\mathcal{O}'$ such that the induced process over states, actions and rewards (in the original MDP M) is the same for both π and π', i.e., for any sequence $\mathcal{H}_t = (s_1, a_1, r_1 \ldots, s_t)$, $\mathbb{P}^\pi(\mathcal{H}_t) = \mathbb{P}^{\pi'}(\mathcal{H}_t)$.*

Proof. For any option $o \in \mathcal{O}$ in the orginal set of options, let's denote by $o' \in \mathcal{O}'$ the same option after forcing $\beta_o(s) = 1$. For any stationary policy π over $\mathcal{O}$, let's define a corresponding stationary policy π' over $\mathcal{O}'$ by: $\pi'(s) = (\pi(s))'$, $\forall s \in \mathcal{S}_\mathcal{O}$. For any option o such that $\pi(s) = o$ and $\beta_o(s) < 1$, the state $s \in \mathcal{S}_{s,o}$ might be visited while o is being executed and o is not stopped in s. But since π_o (policy of option o) is *stationary Markov*, the distribution on the sequence of states and actions visited after s is exactly the same as if the option was first stopped and executed again (in both cases the policy π_o and the starting state s are the same). So the process over states and actions is the same for π and π'. ■

Since the opimal policy over options $\mathcal{O}$ is stationary deterministic (optimal policy of SMDP $M_\mathcal{O}$), Prop. 6.7 implies that Asm. 6.3 is not very restrictive. We are now ready to prove an important lemma.

Lemma 6.6

Under Asm. 6.3, let $\mu_{s,o} \in [0,1]^{\mathcal{S}_o}$ be the unique stationary distribution of the irreducible MC $Q'_{s,o}$ associated to state-option (s,o), then

$$\frac{1}{\tau(s,o)} = \mu_{s,o}(s) \qquad \text{and} \qquad \frac{R(s,o)}{\tau(s,o)} = \sum_{\substack{x \in \mathcal{S}_{s,o} \\ a \in \mathcal{A}_x}} r(x,a)\pi_o(a|x)\mu_{s,o}(x). \tag{6.41}$$

Proof. Under Asm. 6.3, $Q_{s,o}(x,s) = (1-\beta_o(s)) \cdot \sum_a p(s|x,a)\pi_o(a|x) = 0$ for all $x \in \mathcal{S}_{s,o}$ implying that $Q'_{s,o}(x,s) = v_{s,o}(x)$. So state s can only be reached when the option is "reset". $Q'_{s,o}$ has a finite number of states and is thus recurrent positive (see e.g., Thm. 3.3 of Bremaud (1999, Chapter 3)). Moreover, $1/\mu_{s,o}(s)$ corresponds to the *mean return* time in state s, i.e., the expected time to reach s starting from s (see e.g., Theorem 3.2 in Bremaud (1999, Chapter 3)). Finally, $\tau(s,o)$ is the expected time before reaching an absorbing states starting from s in the original absorbing Markov chain $P_{s,o}$. Since all absorbing states of $Q_{s,o}$ are merged with s in MC $Q'_{s,o}$, $1/\mu_{s,o}(s)$ is exactly equal to $\tau(s,o)$ in this case.

Let $(s_t)_{t\in\mathbb{N}}$ be the sequence of states visited while executing $Q'_{s,o}$ starting from s and let $r_t = \sum_{a\in\mathcal{A}_{s_t}} r(s_t,a)\pi_o(a|s_t)$. By the *Ergodic Theorem for Markov chains* (see e.g., Thm. 4.1 of Bremaud (1999, Chapter 3)):

$$\lim_{T\to+\infty} \frac{\sum_{t=0}^{T-1} r_t}{T} = \sum_{\substack{x \in \mathcal{S}_{s,o} \\ a \in \mathcal{A}_x}} r(x,a)\pi_o(a|x)\mu_{s,o}(x) \quad \text{a.s.} \tag{6.42}$$

Let $T_0 = 0, T_1, T_2, ...$ be the successive times of visit to s (random stopping times) i.e., $T_0 := 0$ and $T_{n+1} := \inf\{t > T_n : \ s_t = s\}$. From the *Regenerative Cycle Theorem for Markov chains* (see e.g., Thm. 7.4 of Bremaud (1999, Chapter 2)) we have that the pieces of trajectory $\left(s_{T_n}, ..., s_{T_{n+1}-1}\right)_{n\geq 0}$ are i.i.d. By the *Law of Large Numbers* we thus have:

$$\frac{\sum_{t=0}^{T_n-1} r_t}{n} = \frac{\sum_{k=0}^{n-1}\left(\sum_{t=T_k}^{T_{k+1}-1} r_t\right)}{n} \underset{n\to+\infty}{\longrightarrow} R(s,o) \quad \text{a.s.}$$

The same arguments can be used to show that

$$\frac{T_n}{n} = \frac{\sum_{k=0}^{n-1}(T_{k+1}-T_k)}{n} \underset{n\to+\infty}{\longrightarrow} \tau(s,o) \quad \text{a.s.}$$

By taking the ratio, the term n disappears and we obtain:

$$\frac{\sum_{t=0}^{T_n-1} r_t}{T_n} \underset{n\to+\infty}{\longrightarrow} \frac{R(s,o)}{\tau(s,o)} \quad \text{a.s.} \tag{6.43}$$

All sub-sequences of a convergent sequence converge to the limit of that sequence. Extracting

the subsequence $(T_n)_{n\in\mathbb{N}}$ in (6.42) we obtain:

$$\frac{\sum_{t=0}^{T_n-1} r_t}{T_n} \underset{n\to+\infty}{\longrightarrow} \sum_{\substack{x\in\mathcal{S}_{s,o}\\ a\in\mathcal{A}_x}} r(x,a)\pi_o(a|x)\mu_{s,o}(x) \quad \text{a.s.} \tag{6.44}$$

We then use the *uniqueness* of the limit ((6.43) and (6.44)) to conclude the proof of (6.41). ■

Lem. 6.6 makes explicit the relationship between the stationary distribution of $Q'_{s,o}$ and the key terms appearing in Eq. 6.40. More precisely, we have shown that:

$$\frac{R(s,o)}{\tau(s,o)} + \frac{\alpha}{\tau(s,o)}\left(b(\cdot|s,o)^\intercal h^*_{\text{eq}} - h^*_{\text{eq}}(s)\right) = \sum_{\substack{x\in\mathcal{S}_{s,o}\\ a\in\mathcal{A}_x}} r(x,a)\pi_o(a|x)\mu_{s,o}(x) + \alpha\left(b(\cdot|s,o)^\intercal h^*_{\text{eq}} - h^*_{\text{eq}}(s)\right)\mu_{s,o}(s).$$

This confirms our first intuition that the term $\frac{R(s,o)}{\tau(s,o)} + \frac{\alpha}{\tau(s,o)}\left(b(\cdot|s,o)^\intercal h^*_{\text{eq}} - h^*_{\text{eq}}(s)\right)$ corresponds to a long term average *gain*, namely the gain of the *Markov Reward Process* (MRP) characterized by the MC $Q'_{s,o}$ and the reward function defined by

$$\begin{cases} \sum_{a\in\mathcal{A}_x} r(x,a)\pi_o(a|x) \text{ for } x \neq s, \\ \sum_{a\in\mathcal{A}_s} r(s,a)\pi_o(a|s) + \alpha\left(b(\cdot|s,o)^\intercal h^*_{\text{eq}} - h^*_{\text{eq}}(s)\right) \text{ for } x = s. \end{cases}$$

6.4.2 Optimistic bilevel Bellman operator

Inspired by the mapping between options and irreducible MRPs highlighted in the previous section, we will now define an optimistic Bellman operator $\mathcal{L}^{\text{eq}}_k$ that uses confidence intervals on $b(\cdot|s,o)$, as well as confidence intervals on $Q'_{s,o}$ and $r(x,a)$ (rather than $\tau(s,o)$ and $R(s,o)$). For the rewards we use the same confidence intervals as in UCRLB i.e., $r_k(s,a) \in B^k_r(s,a)$, while for the transition probabilities at the level of options we use the same confidence bounds as in SUCRL (and in UCRLB) i.e., $b_k(\cdot|s,o) \in B^k_p(s,o)$. We also use $B^k_p(s,a)$ for $Q'_{s,o}$.

Inner Bellman operators. We start with the formal definition of a *"local"* extended Bellman operator $\mathcal{L}^{s,o}_k$ characterizing the *inner* dynamics and reward of state-option pair $(s,o) \in \mathcal{S}_\mathcal{O} \times \mathcal{O}$. $\mathcal{L}^{s,o}_k$ takes two inputs: a scalar $c \in \mathbb{R}$ and a vector $u \in \mathbb{R}^{S_{s,o}}$. For all $x \in \mathcal{S}_{s,o}$,

$$\mathcal{L}^{s,o}_k(c,u)(x) := \sum_{a\in\mathcal{A}_x} \pi_o(a|x)\left(\max_{r\in B^k_r(x,a)}\{r\} + \max_{p\in B^k_p(x,a)}\left\{p^\intercal\left((e-\beta_o)\circ u + u(s)\beta_o\right)\right\}\right) + c\mathbb{1}\{x=s\}. \tag{6.45}$$

The vector β_o appearing in Eq. 6.45 corresponds to the stopping condition of option o restricted to the susbset of states $\mathcal{S}_{s,o}$. $\circ$ denotes the Hadamard product i.e., $(e-\beta_o)\circ u = \left((1-\beta_o(x))u(x)\right)_{x\in\mathcal{S}_{s,o}}$. The scalar c appears only in state s.

Although this may not be obvious at first sight, for any fixed $c \in \mathbb{R}$, $\mathcal{L}_k^{s,o}(c,\cdot)$ is an (extended) optimal Bellman operator. The scalar c can indeed be interpreted as an additional reward in state s, while the scalar product $p^\intercal\big((e-\beta_o)\circ u + u(s)\beta_o\big)$ can be expressed as $q^\intercal u$ where $q := p\circ(e-\beta_o)+p^\intercal\beta_o e_s$ is a probability vector (i.e., $q \geq 0$ and $q^\intercal e = 1$). p is also a probability vector and can be easily computed using LPROBA with input vector $(e-\beta_o)\circ u+u(s)\beta_o$ (sorted in decreasing order). Since $B_p^k(x,a)$ is a *polytope*, p and q take values in a finite set that is independent of u, implying that $\mathcal{L}_k^{s,o}(c,\cdot)$ can be expressed as an optimal Bellman operator with *finitely many actions*. Furthermore, the associated MDP $\mathcal{M}_k^{s,o}(c)$ is *communicating*. Due to Prop. 2.4, there *exists* a solution $(g_k^{s,o}(c), h_k^{s,o}(c)) \in \mathbb{R}\times\mathbb{R}^{S_{s,o}}$ to the fixed point equation $h_k^{s,o}(c) + g_k^{s,o}(c)e = \mathcal{L}_k^{s,o}\left(c, h_k^{s,o}(c)\right)$, where $g_k^{s,o}(c)$ is *unique* and can be expressed as:

$$g_k^{s,o}(c) = \mu_k^{s,o}(s)c + \sum_{\substack{x\in\mathcal{S}_{s,o}\\ a\in\mathcal{A}_x}} \pi_o(a|x)\mu_k^{s,o}(x)\cdot \max_{r\in B_r^k(x,a)}\{r\}, \tag{6.46}$$

with $\mu_k^{s,o}$ the stationary distribution of any optimal policy (e.g., a greedy policy w.r.t. $h_k^{s,o}(c)$).

Even though the true Markov Chain $Q'_{s,o}$ is irreducible by construction (and so $\mu_{s,o}$ is unique with $\mu_{s,o}(s) > 0$), it is not necessarily the case for the *optimistic chain*. This chain can happen to contain transient states and/or several recurrent classes. $\mu_k^{s,o}$ is not uniquely defined[9] (but exists) and $\mu_k^{s,o}(s)$ can happen to be 0.

Outer Bellman operator. We define the *"global"* operator $\mathcal{L}_k^{\text{eq}}$ relating all options as follows:

$$\forall v\in\mathbb{R}^{S_\mathcal{O}},\ \forall s\in\mathcal{S}_\mathcal{O},\quad \mathcal{L}_k^{\text{eq}}v(s) := \max_{o\in\mathcal{O}_s}\left\{g_k^{s,o}\left(\alpha\cdot\max_{b\in B_p^k(s,o)}\{b^\intercal v\} - \alpha\cdot v(s)\right)\right\} + v(s). \tag{6.47}$$

$\mathcal{L}_k^{\text{eq}}$ accounts for the *outer* rewards and dynamics at the "SMDP level". Using (6.46), we can rewrite (6.47) as

$$\max_{o\in\mathcal{O}_s}\left\{\max_{\mu}\left\{\underbrace{\sum_{\substack{x\in\mathcal{S}_{s,o}\\ a\in\mathcal{A}_x}}\pi_o(a|x)\mu(x)\cdot\max_{r\in B_r^k(x,a)}\{r\}}_{\text{reward in } [0,r_{\max}]} + \underbrace{\alpha\mu(s)\cdot\max_{b\in B_p^k(s,o)}\{b^\intercal v\} + (1-\alpha\mu(s))\cdot v(s)}_{=p^\intercal v,\text{ with probability vector } p=\alpha\mu(s)b+(1-\alpha\mu(s))e_s}\right\}\right\},$$

where μ is constrained to be a stationary distibution of a (not necessarily irreducible) MC contained in the confidence intervals of $Q'_{s,o}$ (see above). As we explained earlier, q can be constrained to lie in a finite space without impacting the final result, and therefore so does μ. In conclusion, $\mathcal{L}_k^{\text{eq}}$ is an (extended) optimal Bellman operator which can be expressed with only *finitely many actions*. The associated extended MDP $\mathcal{M}_k^{\text{eq}}$ is communicating and so due to Prop. 2.4, there exists a solution $(g_k^{\text{eq}}, h_k^{\text{eq}}) \in [0, r_{\max}]\times\mathbb{R}^{S_\mathcal{O}}$ to the optimality equation $h_k^{\text{eq}} + g_k^{\text{eq}}e = \mathcal{L}_k^{\text{eq}}h_k^{\text{eq}}$. Unlike in the SMDP formulation of SUCRL where the holding times and cumulative rewards must lie in *bounded* confidence intervals, in this new formulation $\mu(s)$ can be equal to 0 (corresponding to an *infinite* holding time and cumulative reward) without

[9] Although $\mu_k^{s,o}$ is not unique, the value in Eq. 6.46 is the same for all possible values of $\mu_k^{s,o}$.

compromising the solution of the optimality equation. Furthermore, this new approach implicitly leverages over the *correlations* between cumulative reward and holding time, which is ignored when estimating $R(s,o)$ and $\tau(s,o)$ separately.

Since $\mathcal{M}_k^{\text{eq}}$ is aperiodic by construction (see Eq. 6.4), Prop. 2.6 implies that EVI converges to a solution of the optimality equation. The limit of the sequence of vectors v_n generated by EVI when started from vector $v_0 = 0$ will be denoted h_k^{eq}. Due to Thm. 3.3, $sp\left(h_k^{\text{eq}}\right) \leq \Lambda_k^{\text{eq}}$. To simplify notations, whenever $c = \alpha \cdot \left(\max_{b \in B_p^k(s,o)} \left\{b^\intercal h_k^{\text{eq}}\right\} - h_k^{\text{eq}}(s)\right)$ we drop the dependency in c in Eq. 6.45 i.e., we simply denote the inner operator by $\mathcal{L}_k^{s,o}$. Let $(g_k^{s,o}, h_k^{s,o})$ be any solution to $h_k^{s,o} + g_k^{s,o} e = \mathcal{L}_k^{s,o} h_k^{s,o}$. We cannot use Thm. 3.3 to bound $sp\left(h_k^{s,o}\right)$ because we have no guarantee that the reward $\sum_{a \in \mathcal{A}_s} \pi_o(a|s) \left(\max_{r \in B_r^k(s,a)}\{r\} + \alpha \cdot \left(\max_{b \in B_p^k(s,o)} \left\{b^\intercal h_k^{\text{eq}}\right\} - \cdot h_k^{\text{eq}}(s)\right)\right)$ is bounded by $r_{\max}$. However, combining the inner and outer optimality equations, we obtain that $g_k^{\text{eq}} = \max_{o \in \mathcal{O}_s} \left\{g_k^{s,o}\right\}$ for all $s \in \mathcal{S}_{\mathcal{O}}$ so that $g_k^{s,o} \leq g_k^{\text{eq}} \leq r_{\max}$. Thm. 3.2 then shows that $sp\left(h_k^{s,o}\right) \leq \Lambda_k^{s,o}$ where $\Lambda_k^{s,o}$ is the *"travel-budget"* of the extended MDP $\mathcal{M}_k^{s,o}$ but with policies *restricted* in $\Pi_{\mapsto y}^{\text{SD}}(\mathcal{M}_k^{s,o})$:

$$\Lambda_k^{s,o} := \max_{x,y \in \mathcal{S}_{s,o}} \min_{\pi \in \Pi_{\mapsto y}^{\text{SD}}(\mathcal{M}_k^{s,o})} \mathbb{E}_{\mathcal{M}_k^{s,o}}^{\pi} \left[\sum_{t=1}^{\tau(y)-1} r_{\max} - r(s_t, a_t) \middle| s_1 = x \right]. \tag{6.48}$$

Gain optimism. We use the same argument as for SUCRL: with high probability, $\mathcal{L}_k^{\text{eq}} h_{\text{eq}}^* \geq L_{\text{eq}} h_{\text{eq}}^*$ implying $g_k^{\text{eq}} \geq g_{eq}^* = g_{M_{\mathcal{O}}}^*$.

Range of optimistic biases. The travel-budget $\Lambda_{s,o}$ of any state-option pair $(s,o) \in \mathcal{S}_{\mathcal{O}} \times \mathcal{O}$ is defined as

$$\Lambda_{s,o} := \max_{x,y \in \mathcal{S}_{s,o}} \mathbb{E}_{Q'_{s,o}} \left[\sum_{t=1}^{\tau(y)-1} r_{\max} - \sum_{a \in \mathcal{A}_{s_t}} r(s_t, a) \pi_o(a|s_t) \middle| s_1 = x \right]. \tag{6.49}$$

$\mathbb{E}_{Q'_{s,o}}$ denotes the expectation in the irreducible Markov Chain $Q'_{s,o}$. Since by construction all states are *positive recurrent*, $\mathbb{P}_{Q'_{s,o}}\left(\tau(y) < +\infty \middle| s_1 = x\right) = 1$ so that $\Lambda_{s,o} < +\infty$. Under the same high probability event for which $g_k^{\text{eq}} \geq g_{eq}^*$, $\Lambda_k^{\text{eq}} \leq \Lambda/\alpha$ (same arguments as in SUCRL). A similar reasoning can be used to show that $\Lambda_k^{s,o} \leq \Lambda_{s,o}$.

6.4.3 FSUCRL: SUCRL with Irreducible Markov Chains

Algorithm

FSUCRL combines the confidence bounds $B_p^k(s,o)$ (of the state-option transition $b(\cdot|s,o)$) used in SUCRL, with the confidence bounds $B_r^k(s,a)$ and $B_p^k(s,a)$ (of the state-action reward $r(s,a)$ and transition $p(\cdot|s,a)$) used in UCRLB. FSUCRL does not build confidence intervals on $\tau(s,o)$ and $R(s,o)$ and so *no prior knowledge* on the distribution of holding times and

cumulative rewards of options is needed[10] (e.g., sub-exponential parameters σ_τ, b_τ, σ_r and b_r). The confidence sets define the extended MDP $\mathcal{M}_k^{\text{eq}}$ described in Sec. 6.4.2.

We will assume that FSUCRL computes h_k^{eq} and $h_k^{s,o}$ *exactly* instead of *approximately* using EVI (see Eq. 6.22 in Alg. 12). The policy π_k played at episode k is therefore a *greedy policy* with respect to h_k^{eq} i.e., $\pi_k = G_k^{\text{eq}} h_k^{\text{eq}}$.[11] Computing h_k^{eq} and $h_k^{s,o}$ exactly allows to bound the range (span) of $h_k^{s,o}$ by $\Lambda_{s,o}$ (see Sec. 6.4.2). It is unclear whether we can approximate h_k^{eq} and $h_k^{s,o}$ using an efficient iterative procedure (similar to EVI) while preserving the property $sp\left(h_k^{s,o}\right) \leq \Lambda_{s,o}$. The main challenge is that we have intricated equations e.g., the term c used in the definition of $\mathcal{L}_k^{s,o}(c, v)$ changes at every iteration. Nevertheless, we will later provide a *convergent algorithm* to approximate h_k^{eq} and $h_k^{s,o}$.

Finally, the *stopping condition* used to end an episode *combines* the stopping conditions used by both UCRLB and SUCRL i.e., an episodes stops whenever *either* $\nu_k(s_t, a_t) \geq N_k^+(s_t, a_t)$ for the last *state-action pair* (s_t, a_t) played *or* $\nu_k(s_i, o_i) \geq N_k^+(s_i, o_i)$ for the last *state-option pair* (s_i, o_i) played. Since when the first condition $\nu_k(s_t, a_t) \geq N_k^+(s_t, a_t)$ is triggered the current option being played o_{N_t} may not be over, FSUCRL *waits for the option to end*.

Regret guarantees

We present two regret bounds for FSUCRL (like for SUCRL). Like in Thm. 6.1 and 6.2, the bounds are composed of *two* distinct terms: one reflects the difficulty to learn the dynamics of the corresponding SMDP $M_{\mathcal{O}}$, while the other characterizes the uncertainty of the options themselves. To simplify the bound, we introduce $\Lambda_{\max} := \max_{s,o} \Lambda_{s,o}$.

Theorem 6.7 (Analogue of Thm. 3.4)

There exists a numerical constant $\beta > 0$ such that for any *communicating MDP M, with probability at least $1-\delta$, it holds that for all initial state distributions $\mu_1 \in \Delta_S$ and for all time horizons $T > 1$:*

$$
\begin{aligned}
\Delta(M, \text{FSUCRL}, \mu_1, T_n) \leq\ & \beta \cdot \max\{r_{\max}, \Lambda_{\mathcal{O}}\} \sqrt{\left(\sum_{s,o} \Gamma(s,o)\right) n \ln\left(\frac{n}{\delta}\right)} \\
& + \beta \cdot \max\{r_{\max}, \Lambda_{\max}\} \sqrt{\left(\sum_{s,a} \Gamma(s,a)\right) T_n \ln\left(\frac{T_n}{\delta}\right)} \\
& + \beta \cdot \max\{r_{\max}, \Lambda_{\mathcal{O}}\} S_{\mathcal{O}}^2 O \ln\left(\frac{n}{\delta}\right) \ln(n) \\
& + \beta \cdot SA \ln\left(\frac{T_n}{\delta}\right) \ln(T_n) \Big(\max\{r_{\max}, \Lambda_{\max}\} S + r_{\max}(\tau_{\max} + \sigma_\tau + d_\tau) \Big).
\end{aligned}
$$

[10] FSUCRL is somehow a *"parameter-Free"* version of SUCRL (hence the acronym).
[11] Unlike in SUCRL and UCRLB, π_k is chosen *deterministic* so that Prop. 6.7 applies.

Theorem 6.8 (Analogue of Thm. 3.5)

There exists a numerical constant $\beta > 0$ such that for any *communicating MDP M, with probability at least $1-\delta$, it holds that for all initial state distributions $\mu_1 \in \Delta_S$ and for all time horizons $T > 1$:*

$$\begin{aligned}\Delta(M, \text{FSUCRL}, \mu_1, T_n) \leq & \beta \cdot \max\left\{r_{\max}, \sqrt{r_{\max}\Lambda_{\mathcal{O}}}\right\}\sqrt{\left(\sum_{s,o}\Gamma(s,o)\right) n \ln\left(\frac{n}{\delta}\right)} \\ & + \beta \cdot \max\left\{r_{\max}, \sqrt{r_{\max}\Lambda_{\max}}\right\}\sqrt{\left(\sum_{s,a}\Gamma(s,a)\right) T_n \ln\left(\frac{T_n}{\delta}\right)} \\ & + \beta \cdot \max\left\{r_{\max}, \frac{\Lambda_{\mathcal{O}}^2}{r_{\max}}\right\} S_{\mathcal{O}}^2 O \ln\left(\frac{n}{\delta}\right)\ln(n) \\ & + \beta \cdot SA \ln\left(\frac{T_n}{\delta}\right)\ln(T_n)\left(\max\left\{r_{\max}, \frac{\Lambda_{\max}^2}{r_{\max}}\right\} S + r_{\max}(\tau_{\max} + \sigma_\tau + d_\tau)\right).\end{aligned}$$

The bounds presented above illustrate how options implicitly implement the *divide-and-conquer* paradigm. The main regret term of UCRLB $\widetilde{\mathcal{O}}\left(\sqrt{r_{\max}\Lambda\Gamma SAT_n}\right)$ sees the travel-budget reduced to $\Lambda_{\max}$ while another term $\widetilde{\mathcal{O}}\left(\sqrt{r_{\max}\Lambda_{\mathcal{O}}\Gamma_{\mathcal{O}}S_{\mathcal{O}}On}\right)$ appears, which only scales with the number of decision steps n instad of the number of time steps T_n. The ratio introduced in Sec. 6.3.6 is now roughly bounded as

$$\mathcal{R}(n) \lesssim \sqrt{\frac{\Lambda_{\mathcal{O}}\Gamma_{\mathcal{O}}S_{\mathcal{O}}On}{\Lambda\Gamma SAT_n}} + \sqrt{\frac{\Lambda_{\max}}{\Lambda}}. \tag{6.50}$$

The conclusions that we can draw from (6.50) are similar to the one of Sec. 6.3.6 except that we removed the dependency in the potentially loose sub-exponential parameters of options. We replace these terms by *intrinsic* and *a priori unknown* properties of options (namely their travel-budget) which provide more insights. It is clear from the bounds that unlike SUCRL, FSUCRL leverages the *inner correlation* between the cumulative reward and duration of a single option, as well as the *outer correlation* between different options that share inner state-action pairs. The worst-case travel-budget of options $\Lambda_{\max}$ is a very *loose* upper-bound in practice but difficult to improve while preserving the readibility and interpretability of the regret bound.

Regret analysis

Stopping condition of episodes. FSUCRL uses two condition to terminate an episodes and so the total number of episodes k_n can be decomposed as $k_n = k_n^1 + k_n^2$, where k_n^1 is the number of episodes for which the first condition $\nu_k(s_t, a_t) \geq N_k^+(s_t, a_t)$ is triggered, while k_n^2 is the number of episodes for which the second condition $\nu_k(s_i, o_i) \geq N_k^+(s_i, o_i)$ is triggered. k_n^2 can be bounded as in SUCRL i.e., $k_n^2 \leq S_{\mathcal{O}}O\log_2\left(\frac{8n}{S_{\mathcal{O}}O}\right)$ (see Prop. 3.8). Moreover, this

stopping condition ensures that at every episode k, $\nu_k(s,o) \leq N_k^+(s,o)$ for all pairs (s,o).

Let's now analyze the first stopping condition. Once the number of visits has doubled in one state-action pair, FSUCRL needs to wait for the option being executed to end before starting the next episode. This can only decrease the number of episodes compared to UCRLB. Indeed, at the end of every episode k, the condition $\nu_k(x,a) \geq N_k^+(x,a)$ is always satisfied *for at least one* state-action pair (x,a) and so Prop. 3.8 (which only relies on this property) also holds i.e., $k_n^1 \leq SA \log_2\left(\frac{8T_n}{SA}\right)$. Although the bound on the number of episodes is unchanged, the condition $\nu_k(x,a) \leq N_k^+(x,a)$ *for all* state-action pairs (x,a) no longer holds and we cannot apply Lem. 3.6 to bound the series $\sum_{k=1}^{k_T} \frac{\nu_k(x,a)}{\sqrt{N_k^+(x,a)}}$ and $\sum_{k=1}^{k_T} \frac{\nu_k(x,a)}{N_k^+(x,a)}$. Nevertheless, this condition can only be violated while executing an option o_i that is the *last* of the episode. There is *at most one* such option in every episode and we will bound the regret in each time step of this option by $r_{\max}$. Using Cor. 6.2 with a union bound over all $i = 1 \dots n$ we have that with probability at least $1-\delta$

$$\forall i = 1 \dots n, \quad \tau_i \leq \tau_{\max} + 2\sigma_\tau \sqrt{\ln\left(\frac{2n}{\delta}\right)} + 4d_\tau \ln\left(\frac{2n}{\delta}\right).$$

This means that with high probability, the total regret incurred while executing the last option in all episodes where the first condition is triggered is (cumulatively) at most (ignoring multiplicative numerical constants)

$$r_{\max} \cdot k_n^1 \cdot \left(\tau_{\max} + \sigma_\tau \sqrt{\ln\left(\frac{n}{\delta}\right)} + d_\tau \ln\left(\frac{n}{\delta}\right)\right). \tag{6.51}$$

During the execution of all other options, we can use Lem. 3.6 to bound $\sum_{k=1}^{k_T} \frac{\nu_k(x,a)}{\sqrt{N_k^+(x,a)}}$ and $\sum_{k=1}^{k_T} \frac{\nu_k(x,a)}{N_k^+(x,a)}$. We will account for the term (6.51) in the final regret bound and in the rest of the proof, we will always assume that the condition $\nu_k(x,a) \leq N_k^+(x,a)$ is *never violated*.

Regret decomposition. The regret of FSUCRL can be decomposed as follows:

$$\Delta(\text{FSUCRL}, T_n) = \sum_{t=1}^{T_n} g_M^* - r_t = T_n \cdot (g_M^* - g_{M_\mathcal{O}}^*) + \sum_{t=1}^{T_n} g_{M_\mathcal{O}}^* - r_t.$$

To bound the sum $\sum_{t=1}^{T_n} g_{M_\mathcal{O}}^* - r_t$ we first follow the same steps as for UCRLB (Sec. 3.5.1 and 3.5.2). More precisely, we use a martingale argument (see Lem. 3.1) to bound $-\sum_{t=1}^{T_n} r_t$ and we use the optimism property to bound $g_{M_\mathcal{O}}^*$. We also introduce the optimistic rewards $r_k(s,a)$ and we use another martingale argument (see Lem. 3.5.4) to bound the cumulative differences $r_k(s,a) - r(s,a)$. We set $T_0 := 1$ and we recall that for all $n \geq 1$, $T_n := \sum_{i=1}^n \tau_i$ where τ_i is the duration of the i-th option played by the learning algorithm (denoted o_i). The state $s_{T_{i-1}}$ visited at *time step* T_{i-1} is the state in which o_i is started and is therefore abbreviated s_i (by analogy with SUCRL). The current episode at decision step i is denoted k_i (like in the analysis of SUCRL). The policy π_k played by FSUCRL at episode k is

deterministic and so $o_i = \pi_{k_i}(s_i)$. There is no particular reason to believe the analysis cannot be extended to randomized policies although it would be slightly more involved since we have to deal with several optimality equations as well as several policies wich can all be stochastic: π_k and $(\pi_o)_{o\in\mathcal{O}}$. In the end we obtain (with high probability and up to multiplicative numerical constants):

$$\sum_{t=1}^{T_n} g^*_{M_\mathcal{O}} - r_t \lesssim \sum_{i=1}^{n}\sum_{t=T_{i-1}}^{T_i-1}\left(g^{\mathrm{eq}}_{k_i} - \sum_{a\in\mathcal{A}_{s_t}} \pi_{o_i}(a|s_t) r_{k_i}(s_t,a)\right) + r_{\max}\sqrt{SAT_n \ln\left(\frac{SAT_n}{\delta}\right)} + r_{\max} SA \ln\left(\frac{SAT_n}{\delta}\right).$$

Unlike in Sec. 6.3.4, a denotes a *primitive action* in the original MDP M as opposed to a *macro-action* in SMDP $M_\mathcal{O}$ (denoted by o). Accordingly, $r_k(s,a)$ denotes the optimistic reward associated to state-action pair (s,a) and lies in $[0, r_{\max}]$. If option o_i is played in state s_i at decision step i then $g^{s_i,o_i}_{k_i} = g^{\mathrm{eq}}_{k_i}$ due to the *outer* optimality equation $g^{\mathrm{eq}}_k = \max_{o\in\mathcal{O}_s}\{g^{s,o}_k\}$ (see Sec. 6.4.2). We now use the *inner* optimality equations $h^{s,o}_k + g^{s,o}_k e = \mathcal{L}^{s,o}_k h^{s,o}_k$ which can be expanded as

$$g^{s_i,o_i}_{k_i} - \sum_{a\in\mathcal{A}_{s_t}} \pi_{o_i}(a|s_t) r_{k_i}(s_t,a) = \sum_{a\in\mathcal{A}_{s_t}} \pi_{o_i}(a|s_t) q^{s_i,o_i}_{k_i}(\cdot|s_t,a)^\intercal h^{s_i,o_i}_{k_i} - h^{s_i,o_i}_{k_i}(s_t) + \alpha\cdot\left(b_k(\cdot|s_i,o_i)^\intercal h^{\mathrm{eq}}_k - h^{\mathrm{eq}}_k(s_i)\right)\cdot \mathbb{1}\{t = T_{i-1}\}.$$

The additional term $\alpha\cdot\left(b_k(\cdot|s_i,o_i)^\intercal h^{\mathrm{eq}}_k - h^{\mathrm{eq}}_k(s_i)\right)$ only appears in the initial state s_i i.e., for $t = T_{i-1}$. For the sake of clarity, we use the simplifying notation $q^{s_i,o_i}_{k_i}(\cdot|s_t)$ to denote $\sum_a \pi_{o_i}(a|s_t) q^{s_i,o_i}_{k_i}(\cdot|s_t,a)$. The main regret term becomes:

$$\sum_{i=1}^{n}\sum_{t=T_{i-1}}^{T_i-1}\left(g^{\mathrm{eq}}_{k_i} - \sum_{a\in\mathcal{A}_{s_t}} \pi_{o_i}(a|s_t) r_{k_i}(s_t,a)\right) = \alpha\sum_{i=1}^{n}\left(b_{k_i}(\cdot|s_i,o_i)^\intercal h^{\mathrm{eq}}_{k_i} - h^{\mathrm{eq}}_{k_i}(s_i)\right) + \sum_{i=1}^{n}\sum_{t=T_{i-1}}^{T_i-1}\left(q^{s_i,o_i}_{k_i}(\cdot|s_t)^\intercal h^{s_i,o_i}_{k_i} - h^{s_i,o_i}_{k_i}(s_t)\right).$$

The first sum (on the left-hand side) is analogue to the main term appearing in Eq. 6.37 in the analysis of SUCRL (with different notations: b replaces p). It can be bounded in the same way (we refer to the analysis of UCRLB). This term quantifies the uncertainty on the dynamics *between options* at SMDP level. The main novelty in the analysis of FSUCRL is the second sum (on the right-hand side) which arises due to the uncertainty *within options*. This new term resembles the first sum: it corresponds to the difference between an "optimistic" expectation of $h^{s_i,o_i}_{k_i}(s_{t+1})$ given state s_t and $h^{s_i,o_i}_{k_i}(s_t)$. We will apply a very similar analysis.

Analysis of the new term. We start by adding and subtracting the true transition probability in the MC $Q'_{s,o}$ i.e., $q_{s_i,o_i}(\cdot|s_t) = \sum_a \pi_{o_i}(a|s_t) q_{s_i,o_i}(\cdot|s_t,a)$.

$$\sum_{i=1}^{n}\sum_{t=T_{i-1}}^{T_i-1}\left(q_{k_i}^{s_i,o_i}(\cdot|s_t)^\intercal h_{k_i}^{s_i,o_i} - h_{k_i}^{s_i,o_i}(s_t)\right) = \sum_{i=1}^{n}\sum_{t=T_{i-1}}^{T_i-1}\left(q_{k_i}^{s_i,o_i}(\cdot|s_t) - q_{s_i,o_i}(\cdot|s_t)\right)^\intercal h_{k_i}^{s_i,o_i} + \sum_{i=1}^{n}\sum_{t=T_{i-1}}^{T_i-1}\left(q_{s_i,o_i}(\cdot|s_t)^\intercal h_{k_i}^{s_i,o_i}(s_t) - h_{k_i}^{s_i,o_i}(s_t)\right) \tag{6.52}$$

The first term in Eq. 6.52 corresponds to the difference between the optimistic and estimated transition probability of irreducible MC $Q'_{s,o}$, amplified by the optimistic bias $h_k^{s,o}$:

$$\sum_{k=1}^{k_n}\sum_{s,o}\sum_{x\in\mathcal{S}_{s,o}}\nu_k(s,o,x)\sum_a \pi_o(a|x)\left(q_k^{s,o}(\cdot|x,a) - q_{s,o}(\cdot|x,a)\right)^\intercal h_k^{s,o}, \tag{6.53}$$

where $\nu_k(s,o,x)$ denote the total number of visits in state x while executing state option (s,o) during episode k. We then use the definition of $q_k^{s,o}$ and $q_{s,o}$ to reveal the optimistic and true transition probabilities p_k and p in the MDP (not the MC $Q'_{s,o}$):

$$\begin{aligned}\left(q_k^{s,o}(\cdot|x,a) - q_{s,o}(\cdot|x,a)\right)^\intercal h_k^{s,o} &= \sum_y \Big[\beta_o(y)\cdot\left(p_k(y|x,a) - p(y|x,a)\right)\cdot h_k^{s,o}(y) \\ &\qquad + \left(1-\beta_o(y)\right)\cdot\left(p_k(y|x,a) - p(y|x,a)\right)\cdot h_k^{s,o}(s)\Big] \\ &\leq \Lambda_{\max}\cdot\min\left\{2, \beta_{p,k}^{xa}\right\}\end{aligned}$$

The term (6.53) is therefore similar to the term Δ_k^{p1} appearing in the regret proof of UCRLB. We can apply Lem. 3.2 to obtain the bound

$$\Lambda_{\max}\sum_{k=1}^{k_n}\sum_{x,a}\nu_k(x,a)\beta_{p,k}^{xa} + 4\Lambda_{\max}\sqrt{T_n\ln\left(\frac{5T_n}{\delta}\right)}.$$

For a tighter bound, we can apply Lem. 3.7 and the decomposition of Sec. 3.6 instead. The final bound is obtained as in UCRLB.

The second term in Eq. 6.52 is the difference between $q_{s_i,o_i}(\cdot|s_t)^\intercal h_{k_i}^{s_i,o_i}$ and $h_{k_i}^{s_i,o_i}(s_t)$. We define the process $(x_t)_{t\in[T_{i-1},T_i-1]}$ by $x_t = s_t$ if $T_{i-1}\leq t < T_i$ and $x_{T_i} = s_i = x_{T_{i-1}}$. The process (x_t) follows the dynamics of option o_i until the stopping condition is triggered in which case x_t goes back to the initial state of the option s_i. In other words, (x_t) follows the distribution of Markov Chain $Q'_{s,o}$. We can then write

$$\begin{aligned}\sum_{t=T_{i-1}}^{T_i-1}\left(q_{s_i,o_i}(\cdot|s_t)^\intercal h_{k_i}^{s_i,o_i} - h_{k_i}^{s_i,o_i}(s_t)\right) &= \sum_{t=T_{i-1}}^{T_i-1}\left(q_{s_i,o_i}(\cdot|s_t)^\intercal h_{k_i}^{s_i,o_i} - h_{k_i}^{s_i,o_i}(x_{t+1})\right) \\ &\quad + \sum_{t=T_{i-1}}^{T_i-1}\left(h_{k_i}^{s_i,o_i}(x_{t+1}) - h_{k_i}^{s_i,o_i}(s_t)\right) \\ &= \sum_{t=T_{i-1}}^{T_i-1}\left(q_{s_i,o_i}(\cdot|x_t)^\intercal h_{k_i}^{s_i,o_i} - h_{k_i}^{s_i,o_i}(x_{t+1})\right).\end{aligned}$$

The *telescopic sum* appearing after adding $h_{k_i}^{s_i,o_i}(x_{t+1})$ is zero because $x_{T_i} = s_i = x_{T_{i-1}}$. Since

Algorithm 13 Nested (Relative) Value Iteration

Input: Operators $L_{s,o}(\cdot,\cdot) : \mathbb{R} \times \mathbb{R}^{S_{s,o}} \mapsto \mathbb{R}^{S_{s,o}}$ and $G_{s,o} : \mathbb{R}^S \mapsto D^{\mathrm{MR}}$, confidence intervals $(B_p(s,o,s'))_{s' \in \mathcal{S}_\mathcal{O}}$, accuracies $(\varepsilon_n)_{n \geq -1} \in]0, r_{\max}[^{\mathbb{N}}$, initial vector $v_0 \in \mathbb{R}^S$, arbitrary reference state $\overline{s} \in \mathcal{S}$

Output: Gain $g \in [0, r_{\max}]$, bias vectors $h \in \mathbb{R}^S$ and $h_{s,o} \in \mathbb{R}^{S_{s,o}}$ and stationary deterministic policy $\pi \in \Pi^{\mathrm{SD}}$

Initialize $n = -1$, $v_{-1} := -\infty$
while $sp\,(v_{n+1} - v_n) > \varepsilon_{-1} - \frac{3}{2}\varepsilon_{n+1}$ **do**
  Increment $n \leftarrow n + 1$
  Shift $v_n \leftarrow v_n - v_n(\overline{s})e$
  for $s \in \mathcal{S}$ **do**
    for $o \in \mathcal{O}$ **do**
      $c \leftarrow \alpha \cdot \Big(\textsc{LProba}\Big(v_n, (B_p(s,o,s'))_{s' \in \mathcal{S}_\mathcal{O}}\Big) - v_n(s)\Big)$
      $\Big(\widehat{g}_{s,o}, \widehat{h}_{s,o}, \widehat{\pi}_{s,o}\Big) \leftarrow \mathrm{EVI}\Big(L_{s,o}(c,\cdot), G_{s,o}, \varepsilon_{n+1}, 0, s\Big)$ ▷ Inner value iteration
    end for
    $v_{n+1}(s) := v_n(s) + \max_{o \in \mathcal{O}} \widehat{g}_{s,o}$ ▷ Outer value iteration
  end for
  $d_n := Gv_n$
end while
Set $g := \frac{1}{2}\Big(\max\{v_{n+1} - v_n\} + \min\{v_{n+1} - v_n\}\Big)$, $h := v_n$ and $\pi := (d_n)^\infty$

(x_t) follows the distribution of MC $Q'_{s,o}$, the remaining term (summed over $i = 1 \dots n$):

$$\sum_{i=1}^{n} \sum_{t=T_{i-1}}^{T_i - 1} \left(q_{s_i,o_i}(\cdot|s_t)^\intercal h_{k_i}^{s_i,o_i} - h_{k_i}^{s_i,o_i}(x_{t+1})\right) = \sum_{t=1}^{T_n} \left(q_{s_{N_t},o_{N_t}}(\cdot|x_t)^\intercal h_{k_{N_t}}^{s_{N_t},o_{N_t}} - h_{k_{N_t}}^{s_{N_t},o_{N_t}}(x_{t+1})\right),$$

is an MDS. It can bounded like the sum $\sum_{k=1}^{k_T} \Delta_k^{p4}$ appearing in the analysis of UCRLB (see Lem. 3.3 and 3.10), knowing that $sp\left(h_{k_{N_t}}^{s_{N_t},o_{N_t}}\right) \leq \Lambda_{s_t,o_t} \leq \Lambda_{\max}$.

Nested value iteration

If EVI is run with the *exact* Bellman operator $\mathcal{L}_k^{\mathrm{eq}}$, both Prop. 2.6 and 2.7 hold and so we obtain an efficient and *convergent* algorithm. The main challenge is that applying $\mathcal{L}_k^{\mathrm{eq}}$ requires computing the optimal gains $g_k^{s,o}(c)$ of extended MDPs $\mathcal{M}_k^{s,o}(c)$. EVI can be used to approximate these gains with an *arbitrary accuracy* $\varepsilon > 0$. We therefore propose the *nested iterative scheme* of Alg. 13 with operators $\mathcal{L}_k^{s,o}$, confidence intervals $B_p^k(s,o,s')$, and initial vector 0 as inputs (we call this algorithm NEVI for Nested Extended Value Iteration). Operator $G_{s,o}$ can be any greedy operator associated to $\mathcal{L}_k^{s,o}$. We pick a sequence of accuracies $(\varepsilon_n)_{n \geq 0}$ such that $\sum_{n \geq 0} \varepsilon_n < +\infty$. With such a sequence, we can prove the following theorem.

Theorem 6.9

If Nested Value Iteration (Alg. 13) is run with operators $\mathcal{L}_k^{s,o}$, confidence intervals $B_p^k(s,o,s')$ and if $\sum_{n \geq 0} \varepsilon_n < +\infty$, there exists $h_k^{eq} \in \mathbb{R}^{S_\mathcal{O}}$ such that $\lim_{n \to +\infty} v_n = h_k^{eq}$ and $\mathcal{L}_k^{eq} h_k^{eq} = h_k^{eq} + g_k^{eq} e$.

Proof. To simplify notations, we denote $\mathcal{L}_k^{\text{eq}}$ by $\mathcal{L}$ and we define (u_n) the sequence obtained using the same algorithm without line 4 (shift) i.e., $u_0 = v_0$ and $v_n = u_n - u_n(\overline{s})e$ for all $n \geq 1$. Prop. 2.7 shows that $\widehat{g}_{s,o}$ is an $\varepsilon_{n+1}/2$-approximation to $g_k^{s,o}$ and so for all $n \geq 0$, $\|u_{n+1} - \mathcal{L}u_n\|_\infty \leq \varepsilon_{n+1}/2$. Since $\mathcal{L}$ is non-expansive in ℓ_∞-norm (see Prop. 2.5 (b)) we have

$$\begin{aligned}\|u_{n+2} - \mathcal{L}^2 u_n\|_\infty &\leq \|u_{n+2} - \mathcal{L}u_{n+1}\|_\infty + \|\mathcal{L}u_{n+1} - \mathcal{L}^2 u_n\|_\infty \\ &\leq \varepsilon_{n+2}/2 + \|u_{n+1} - \mathcal{L}u_n\|_\infty \leq \frac{\varepsilon_{n+2} + \varepsilon_{n+1}}{2}.\end{aligned}$$

By trivial induction, $\|u_{n+k} - \mathcal{L}^k u_n\|_\infty \leq \frac{1}{2}\sum_{i=n+1}^{n+k} \varepsilon_i$ for all $n, k \geq 0$ and so

$$\begin{aligned}\|v_{n+k} - v_n\|_\infty &= \|\,(u_{n+k} - u_{n+k}(\overline{s})e) - (u_n - u_n(\overline{s})e)\,\|_\infty \\ &\leq \|\mathcal{L}^k u_n - \mathcal{L}^k u_n(\overline{s})e - (u_n - u_n(\overline{s})e)\,\|_\infty + \sum_{i=n+1}^{n+k} \varepsilon_i. \qquad (6.54)\end{aligned}$$

We know from Prop. 2.6 that $\mathcal{L}^k u_n - \mathcal{L}^k u_n(\overline{s})e$ converges as $k \to +\infty$ (as an instance of relative value iteration with initial vector u_n). A convergent sequence is a *Cauchy sequence* which means that (by definition)

$$\sup_{k \geq 0} \|\mathcal{L}^k u_n - \mathcal{L}^k u_n(\overline{s})e - (u_n - u_n(\overline{s})e)\,\|_\infty \underset{n \to +\infty}{\longrightarrow} 0.$$

Conversely, in a Banach space such that $\mathbb{R}^{S \mathcal{O}}$, any Cauchy sequence converges. Since by assumption $\sum_{n \geq 0} \varepsilon_n < +\infty$, necessarily $\sup_{k \geq 0} \left\{\sum_{i=n+1}^{n+k} \varepsilon_i\right\} = \sum_{i=n+1}^{+\infty} \varepsilon_i \underset{n \to +\infty}{\longrightarrow} 0$ and we conclude from Eq. 6.54 that (v_n) is a Cauchy sequence, and thus converges. Because $\lim_{n \to +\infty} \varepsilon_n = 0$ (otherwise the series $\sum_{n \geq 0} \varepsilon_n$ diverges), the limit of (v_n) must satisfy the optimality equation of $\mathcal{L}$. ■

One of the interesting features of NEVI is its hierarchical structure. NEVI is operating on two different *time scales* by iteratively considering every option as an independent optimistic planning sub-problem (line 8 of Alg. 13) and gathering all the results into a higher level planning problem (line 10 of Alg. 13). This idea is at the core of the *hierarchical approach* in RL, but it is not always present in the algorithmic structure, while NEVI naturally arises from decomposing EVI in two value iteration algorithms.

6.5 Numerical Experiments

In this section we compare the regrets of FSUCRL, SUCRL and UCRLB to empirically demonstrate the advantage of *temporal abstraction*.

6.5.1 Simple grid world.

In order to isolate temporal abstraction from other potential sources of improvements (e.g., number of states, diameter, etc.), we first design a domain that preserves most parameters.

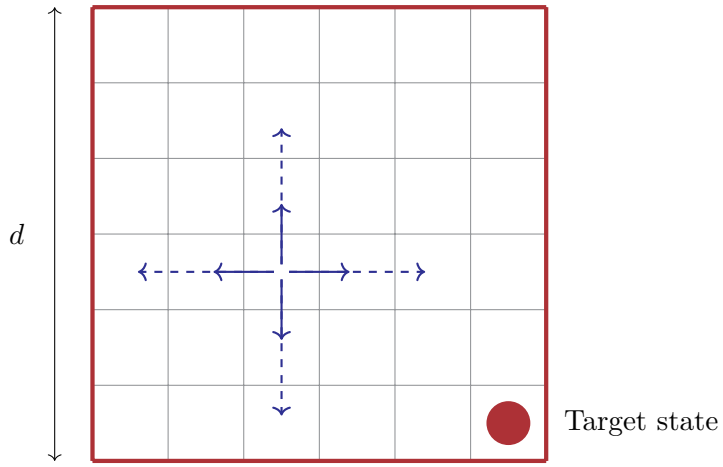


Figure 6.4: Navigation problem with the four cardinal actions represented as continuous arrows and options (temporally extended actions) of length 2 as dashed arrows.

We consider the simple navigation problem of Fig. 6.4. In any of the d^2 states of the grid except the target, the four cardinal actions are available, each of them being successful with probability 1. If the agent hits a wall then it stays in its current position with probability 1. When the target state is reached, the state is reset to any other state with uniform probability. The reward of any transition is 0 except when the agent leaves the target in which case it equals $r_{\max}$. The optimal policy simply takes the shortest path from any state to the target state. The travel-budget Λ of the MDP is equal to $r_{\max} D$ in this domain and $D = 2(d-1)$.

Let m be any non-negative integer smaller than d and in every state but the target we define four macro-actions: *LEFT*, *RIGHT*, *UP* and *DOWN* (dashed arrows in the figure). When *LEFT* is taken, primitive action *left* is applied up to m times (similar for the other three options). For any state s' which is $k \leq m$ steps on the left of the starting state s, we set $\beta_o(s') = 1/(m-k+1)$ so that the probability of the option to be interrupted after any $k \leq m$ steps is $1/m$. If the starting state s is l steps close to the left border with $l < m$ then we set $\beta_o(s') = 1/(l-k+1)$ for any state s' which is $k \leq l$ steps on the left. As a result, for all options started m steps far from any wall, $t_{\max} = m$ and $\tau_{\max} = (m+1)/2$, (while it is respectively l and $(l+1)/2$ for an option started $l < m$ step from the wall and moving towards it). More precisely, all options have an expected duration of $\tau_{\max}$ in all but in $m \times d$ states, which is small compared to the total number of d^2 states if $m \ll d$. The SMDP formed with this set of options preserves the number of state-action pairs ($S_{\mathcal{O}} = S = d^2$ and $O = A = 4$) as well as the optimal average reward $g^* = g^*_{\mathcal{O}}$, while it slightly perturbs the diameter $D_{\mathcal{O}} \leq D + m(m+1)$ (Fruit and Lazaric, 2017, Appendix F). Finally, to remove the impact of the support Γ, we consider Hoeffding rather than empirical Bernstein bounds for the transition probabilities (for all algorithms). In conclusion: ignoring the impact of temporal abstraction, the two problems seem to be almost equally hard to learn.

While a rigorous analysis of the ratio between the number of *option decision steps* n and number of *primitive actions* T_n is difficult, we notice that as d increases w.r.t. m, the chance of executing options close to a wall decreases, since for any option only $m \times d$ out of d^2 states will lead to a duration smaller than $\tau_{\max}$ and thus we can conclude that n/T_n tends to $1/\tau_{\max} = 2/(m+1)$ as n and d grow. This suggests that if d is big enough, there is always

Algorithms	Description (level of prior knowledge)
FSUCRL	Uses nested EVI to achieve optimism (no prior knowledge)
SUCRLv1	Maximal reward $r_{\max}$ and actual duration $t_{\max}$
SUCRLv2	Maximal expected duration $\tau_{\max}$, maximal variance of holding time $\sigma_\tau = \max_{s,o} \sigma_\tau(s,o)$ and reward $\sigma_R = r_{\max}\sqrt{\tau_{\max} + \sigma_\tau^2}$
SUCRLv3	$\tau_{\max}$ and $\forall s, o$, $\sigma_\tau(s,o)$ and $\sigma_R(s,o) = r_{\max}\sqrt{\tau(s,o) + \sigma_\tau(s,o)^2}$
SUCRLv4	Same as SUCRLv2 with $\sigma_R = 0$
SUCRLv5	Same as SUCRLv3 with $\sigma_R(s,o) = 0$

Table 6.1: Detailed description of the different algorithms used for the experiments. The SUCRL-like algorithms are sorted by ascending level of prior knowledge. $\sigma_\tau(s,o)$ can easily be computed exactly using an analytical formula. Note that the options are all almost surely bounded so that $\max_{s,o} b_R(s,o) = \max_{s,o} b_\tau(s,o) = 0$. All options have 0 reward so that the tightest prior knowledge we can have corresponds to $\sigma_R = 0$.

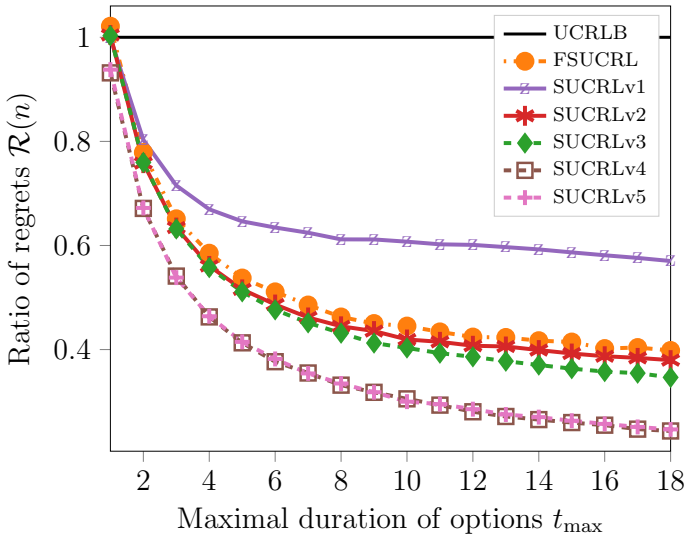


Figure 6.5: Ratio of regrets after $T_n = 2 \cdot 10^9$ steps normalized for different option durations $t_{\max}$ in a 20×20 grid-world.

an appropriate choice of m for which learning with options becomes significantly better than learning with primitive actions.

In Fig. 6.5 we plot the ratio between the regrets of SUCRL/FSUCRL and the regret of UCRLB, as $t_{\max} = m$ varies and $d = 20$. The value of T_n is fixed and chosen big enough for all d. The versions of SUCRL appearing on the plot differ in the amount of prior knowledge given to the algorithm to construct the parameters σ_R and σ_τ that are used in building the confidence intervals (see table 6.1). Unlike FSUCRL which is "parameter-free", SUCRL is highly sensitive to the prior knowledge about options and in theory, could perform even worse than UCRL2. The ratio $\mathcal{R}(n)$ decreases as m increases showing that temporal abstraction improves as $t_{\max}$ increases. This behaviour matches the theoretical predictions.

Discussion. Despite its simplicity, the most interesting aspect of this example is that the improvement on the regret is not obtained by trivially reducing the number of state-action pairs, but it is intrinsic in the way options change the dynamics of the exploration process.

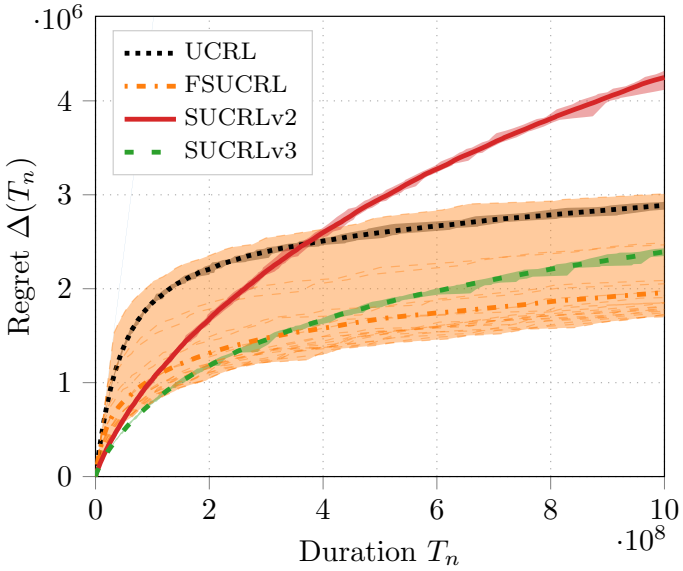


Figure 6.6: Evolution of the regret as T_n increases for a 14x14 four-rooms maze.

The two key elements in designing a successful set of options $\mathcal{O}$ is to preserve the average reward of the optimal policy and the travel-budget. The former is often a weaker condition than the latter. In this example, we achieved both conditions by designing a set $\mathcal{O}$ where the termination conditions allow any option to end after only one step. This preserves the travel-budget of the original MDP (up to a small additive term), since the agent can still navigate at the level of granularity of primitive actions. Consider a slightly different set of options $\mathcal{O}'$, where each option moves exactly by m steps (no intermediate interruption). The number of steps to the target remains unchanged from any state and thus we can achieve the optimal performance. Nonetheless, having π^* in the set of policies that can be represented with $\mathcal{O}'$ does not guarantee that the UCRL-SMDP would be as efficient in learning the optimal policy as UCRL2. In fact, the expected number of steps needed to go from a state s to an adjacent state s' may significantly increase. Despite being only one primitive action apart, there may be no sequence of options that allows to reach s' from s without relying on the random restart triggered by the target state. A careful analysis of this case shows that the travel-budget is as large as $D_{\mathcal{O}'} = D(1 + m^2)$ (Fruit and Lazaric, 2017, Appendix F).

6.5.2 Four -room maze.

We now consider the classical 4-room maze that was initially introduced by Sutton et al. (1999) to illustrate the concept of options. The domain is a grid-world of dimension 14×14 with walls separating each 7×7 "room" (see Fig. 2.1). The four cardinal actions fail with probability 0.2 (uniformly in any other direction). In every state of every room, we define four options: two are leading to the two exit doors, one is leading to the center of the room, and the last one leads to the unique corner of the grid in the room. Thus, the number of state-options is slightly bigger than the number of state-actions. The optimal policy takes the shortest path to the target state which is located in one of the 4 corners of the grid and the rewards are the same as in the previous experiment. Once the target is reached, the next

state is chosen uniformly at random in the grid.

On Fig. 6.6, we plot the regret $\Delta(\mathfrak{A}, n)$ as a function of T_n for $\mathfrak{A} \in \{$UCRL2, SUCRL, FSUCRL $\}$. The two versions of SUCRL are exactly the same as in the previous experiments: SUCRLv2 uses $\max_{s,o} \sigma_\tau(s,o)$ while SUCRLv3 uses $\sigma_\tau(s,o)$. Note that the other versions of SUCRL are not valid in this domain since the options are not almsot surely bounded. We use Bernstein bounds (as in the original versions of the algorithms presented in this thesis).

Version 2 of SUCRL fails to beat UCRL2, and it is likely that version 3 will also eventually suffer higher regret. For FSUCRL, we plot all 20 runs (as well as the average in bold). Except in one run, FSUCRL always outperforms UCRL2. The variance is clearly higher than for any other algorithm. The choice of options is probably not the best.

In both experiments, UCRL and FSUCRL had similar running times meaning that the improvement in cumulative regret is not at the expense of the computational complexity. More experiments can be found in (Fruit et al., 2017).

6.6 Conclusion

In this chapter, we started by deriving upper and lower-bounds on the regret of learning in SMDPs and we showed how these results apply to learning with options in MDPs. Comparing the regret bounds of SUCRL with UCRLB, we provided sufficient conditions on the set of options and the MDP (i.e., similar travel-budget and average reward) to reduce the regret w.r.t. learning with primitive actions. To the best of our knowledge, this is the first attempt of explaining when and how options affect the learning performance.

Then, we introduced FSUCRL, a parameter-free algorithm to learn in MDPs with options by combining the SMDP view to estimate the transition probabilities at the level of options –$b(\cdot|s,o)$– and the MDP structure of options to estimate the stationary distribution of an associated irreducible MC which allows to compute the optimistic policy at each episode. We show both theoretically and empirically that FSUCRL is actually competitive with SUCRL and it retains the advantage of temporal abstraction w.r.t. learning without options. Since FSUCRL does not require strong prior knowledge about options and its regret bound is partially computable, we believe the results of this chapter could be used as a basis to construct more principled option discovery algorithms that explicitly optimize the exploration-exploitation performance of the learning algorithm (e.g., in a transfer setting). Although FSUCRL does not require prior knowledge on sub-exponential parameters, it needs to know the outer state space $\mathcal{S}_\mathcal{O}$ (reachable states from μ_1 using only options) as well as inner state spaces $\mathcal{S}_{s,o}$ (states reachable while executing state-option (s,o)). If additional states are added to these sets, we face the same problem as with non-communicating MDPs (infinite diameter and travel-budget). Nevertheless, in this case we can apply the techniques developped in Chap. 4 for infinite diameter/travel-budget.

As future work, it would be interesting to extend the current analyses to more sophisticated hierarchical approaches to RL such as MAXQ (Dietterich, 2000).

A Appendix of Chapter 3

A.1 Bias and travel-budget

A.1.1 Proof of Thm. 3.2

If $h^* + g^* e = Lh^*$ then $L^2 h^* = L(h^* + g^* e) = Lh^* + g^* e = h^* + 2g^* e$ using the "linearity" of L (Prop. 2.5). So by induction we have $L^n h^* = h^* + ng^* e$ for all $n \geq 1$.

As shown in Prop. 2.1, for any vector $v \in \mathbb{R}^S$,

$$L^n v(s) = \max_{\pi \in \Pi^{HR}} \mathbb{E}^\pi \left[\sum_{t=1}^{n} r_t + v(s_{n+1}) \middle| s_1 = s \right] \tag{A.1}$$

Note that the maximum in (A.1) is over all history-dependent randomized policies.

Fix an arbitrary state $s' \neq s$ and define the policy $\pi' \in \Pi^{HR}$ that executes an arbitrary stationary randomized policy $\pi \in \Pi^{\text{SD}}$ as long as $t < \tau(s')$ and a greedy policy $\pi^* = (d^*)^\infty \in \Pi^{\text{SD}}$ s.t. $Lh^* = L_{d^*} h^*$ for $t \geq \tau(s')$. We denote by $n \wedge (\tau(s') - 1) := \min\{n, \tau(s') - 1\}$ the minimum between n and $\tau(s') - 1$. Due to Eq. A.1 we have:

$$\begin{aligned}
L^n h^*(s) &\geq \mathbb{E}^{\pi'} \left[\sum_{t=1}^{n} r_t + h^*(s_{n+1}) \middle| s_1 = s \right] \\
&= \mathbb{E}^{\pi'} \left[\sum_{t=1}^{n \wedge (\tau(s')-1)} r_t \middle| s_1 = s \right] + \mathbb{E}^{\pi'} \left[\sum_{t=n \wedge (\tau(s')-1)+1}^{n} r_t + h^*(s_{n+1}) \middle| s_1 = s \right] \\
&= \underbrace{\mathbb{E}^{\pi} \left[\sum_{t=1}^{n \wedge (\tau(s')-1)} r_t \middle| s_1 = s \right]}_{(1)} + \underbrace{\mathbb{E}^{\pi'} \left[\sum_{t=n \wedge (\tau(s')-1)+1}^{n} r_t + h^*(s_{n+1}) \middle| s_1 = s \right]}_{(2)}
\end{aligned} \tag{A.2}$$

The fact that we can change π' into π in the first expectation is because the MRP has the same distribution under π and π' for $t < \tau(s')$ by definition. We now analyze the second term in (A.2). Due to the Markov property, what happens for $t \geq \tau(s')$ depends only on $s_{\tau(s')} = s'$ and π^*, and not on the states, actions and rewards observed before $\tau(s')$. Mathematically,

this means that

$$\begin{aligned}(2) &= \mathbb{E}^{\pi'}\left[\sum_{t=n\wedge(\tau(s')-1)+1}^{n} r_t + h^*(s_{n+1})\middle| s_1 = s\right] \\ &= \mathbb{E}^{\pi}\left[\mathbb{E}^{\pi^*}\left[\sum_{l=1}^{n-\tau(s')+1} r_l + h^*(s_{n-\tau(s')+2})\middle| s_1 = s', \tau(s')\right]\cdot \mathbb{1}\left\{\tau(s') \leq n+1\right\}\right. \\ &\qquad\qquad \left. + h^*(s_{n+1})\cdot \mathbb{1}\left\{\tau(s') > n+1\right\}\middle| s_1 = s\right] \\ &= \mathbb{E}^{\pi}\left[L^{n-\tau(s')+1}h^*(s')\cdot \mathbb{1}\left\{\tau(s') \leq n+1\right\} + h^*(s_{n+1})\cdot \mathbb{1}\left\{\tau(s') > n+1\right\}\middle| s_1 = s\right]\end{aligned}$$

Note that it is possible to condition on $\tau(s')$ since $\tau(s')$ is a stopping time and so the sigma-algebra at stopping time $\tau(s')$ is well-defined. Since $L^n h^* = h^* + ng^* e$ for all $n \geq 1$, we have (a.s.)

$$L^{n-\tau(s')+1}h^*(s') = h^*(s') + (n - \tau(s') + 1)\cdot g^*$$

Combining these last two equalities and using the law of total expectation, we can write:

$$\begin{aligned}(2) = h^*(s')\cdot \mathbb{P}^{\pi}\left(\tau(s') \leq n+1 | s_1 = s\right) &+ \mathbb{E}^{\pi}\left[h^*(s_{n+1})\cdot \mathbb{1}\left\{\tau(s') > n+1\right\}\middle| s_1 = s\right] \\ + \; g^*\cdot \mathbb{E}^{\pi}&\left[\Big(n - \tau(s') + 1\Big)\cdot \mathbb{1}\left\{\tau(s') \leq n+1\right\}\middle| s_1 = s\right]. \end{aligned} \tag{A.3}$$

Replacing $L^n h^*(s)$ by $h^*(s) + ng^*$ in inequality A.2 and plugging (A.3) we have for all $n \geq 1$:

$$\begin{aligned}h^*(s) \geq \; \mathbb{E}^{\pi}&\left[\sum_{t=1}^{n\wedge(\tau(s')-1)} r_t\middle| s_1 = s\right] + \mathbb{E}^{\pi}\left[h^*(s_{n+1})\mathbb{1}\left\{\tau(s') > n+1\right\}\middle| s_1 = s\right] \\ &+ g^*\cdot \mathbb{E}^{\pi}\left[(n - \tau(s') + 1)\cdot \mathbb{1}\left\{\tau(s') \leq n+1\right\} - n\middle| s_1 = s\right] \\ &+ h^*(s')\cdot \mathbb{P}^{\pi}\left(\tau(s') \leq n+1 | s_1 = s\right).\end{aligned} \tag{A.4}$$

We notice that $(n - \tau(s') + 1)\cdot \mathbb{1}\left\{\tau(s') \leq n+1\right\} = n - n\wedge(\tau(s') - 1)$ and so (A.4) becomes:

$$\begin{aligned}h^*(s) \geq \; \mathbb{E}^{\pi}&\left[\sum_{t=1}^{n\wedge(\tau(s')-1)} r_t - g^*\middle| s_1 = s\right] + \mathbb{E}^{\pi}\left[h^*(s_{n+1})\mathbb{1}\left\{\tau(s') > n+1\right\}\middle| s_1 = s\right] \\ &+ \; h^*(s')\cdot \mathbb{P}^{\pi}\left(\tau(s') \leq n+1 | s_1 = s\right).\end{aligned} \tag{A.5}$$

If $\pi \in \Pi^{\mathrm{SD}}_{\mapsto s'}$ then $\tau(s')$ is a.s. finite by definition i.e., $\mathbb{P}^{\pi}(\tau(s') < +\infty) = 1$. As a consequence,

$$\lim_{n\to+\infty} \mathbb{P}^{\pi}\left(\tau(s') \leq n+1 | s_1 = s\right) = 1.$$

Since $h^*(s_{n+1})$ is bounded (by $\|h^*\|_\infty$) it also holds that

$$\left|\mathbb{E}^{\pi}\left[h^*(s_{n+1})\mathbb{1}\left\{\tau(s') > n+1\right\}\middle| s_1 = s\right]\right| \leq \|h^*\|_\infty \cdot \mathbb{P}^{\pi}\left(\tau(s') > n+1 | s_1 = s\right) \underset{n\to+\infty}{\longrightarrow} 0.$$

Finally, the term $\mathbb{E}^{\pi}\left[\sum_{t=1}^{n\wedge(\tau(s')-1)} r_t - g^*\middle| s_1 = s\right]$ tends to $\mathbb{E}^{\pi}\left[\sum_{t=1}^{\tau(s')-1} r_t - g^*\middle| s_1 = s\right]$ as n tends to infinity. We conclude the proof by taking $n \to +\infty$ in (A.5).

A.1.2 Proof of Thm. 3.3

We use the same arguments as in the previous section (App. A.1.1, proof of Thm. 3.2). We consider a policy $\pi' \in D^{\mathrm{HR}}$ which first executes $\pi \in \Pi^{\mathrm{SD}}$ until s' is visited for the first time, and then executes the non-statinary policy $\pi^+ = (d_1, \dots, d_n, \dots) \in (\Pi^{\mathrm{SR}})^{\mathbb{N}}$ such that $L_{d_{n+1}} v_n = L v_n$ for all $n \geq 1$, with $v_0 := 0$ and $v_{n+1} := L v_n$. We can write (see Eq. A.2):

$$v_n(s) := L^n v_0(s) \geq \underbrace{\mathbb{E}^{\pi}\left[\sum_{t=1}^{n\wedge(\tau(s')-1)} r_t \middle| s_1 = s\right]}_{(1)} + \underbrace{\mathbb{E}^{\pi'}\left[\sum_{t=n\wedge(\tau(s')-1)+1}^{n} r_t \middle| s_1 = s\right]}_{(2)}$$

Since $r_t \leq r_{\max}$ for all $t \geq 1$ then we can bound the first term as follows:

$$\begin{aligned}
(1) \geq \quad & \mathbb{E}^{\pi}\left[\sum_{t=1}^{\tau(s')-1} r_t - \sum_{t=n\wedge(\tau(s')-1)+1}^{\tau(s')-1} r_{\max} \middle| s_1 = s\right] \\
&= \mathbb{E}^{\pi}\left[\sum_{t=1}^{\tau(s')-1} (r_t - r_{\max}) - \sum_{t=1}^{n\wedge(\tau(s')-1)} r_{\max} \middle| s_1 = s\right] \\
&= \mathbb{E}^{\pi}\left[\sum_{t=1}^{\tau(s')-1} (r_t - r_{\max}) \middle| s_1 = s\right] + r_{\max}\mathbb{E}^{\pi}\left[n \wedge (\tau(s') - 1) \middle| s_1 = s\right].
\end{aligned}$$

Note that all the inequalities and equalities remain true even when $\tau(s')$ is not almost surely finite. In this case, the terms on the right-hand side may either be finite (convergent series) or be equal to $-\infty$, but this is a trivial lower bound to (1).

Similarly to (A.3) the second term can be exressed as follows:

$$\begin{aligned}
(2) &= \mathbb{E}^{\pi}\left[\mathbb{E}^{\pi^+}\left[\sum_{l=1}^{n-n\wedge(\tau(s')-1)} r_l \middle| s_1 = s', \tau(s')\right] \middle| s_1 = s\right] \\
&\geq \mathbb{E}^{\pi}\left[\underbrace{\mathbb{E}^{\pi^+}\left[\sum_{l=1}^{n} r_l \middle| s_1 = s'\right]}_{=v_n(s')} - r_{\max} \cdot \Big(n \wedge (\tau(s') - 1)\Big) \middle| s_1 = s\right] \\
&= v_n(s') - r_{\max} \cdot \mathbb{E}^{\pi}\left[n \wedge (\tau(s') - 1) \middle| s_1 = s\right]
\end{aligned}$$

Summing (1) and (2), the term $n \wedge (\tau(s') - 1)$ cancels and so we have

$$v_n(s) \geq v_n(s') - \mathbb{E}^{\pi}\left[\sum_{t=1}^{\tau(s')-1} r_t - r_{\max} \middle| s_1 = s\right]$$

which concludes the proof.

A.2 Concentration bounds using a martingale argument

For any $t \geq 0$, the σ-algebra induced by the past history of state-action pairs and rewards up to time t (included) is denoted $\mathcal{F}_t = \sigma(s_1, a_1, r_1, \ldots, s_t, a_t, r_t, s_{t+1})$ where by convention $\mathcal{F}_0 = \sigma(\emptyset)$ and $\mathcal{F}_\infty := \cup_{t\geq 0}\mathcal{F}_t$. Trivially, for all $t \geq 0$, $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$ and the filtration $(\mathcal{F}_t)_{t\geq 0}$ is denoted by $\mathbb{F}$. We recall that k_t is the integer-valued r.v. indexing the current episode at time t (3.12). It is immediate from the termination condition of episodes that for all $t \geq 1$, k_t is $\mathcal{F}_{t-1}$-measurable i.e., the past sequence $(s_1, a_1, r_1, \ldots, s_{t-1}, a_{t-1}, r_{t-1}, s_t)$ fully determines the ongoing episode at time t. As a consequence, the stationary (randomized) policy π_{k_t} executed at time t is also $\mathcal{F}_{t-1}$-measurable.

A.2.1 Proofs of Lem. 3.1 and 3.4

Let's consider the stochastic stochastic process $X_t := r_t(s_t, a_t) - \sum_{a\in\mathcal{A}_{s_t}} \pi_{k_t}(a|s_t) r(s_t, a)$. The term $\sum_{a\in\mathcal{A}_{s_t}} r(s_t, a)\pi_{k_t}(a|s_t)$ is $\mathcal{F}_{t-1}$-measurable and moreover

$$\mathbb{E}[r_t(s_t, a_t)|\mathcal{F}_{t-1}] = \sum_{a\in\mathcal{A}_{s_t}} \pi_{k_t}(a|s_t) r(s_t, a)$$

so that $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$. Since in addition $|X_t| \leq r_{\max}$, $(X_t, \mathcal{F}_t)_{t\geq 1}$ is a Martingale Difference Sequence (MDS) and we can apply Azuma's inequality (Prop. 3.7):

$$\mathbb{P}\left(\sum_{t=1}^{T} r_t(s_t, a_t) \leq \sum_{t=1}^{T}\sum_{a\in\mathcal{A}_{s_t}} \pi_{k_t}(a|s_t) r(s_t, a) - r_{\max}\sqrt{4T\ln\left(\frac{4T}{\delta}\right)}\right) \leq \left(\frac{\delta}{4T}\right)^2 \leq \frac{\delta}{16T^2}. \quad \text{(A.6)}$$

After taking a union bound over all possible values of $T \geq 1$, we obtain that with probability at least $1 - \sum_{T=1}^{+\infty} \frac{\delta}{16T^2} = 1 - \frac{\pi^2\delta}{96} \geq 1 - \frac{\delta}{6}$

$$\forall T \geq 1, \quad \sum_{t=1}^{T} r_t(s_t, a_t) \geq \sum_{t=1}^{T}\sum_{a\in\mathcal{A}_{s_t}} \pi_{k_t}(a|s_t) r(s_t, a) - 2r_{\max}\sqrt{T\ln\left(\frac{4T}{\delta}\right)}. \quad \text{(A.7)}$$

To prove Lem. 3.4 we consider similar stochastic processes: $r(s_t, a_t) - \sum_{a\in\mathcal{A}_{s_t}} \pi_{k_t}(a|s_t) r(s_t, a)$ and $r_{k_t}(s_t, a_t) - \sum_{a\in\mathcal{A}_{s_t}} \pi_{k_t}(a|s_t) r_{k_t}(s_t, a)$. Both are also MDS bounded by $r_{\max}$ and so we can apply Azuma's inequality, use a union bound and take the difference.

A.2.2 Proofs of Lem. 3.2 and 3.7

Let's consider the stochastic process

$$\begin{aligned} X_t &:= \alpha \sum_{a,s'} \pi_{k_t}(a|s_t) p_{k_t}(s'|s_t, a) h_{k_t}(s') - \alpha \sum_{s'} p_{k_t}(s'|s_t, a_t) h_{k_t}(s') \\ &= \alpha \sum_{a,s'} \pi_{k_t}(a|s_t) p_{k_t}(s'|s_t, a) w_{k_t}(s') - \alpha \sum_{s'} p_{k_t}(s'|s_t, a_t) w_{k_t}(s'). \end{aligned}$$

Since π_{k_t} is $\mathcal{F}_{t-1}$-measurable, $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$ and moreover $|X_t| \leq 2\alpha\|w_{k_t}\|_\infty \leq \Lambda$ a.s. for all t. $(X_t, \mathcal{F}_t)_{t\geq 1}$ is an MDS and using Azuma's inequality (Prop. 3.7):

$$\mathbb{P}\left(\sum_{t=1}^{T} X_t \geq 2\Lambda\sqrt{T\ln\left(\frac{6T}{\delta}\right)}\right) \leq \frac{\delta}{36T^2}.$$

We then notice that

$$\sum_{t=1}^{T} X_t = \alpha\sum_{k=1}^{k_T}\sum_{s,a,s'}\nu_k(s)\pi_k(a|s)p_k(s'|s,a)h_k(s') - \alpha\sum_{k=1}^{k_T}\sum_{s,a,s'}\nu_k(s,a)p_k(s'|s,a)h_k(s').$$

We proceed similarly with the stochastic process

$$X_t := \sum_{a,s'}\pi_{k_t}(s_t,a)p(s'|s_t,a)h_{k_t}(s') - \sum_{s'}p(s'|s_t,a_t)h_{k_t}(s')$$

where p_k is replaced by p and take a union bound to conclude the proof of Lem. 3.2.

To prove Lem. 3.7, we consider the same stochastic processes but we apply Freedman's inequality (Prop. 3.9) instead of Azuma's.

Let's define $\lambda_t := -\sum_{a,s'}\pi_{k_t}(a|s_t)p_{k_t}(s'|s_t,a)h_{k_t}(s')$ and $w_t = h_{k_t} + \lambda_t e$. Since by definition $\sum_{s'}p_{k_t}(s'|s_t,a_t) = 1$, we have

$$X_t = -\alpha\sum_{s'}p_{k_t}(s'|s_t,a_t)w_t(s').$$

Since $\mathbb{E}\left[X_t|\mathcal{F}_{t-1}\right] = 0$ we have:

$$\mathbb{V}\left(X_t|\mathcal{F}_{t-1}\right) = \sum_{a}\pi_{k_t}(a|s_t)\left(\alpha\sum_{s'}p_{k_t}(s'|s_t,a)w_t(s')\right)^2.$$

Proposition A.1

For any $n \geq 1$ and any n-tuple $(a_1,\ldots,a_n) \in \mathbb{R}^n$, $\left(\sum_{i=1}^n a_i\right)^2 \leq n\left(\sum_{i=1}^n a_i^2\right)$.

Proof. The statement is trivially true for $n = 1$. For $n = 2$ we have $(a_1 - a_2)^2 = a_1^2 + a_2^2 - 2a_1a_2 \geq 0$ implying that $2a_1a_2 \leq a_1^2 + a_2^2$. Therefore, $(a_1 + a_2)^2 = a_1^2 + a_2^2 + 2a_1a_2 \leq 2(a_1^2 + a_2^2)$ and so the result holds. We prove the result for $n \geq 2$ by induction. Assumed that it is true for any $n \geq 2$. Then we have:

$$\begin{aligned}\left(\sum_{i=1}^{n+1} a_i\right)^2 &= \underbrace{\left(\sum_{i=1}^{n} a_i\right)^2}_{\leq n\left(\sum_{i=1}^n a_i^2\right)} + a_{n+1}^2 + 2a_{n+1}\sum_{i=1}^{n} a_i \\ &\leq n\left(\sum_{i=1}^{n} a_i^2\right) + a_{n+1}^2 + \sum_{i=1}^{n}\underbrace{2a_ia_{n+1}}_{\leq a_i^2 + a_{n+1}^2} \leq (n+1)\cdot\left(\sum_{i=1}^{n+1} a_i^2\right)\end{aligned}$$

where the first inequality follows from the induction hypothesis and the second inequality follows from the inequality for $n = 2$ that we proved. This concludes the proof. ■

For the sake of clarity we will now use the notation $p_k(s'|s) := \sum_{a \in \mathcal{A}_s} \pi_k(a|s) p_k(s'|s,a)$ for every $s, s' \in \mathcal{S}$ and every $k \geq 1$. Using Prop. A.1 we have that

$$\begin{aligned}\mathbb{V}\left(X_t|\mathcal{F}_{t-1}\right) &\leq \alpha^2 S \sum_{a,s'} \pi_{k_t}(a|s_t) \underbrace{p_{k_t}(s'|s_t,a)^2}_{\leq p_{k_t}(s'|s_t,a)} w_{k_t}(s')^2 \\ &\leq \alpha^2 S \sum_{a,s'} \pi_{k_t}(a|s_t) p_{k_t}(s'|s_t,a) w_{k_t}(s')^2 = S \cdot \mathbb{V}_{p_{k_t}(\cdot|s_t)}\left(\alpha h_{k_t}\right)\end{aligned}$$

After applying Freedman's inequality (Prop. 3.9) to the MDS $(X_t, \mathcal{F}_t)_{t \geq 1}$ we obtain that with probability at least $1 - \frac{\delta}{12}$, for all $T \geq 1$:

$$\begin{aligned}\alpha \sum_{k=1}^{k_T} \sum_{s,a,s'} \nu_k(s)\pi_k(s,a)p_k(s'|s,a)h_k(s') \leq &\alpha \sum_{k=1}^{k_T} \sum_{s,a,s'} \nu_k(s,a)p_k(s'|s,a)h_k(s') + 2\Lambda \ln\left(\frac{48T}{\delta}\right) \\ &+ 2\sqrt{S \ln\left(\frac{48T}{\delta}\right) \sum_{t=1}^{T} \mathbb{V}_{p_{k_t}(\cdot|s_t)}\left(\alpha h_{k_t}\right)} \qquad \text{(A.8)}\end{aligned}$$

As we did before, we can do exactly the same analysis with p_k replaced by p so that with probability at least $1 - \frac{\delta}{12}$, for all $T \geq 1$:

$$\begin{aligned}-\alpha \sum_{k=1}^{k_T} \sum_{s,a,s'} \nu_k(s)\pi_k(s,a)p(s'|s,a)h_k(s') \leq &- \alpha \sum_{k=1}^{k_T} \sum_{s,a,s'} \nu_k(s,a)p(s'|s,a)h_k(s') + 2\Lambda \ln\left(\frac{48T}{\delta}\right) \\ &+ 2\sqrt{S \ln\left(\frac{48T}{\delta}\right) \sum_{t=1}^{T} \mathbb{V}_{\overline{p}_{k_t}(\cdot|s_t)}\left(\alpha h_{k_t}\right)} \qquad \text{(A.9)}\end{aligned}$$

with the notation $\overline{p}_k(s'|s) := \sum_{a \in \mathcal{A}_s} \pi_k(a|s) p(s'|s,a)$ for every $s, s' \in \mathcal{S}$ and $k \geq 1$. To conclude the proof of Lem. 3.7 we take a union bound.

A.2.3 Proofs of Lem. 3.3 and 3.10

Let's consider the stochastic process

$$X_t := \alpha \sum_{a,s'} \pi_{k_t}(a|s_t) p(s'|s_t,a) w_{k_t}(s') - \alpha w_{k_t}(s_{t+1}).$$

Once action $a_t \sim \pi_{k_t}(a|s_t)$ has been sampled, the next state is sampled according to the distrbutiion $s_{t+1} \sim p(\cdot|s_t,a)$. Thus, $\mathbb{E}\left[X_t|\mathcal{F}_{t-1}\right] = 0$ and $|X_t| \leq 2\alpha\|w_{k_t}\|_\infty \leq \Lambda$. Using Azuma's inequality (Prop. 3.7):

$$\mathbb{P}\left(\sum_{t=1}^{T} X_t \geq 2\Lambda\sqrt{T \ln\left(\frac{4T}{\delta}\right)}\right) \leq \frac{\delta}{16T^2}.$$

and we conclude the proof of Lem. 3.3 as usual (see previous sections).

The conditional variance can be written as

$$\mathbb{V}\left(X_t|\mathcal{F}_{t-1}\right) = \mathbb{V}_{\overline{p}_{k_t}(\cdot|s_t)}\left(\alpha h_k\right)$$

Using Freedman's inequality we have that with probability at least $1-\frac{\delta}{6}$:

$$\sum_{t=1}^{T}\Delta_k^{p4} \leq 2\sqrt{\left(\sum_{t=1}^{T}\mathbb{V}_{\overline{p}_{k_t}(\cdot|s_t)}\left(h_k\right)\right)\cdot\ln\left(\frac{24T}{\delta}\right)} + 4\Lambda\ln\left(\frac{24T}{\delta}\right) \tag{A.10}$$

which concludes the proof of Lem. 3.10.

A.2.4 Proofs of Lem. 3.9

Let's now consider the stochastic process

$$X_t := \mathbb{V}_{\widehat{p}_{k_t}(\cdot|s_t,a_t)}\left(\alpha h_{k_t}\right) - \sum_a \pi_{k_t}(a|s_t)\mathbb{V}_{\widehat{p}_{k_t}(\cdot|s_t,a)}\left(\alpha h_{k_t}\right).$$

$(X_t,\mathcal{F}_t)_{t\geq 1}$ is an MDS. Since $\mathbb{V}_{\widehat{p}_{k_t}(\cdot|s_t,a_t)}\left(\alpha h_{k_t}\right) \geq 0$ and $sp\left(h_{k_t}\right) \leq \Lambda/\alpha$, it follows from Prop. 3.10 that $|X_t| \leq \Lambda^2/4$. Applying Azuma's inequality (Prop. 3.7), we have that with probability at least $1-\frac{\delta}{6}$, for all $T \geq 1$:

$$\sum_{t=1}^{T}\mathbb{V}_{\widehat{p}_{k_t}(\cdot|s_t,a_t)}\left(h_{k_t}\right) \leq \sum_{t=1}^{T}\mathbb{V}_{\widehat{p}_{k_t}(\cdot|s_t)}\left(h_{k_t}\right) + 2\Lambda^2\sqrt{T\ln\left(\frac{4T}{\delta}\right)}.$$

A.3 Proofs of Lem. 3.5 and 3.8 (Cauchy-Schwartz)

Denote by $\mathcal{S}_k(s,a) := \{s' \in \mathcal{S} : \ \widehat{p}_k(s'|s,a) > 0\}$ the set of observed next states starting from s when playing a, and $\Gamma_k(s,a) := |\mathcal{S}_k(s,a)| = \|\widehat{p}_k(s'|s,a)\|_0$ the cardinal of $\mathcal{S}_k(s,a)$. By Cauchy-Schwartz inequality[1]

$$\begin{aligned}
\sum_{s'\in\mathcal{S}}\sqrt{\widehat{p}_k(s'|s,a)(1-\widehat{p}_k(s'|s,a))} &= \sum_{s'\in\mathcal{S}_k(s,a)}\sqrt{\widehat{p}_k(s'|s,a)(1-\widehat{p}_k(s'|s,a))} \\
&\leq \sqrt{\left(\sum_{s'\in\mathcal{S}_k(s,a)}\widehat{p_k}(s'|s,a)\right)\cdot\left(\sum_{s'\in\mathcal{S}_k(s,a)}1-\widehat{p_k}(s'|s,a)\right)} \\
&\leq \sqrt{\Gamma_k(s,a)-1} \leq \sqrt{\Gamma_k(s,a)-1}.
\end{aligned}$$

Note that the observed next states $s' \in \mathcal{S}_k(s,a)$ necessarily satisfy $p(s'|s,a) > 0$ and so $\Gamma_k(s,a) \leq \Gamma(s,a)$. This concludes the prooff of Lem. 3.5.

[1] The inequality obtained is somehow tight since when $\widehat{p}_k(\cdot|s,a)$ is uniform on its support, it becomes an equality.

Using Cauchy-Schartz inequality we have:

$$\sum_{s'\in\mathcal{S}}\sqrt{\widehat{p}_k(s'|s,a)(1-\widehat{p}_k(s'|s,a))w_k^s(s')^2} = \sum_{s'\in\mathcal{S}_k(s,a)}\sqrt{\widehat{p}_k(s'|s,a)(1-\widehat{p}_k(s'|s,a))w_k^s(s')^2}$$
$$\leq \sqrt{\left(\sum_{s'\in\mathcal{S}_k(s,a)} 1-\widehat{p}_k(s'|s,a)\right)\cdot\left(\sum_{s'\in\mathcal{S}_k(s,a)}\widehat{p}_k(s'|s,a)w_k^s(s')^2\right)}$$
$$= \sqrt{(\Gamma_k(s,a)-1)\cdot\left(\sum_{s'\in\mathcal{S}}\widehat{p}_k(s'|s,a)w_k^s(s')^2\right)} \leq \sqrt{(\Gamma(s,a)-1)\cdot\sum_{s'\in\mathcal{S}}\widehat{p}_k(s'|s,a)w_k^s(s')^2}$$

By definition, $\alpha^2\sum_{s'\in\mathcal{S}}\widehat{p}_k(s'|s,a)w_k^s(s')^2 = V_k(s,a)$ which concludes the proof of Lem. 3.8.

A.4 Proof of Lem. 3.6

We slightly change our notations and denote by $N_t(s,a)$ the number of visits in state-action pair (s,a) *strictly* before t (i.e., t *not* included). With this convention, what was denoted $N_k(s,a)$ (3.14) actually corresponds to $N_{t_k}(s,a)$. The stopping condition of episodes ensures that for all $t\geq 1$, $N_t(s,a)\leq 2N_{k_t}(s,a)$. Therefore, similarly to what is done in (Ouyang et al., 2017a, Proof of Lemma 5)

$$\sum_{k=1}^{k_T}\frac{\nu_k(s,a)}{N_k^+(s,a)} \leq 2\sum_{t=1}^{T}\frac{\mathbb{1}\{s_t=s,a_t=a\}}{N_t^+(s,a)}$$
$$= 2\Bigg[\mathbb{1}\{N_{T+1}(s,a)\geq 1\} + \underbrace{\sum_{j=1}^{N_{T+1}(s,a)-1}\frac{1}{j}}_{\leq 1+\ln(N_{T+1}(s,a))\mathbb{1}\{N_{T+1}(s,a)\geq 1\}}\Bigg]$$
$$\leq 2+2\ln\left(N_{T+1}^+(s,a)\right) \tag{A.11}$$

where (A.11) follows from the rate of divergence of an harmonic series.

We proceed similarly for the second series:

$$\sum_{k=1}^{k_T}\frac{\nu_k(s,a)}{\sqrt{N_k^+(s,a)}} \leq \sqrt{2}\sum_{t=1}^{T}\frac{\mathbb{1}\{s_t=s,a_t=a\}}{\sqrt{N_t^+(s,a)}}$$
$$= \sqrt{2}\left(\mathbb{1}\{N_{T+1}(s,a)\geq 1\} + \underbrace{\sum_{j=1}^{N_{T+1}(s,a)-1}\frac{1}{\sqrt{j}}}_{\leq 2\sqrt{N_{T+1}(s,a)-1}-1}\right)$$
$$\leq 2\sqrt{2}\sqrt{N_{T+1}(s,a)-1} \leq 3\sqrt{N_{T+1}(s,a)}.$$

B Appendix of Chap. 4

B.1 Number of episodes

The stopping condition of episodes used by TUCRL (4.11) combines the original stopping condition of UCRLB with the condition $s_{t+1} \in \mathcal{S}^{\mathtt{T}}_{k_t}$. Using only the fact that $\nu_k(s,a) \geq N_k(s,a)$ for at least one pair (s,a), Jaksch et al. (2010, Proposition 18) proved that for $T \geq SA$, the number of episodes is bounded by $\log_2\left(\frac{8T}{SA}\right)$ (Prop. 3.8). The total number of episodes in TUCRL can be bounded by the same quantity (with S replaced by $S^{\mathtt{C}}$ since no sate in $\mathcal{S}^{\mathtt{T}}$ is ever visited) plus the number of times the event $s_{t+1} \in \mathcal{S}^{\mathtt{T}}_{k_t}$ occurs. Since whenever $s_{t+1} \in \mathcal{S}^{\mathtt{T}}_{k_t}$ state s_{t+1} is removed from $\mathcal{S}^{\mathtt{T}}_{k_t+1}$ and s_{t+1} necessarily belongs to $\mathcal{S}^{\mathtt{C}}$ (by definition), this event can happen at most $\mathcal{S}^{\mathtt{C}}$ times. We thus have:

$$\forall T \geq SA, \quad k_T \leq S^{\mathtt{C}} A \log_2\left(\frac{8T}{S^{\mathtt{C}}A}\right) + S^{\mathtt{C}}. \tag{B.1}$$

B.2 Proof of Thm. 4.2

We prove the following lemma used in the proof of Thm. 4.2.

Lemma B.1

For all $x \in]0, 1/10]$, we have $(1-x)^{1/x} \geq 1/3$.

Proof. It is easy to verify that the derivative of $x \longmapsto (1-x)^{1/x}$ is:

$$\forall x \in]0, 1/10], \quad \frac{d}{dx}\left((1-x)^{1/x}\right) = -\underbrace{\frac{(1-x)^{1/x-1}}{x^2}}_{\geq 0} \cdot ((1-x)\ln(1-x) + x)$$

It is well known that for all $x \in]0,1[$, $x < -\ln(1-x) < \frac{x}{1-x}$ implying that $(1-x)\ln(1-x)+x$ is positive. Therefore, $\frac{d}{dx}\left((1-x)^{1/x}\right)$ is negative on $]0, 1/10]$ implying that $x \longmapsto (1-x)^{1/x}$ is decreasing. As a result: $\forall x \in]0, 1/10], \quad (1-x)^{1/x} \geq 0.9^{10} > 1/3$. ■

C Appendix of Chap. 5

C.1 Projection on a semi-ball (proof of Lem. 5.4)

Let $v \in \mathbb{R}^S$ and define $u = \Gamma_c v$. If $sp(v) \leq c$ then $u = v$ and so the result holds.
If $sp(v) > c$ then for all $s \in \mathcal{S}$ such that $v(s) > \min_x v(x) + c$ we have $u(s) = \min_x v(x) + c$ (there exists at least one such state since $sp(v) > c$) while for all other states we have $u(s) = v(s)$ (there also exists at least one such state). This implies that $\max\{v - \Gamma_c v\} = \max_x v(x) - \min_x v(x) - c = sp(v) - c$ and $\min\{v - \Gamma_c v\} = 0$. As a result, we also have

$$sp(u - v) = sp(v - u) = \max\{v - \Gamma_c v\} - \min\{v - \Gamma_c v\} = sp(v) - c - 0 = sp(v) - c.$$

For any vector $z \in \mathcal{B}_c$ i.e., such that $sp(z) \leq c$, by reverse triangle inequality[1] we have that:

$$sp(z - u) \geq sp(u) - sp(z) \geq sp(u) - c = sp(w - u)$$

which concludes the proof.

C.2 Aperiodicity transformation (proof of Lem. 5.3)

We prove a slightly more general result.

Theorem C.1

Let P be any stochastic matrix and H_P its associated deviation matrix i.e., the Drazin inverse of $I - P$: $H_P := (I - P + P^*)^{-1}(I - P^*)$ (see Sec. 2.2). For any $0 \leq \alpha < 1$ we denote by $P_\alpha := (1 - \alpha)P + \alpha I$ the aperiodic transform of P with parameter α. The deviation matrix of P_α can be expressed as $H_{P_\alpha} = 1/(1 - \alpha)H_P$.

Proof. Let A be a square matrix and assume there exists a matrix $A^\#$ that satisfies the following properties:

[1] The triangle inequality for the span is proved in (Puterman, 1994, Section 6.6.1).

- $AA^{\#}A = A$
- $AA^{\#} = A^{\#}A$
- $A^{\#}AA^{\#} = A^{\#}$

then $A^{\#}$ is the Drazin inverse of A. We know from App. A of Puterman (1994) that these properties hold for $A = P$ and $A^{\#} = H_P$.
By definition: $I - P_\alpha = (1-\alpha)P + \alpha I - I = (1-\alpha)(I-P)$. Based on this result and using the properties of P and H_P, we can derive the same relations for P_α and H_{P_α}:

$$\begin{aligned}(I-P_\alpha)\,(1/(1-\alpha)H_P)\,(I-P_\alpha) &= (1-\alpha)(I-P)\cdot(1/(1-\alpha)H_P)\cdot(1-\alpha)(I-P)\\ &= (1-\alpha)(I-P) = (I-P_\alpha)\end{aligned}$$

$$\begin{aligned}(I-P_\alpha)\,(1/(1-\alpha)H_P) &= (1-\alpha)(I-P)\cdot(1/(1-\alpha)H_P) = (I-P)H_P = H_P(I-P)\\ &= (1/(1-\alpha)H_P)\,(1-\alpha)(I-P) = (1/(1-\alpha)H_P)\,(I-P_\alpha)\end{aligned}$$

$$\begin{aligned}(1/(1-\alpha)H_P)\,(I-P_\alpha)\,(1/(1-\alpha)H_P) &= (1/(1-\alpha)H_P)\cdot(1-\alpha)(I-P)\cdot(1/(1-\alpha)H_P)\\ &= 1/(1-\alpha)H_P(I-P)H_P = 1/(1-\alpha)H_P\end{aligned}$$

In conclusion: $H_{P_\alpha} = 1/(1-\alpha)H_P$. ■

As a consequence of Thm. C.1 and by definition of the bias (Eq. 6.3), for any $\pi = d^\infty \in \Pi^{\mathrm{SR}}$ we have: $h^\pi_M = H_{P_d} r_d$ and $h^\pi_{M_\alpha} = H_{P^d_\alpha} r^d_\alpha$. The aperiodicity transformation applies only to the transition kernel of the MDP not the reward, so $r^d_\alpha = r_d$ and $h^\pi_{M_\alpha} = 1/(1-\alpha)h^\pi_M$.

C.3 Operator of SCAL* (proof of Lem. 5.13)

We start the proof of Lem. 5.13 with a simple definition.

Definition C.1

Let $B \subseteq \Delta_S$ be a non-empty compact convex subset of the probability simplex, $q \in B$ a probability vector in B and $\beta \geq 0$ a positive scalar. For all vectors $v \in \mathbb{R}^S$ we define $B^q_\beta(v) := \{p \in B :\ p^\intercal v \leq q^\intercal v + \beta\} \subseteq B$.

Since $q \in B$ and $\beta \geq 0$ by assumption, $q \in B^q_\beta(v)$ for all $v \in \mathbb{R}^S$ and so $B^q_\beta(v)$ is never empty.

For any vector $v \in \mathbb{R}^S$, we define $p_v \in \arg\max_{p \in B^q_\beta(v)} p^\intercal v$ (we drop the dependency in β and q for simplicity) and $\overline{p}_v \in \arg\max_{p \in B} p^\intercal v$. Since $B \subseteq B_\beta(v)$, $p_v^\intercal v \leq \overline{p}_v^\intercal v$. The following lemma provides a sufficient condition for the equality to hold.

Lemma C.1

If $p_v^\intercal v < q^\intercal v + \beta$ *then* $p_v^\intercal v = \overline{p}_v^\intercal v$.

Proof. We define the function $f : [0,1] \mapsto \mathbb{R}$ mapping x to $((1-x)\cdot p_v + x \cdot \overline{p}_v)^\intercal v$. Since B is convex, for all $x \in [0,1]$, $f(x) \in B$. By assumption, $f(0) = p_v^\intercal v < q^\intercal v + \beta$. If in addition we assume that $\overline{p}_v^\intercal v > p_v^\intercal v$ then f is strictly increasing. If $f(1) = \overline{p}_v^\intercal v \leq q^\intercal v + \beta$ then $\overline{p}_v \in B_\beta^q(v)$ by definition implying that $p_v^\intercal v = \overline{p}_v^\intercal v$ which contradicts the assumption that $\overline{p}_v^\intercal v > p_v^\intercal v$ and so $f(1) > q^\intercal v + \beta$. By the *intermediate value theorem*, $\exists \overline{x} \in [0,1]$ s.t. $f(\overline{x}) = q^\intercal v + \beta$ and so obviously $(1-\overline{x})\cdot p_v + \overline{x} \cdot \overline{p}_v \in B_\beta^q(v)$. Since p_v achieves the maximum value of $p^\intercal v$ for all $p \in B_\beta^q(v)$, this contradicts the assumption that $p_v^\intercal v < q^\intercal v + \beta$, implying that $p_v^\intercal v = \max_{p \in B} p^\intercal v = \overline{p}_v^\intercal v$. In conclusion, under the assumption that $p_v^\intercal v < q^\intercal v + \beta$, necessarily $p_v^\intercal v = \overline{p}_v^\intercal v$. ■

Thanks to Lem. C.1, we know that whenever the constraint $p^\intercal v \leq q^\intercal v + \beta$ is strict, the maximum of $p^\intercal v$ over $B_\beta^q(v)$ matches the maximum over B. We deduce the following lemma.

Lemma C.2

If $u, v \in \mathbb{R}^S$ *and* $v \leq u$ *then*

$$\max_{p \in B_\beta^q(v)} p^\intercal v \leq \max_{p \in B_\beta^q(u)} p^\intercal u.$$

Proof. We distinguish two possible cases:

1. If $p_u^\intercal u < q^\intercal u + \beta$:
   From Lem. C.1, we have that $p_u^\intercal u = \overline{p}_u^\intercal u \geq p_v^\intercal u \geq p_v^\intercal v$. The first inequality follows from the fact that $\overline{p}_u$ is the argmax over all $p \in B$ and $p_v \in B$, while the second inequality follows from the fact that $u \geq v$ (by assumption).

2. If $p_u^\intercal u = q^\intercal u + \beta$:
   $p_u^\intercal u = q^\intercal u + \beta \geq q^\intercal v + \beta \geq p_v^\intercal v$ where the first inequality follows from the assumption $u \geq v$ and the second inequality is a consequence of the fact that $p_v \in B_\beta^q(v)$ by definition.

■

If we take $B \leftarrow B_p^k(s,a)$, $q \leftarrow \widehat{p}_k(\cdot|s,a)$ and $\beta \leftarrow c\beta_k^{sa}$ all the requirements of Def. C.1 are satisifed and all the above lemmas hold with $B_\beta^q(v) \leftarrow B_p^k(s,a) \cap \Theta_p^k(s,a,v)$. Given the definition of $\mathfrak{L}_k$ (see Eq. 5.34), it is immediate to see that the monotonicity of $\mathfrak{L}_k$ is a direct consequence of Lem. C.2. The linearity simply follows from the fact that $p^\intercal(v+\lambda e) = p^\intercal v + \lambda e$ for all $p \in \Delta_S$ and $\lambda \in \mathbb{R}$. To prove the non-expansiveness of $\mathfrak{L}_k$, we denote $v(s^+) - u(s^+) := \max_{s \in \mathcal{S}} \{v(s) - u(s)\}$ and $v(s^-) - u(s^-) := \min_{s \in \mathcal{S}} \{v(s) - u(s)\}$. By definition,

$$u + (v(s^-) - u(s^-))e \leq v \leq u + (v(s^+) - u(s^+))e$$
$$\implies \mathfrak{L}_k u + (v(s^-) - u(s^-))e \leq \mathfrak{L}_k v \leq \mathfrak{L}_k u + (v(s^+) - u(s^+))e$$

where the implication is a direct application of the monotonicity and linearity of $\mathfrak{L}_k$. It follows that:

$$\max_{s\in\mathcal{S}}\{\mathfrak{L}_k v(s) - \mathfrak{L}_k u(s)\} \leq v(s^+) - u(s^+),$$
$$\min_{s\in\mathcal{S}}\{\mathfrak{L}_k v(s) - \mathfrak{L}_k u(s)\} \geq v(s^-) - u(s^-).$$

In conclusion, $sp\,(\mathfrak{L}_k v - \mathfrak{L}_k u) \leq sp\,(v-u)$. Using the fact that

$$\|v-u\|_\infty = \max\{v(s^+) - u(s^+), u(s^-) - v(s^-)\},$$

we deduce that $\|\mathfrak{L}_k v - \mathfrak{L}_k u\|_\infty \leq \|v-u\|_\infty$.

If we replace $B_p^k(s,a)$ by $\widetilde{B}_p^k(s,a)$ and $\widehat{p}_k(\cdot|s,a)$ by $\widetilde{p}_k(\cdot|s,a)$, the requirements of Def. C.1 are still satisifed and so we can prove the same results for $\widetilde{\mathfrak{L}}_k$.

C.4 Perturbation of SCAL* operator (proof of Lem. 5.14)

We use the same notations as in App. C.3 above.

To prove that $\widetilde{\mathfrak{L}}_k$ is a $(1-\eta_k)$-contraction we first prove the following lemma.

Lemma C.3

For all $u, v \in \mathbb{R}^S$ there exists $p_{u,v} \in B$ such that

$$\max_{p\in B_\beta^q(v)} p^\intercal v - \max_{p\in B_\beta^q(u)} p^\intercal u \leq p_{u,v}^\intercal (v-u).$$

Proof. We distinguish between two cases:

1. If $p_u^\intercal u < q^\intercal u + \beta$:
   From Lem. C.1, we have that $p_u^\intercal u = \overline{p}_u^\intercal u \geq p_v^\intercal u$. We deduce that $p_v^\intercal v - p_u^\intercal u \leq p_v^\intercal (v-u)$. Since $p_v \in B$, we can take $p_{u,v} \leftarrow p_v$.
2. If $p_u^\intercal u = q^\intercal u + \beta$:
   $p_v^\intercal v - p_u^\intercal u = p_v^\intercal v - (q^\intercal u + \beta) \leq q^\intercal v + \beta\!\!\!/ - q^\intercal u - \beta\!\!\!/ = q^\intercal (v-u)$ where the inequality holds because $p_v \in B_\beta^q(v)$. Since $q \in B$, we can take $p_{u,v} \leftarrow q$.

■

Just like Lem. C.2, Lem. C.3 can be applied to operators $\mathfrak{L}_k$ and $\widetilde{\mathfrak{L}}_k$. In the case of $\widetilde{\mathfrak{L}}_k$, $B = \widetilde{B}_p^k(s,a) \subseteq \{p \in \Delta_S : p(\bar{s}) \geq \eta\}$ where $\bar{s} \in \mathcal{S}$ is an arbitrary reference state and $\eta > 0$. We then use similar arguments as Puterman (1994, Theorem 6.6.6). Let's denote $\widetilde{\mathfrak{L}}_k$ by L (for the sake of clarity) and $Lv(s^+) - Lu(s^+) := \max_{s\in\mathcal{S}}\{Lv(s) - Lu(s)\}$ and $Lv(s^-) - Lu(s^-) := \min_{s\in\mathcal{S}}\{Lv(s) - Lu(s)\}$. Applying Lem. C.2, we obtain that

$$Lv(s^+) - Lu(s^+) \leq {p_{u,v}^+}^\intercal (v-u) \quad \text{and} \quad Lu(s^-) - Lv(s^-) \leq {p_{v,u}^-}^\intercal (u-v)$$

where $p^+_{u,v},\ p^-_{v,u} \in B$ and in particular $p^+_{u,v}(\overline{s}),\ p^-_{v,u}(\overline{s}) \geq \eta$. More generally, for any $s \in \mathcal{S}$, we can bound $Lv(s) - Lu(s)$ using corresponding vectors $p^s_{u,v}$ and $p^s_{v,u}$. If we concatenate all the S probability vectors, we obtain two transition matrices $P_{u,v}$ and $P_{v,u}$. Like in the proof of Theorem 6.6.6 of Puterman (1994) we have

$$\begin{aligned} sp\left(Lv - Lu\right) \leq {p^+_{u,v}}^\intercal (v-u) - {p^-_{u,v}}^\intercal (v-u) &\leq \max_{s\in\mathcal{S}} {p^s_{u,v}}^\intercal (v-u) - \min_{s\in\mathcal{S}} {p^s_{v,u}}^\intercal (v-u) \\ &\leq sp\left(\begin{bmatrix} P_{u,v} \\ P_{v,u} \end{bmatrix} (v-u)\right) \leq (1-\eta) sp\left(v-u\right). \end{aligned}$$

The last inequality follows from Proposition 6.6.1 of Puterman (1994) and the fact that $B \subseteq \{p \in \Delta_S : p(\overline{s}) \geq \eta\}$.

To quantify the impact of the perturbation, we rely on the proof of Lem. 5.10 (Fruit et al., 2018b, Lemma 19, Appendix E). We denote by $\widetilde{p}_v \in \arg\max_{p\in\widetilde{B}^q_\beta(v)} p^\intercal v$ and $\widetilde{\widetilde{p}}_v \in \arg\max_{p\in\widetilde{B}} p^\intercal v$ (a tilde indicates an η-perturbation). We bound the difference $p_v^\intercal v - \widetilde{p}_v^\intercal v$ (the opposite can be bounded in the same way).

1. If $\widetilde{p}_v^\intercal v < \widetilde{q}^\intercal v + \beta$:
   Lem. C.2 shows that $\widetilde{p}_v^\intercal v = \widetilde{\widetilde{p}}_v^\intercal v$ and so

   $$p_v^\intercal v - \widetilde{p}_v^\intercal v = p_v^\intercal v - \widetilde{\widetilde{p}}_v^\intercal v \leq \overline{p}_v^\intercal v - \widetilde{\widetilde{p}}_v^\intercal v \leq \|p_v - \widetilde{p}_v\|_1 \times \frac{sp\left(v\right)}{2} \leq \eta \cdot sp\left(v\right)$$

   where the last inequality is proved in (Fruit et al., 2018b, Lemma 19, Appendix E).

2. If $\widetilde{p}_v^\intercal v = \widetilde{q}^\intercal v + \beta$:

   $$p_v^\intercal v - \widetilde{p}_v^\intercal v = p_v^\intercal v - \widetilde{q}^\intercal v - \beta \leq q^\intercal v + \cancel{\beta} - \widetilde{q}^\intercal v - \cancel{\beta} \leq \|q - \widetilde{q}\|_1 \times \frac{sp\left(v\right)}{2} \leq \eta \cdot sp\left(v\right)$$

   where the last inequality follows from the fact that $\widetilde{q}$ is an ℓ_1-projection of q onto $\widetilde{B}$ (see Sec. 5.8).

D Appendix of Chapter 6

D.1 Sub-exponential options (proof of Lem.6.1)

We use the second definition of sub-exponential r.v. in Def. 6.2. In the following we drop the notation s, o and denote by τ the random realization of the holding time and $\overline{\tau}$ its expectation. Using eq. 6.11, the Laplace transform of the holding time can be computed as follows:

$$\mathbb{E}\left[e^{\lambda(\tau-\overline{\tau})}\right] = \sum_{k=1}^{\infty} e^{\lambda(k-\overline{\tau})} e_s^\intercal Q^{k-1} V e = e^{\lambda(1-\overline{\tau})} e_s^\intercal \left[\sum_{k=0}^{\infty} \left(e^{\lambda} Q\right)^k\right] V e$$

The term $\sum_{k=0}^{\infty} \left(e^{\lambda} Q\right)^k$ is finite if and only if $e^{\lambda}\rho(Q) < 1$, in which case we have:

$$\mathbb{E}\left[e^{\lambda(\tau-\overline{\tau})}\right] = e^{\lambda(1-\overline{\tau})} e_s^\intercal \left(I - e^{\lambda} Q\right)^{-1} V e,$$

and otherwise $\mathbb{E}\left[e^{\lambda(\tau-\overline{\tau})}\right] = +\infty$. Note that $e^{\lambda}\rho(Q) < 1$ if and only if either $\lambda < -\log\left(\rho(Q)\right)$ or $\rho(Q) = 0$. We will now analyse the two cases separately:

1. $\rho(Q) = 0$ if and only if all the eigenvalues of Q in $\mathbb{C}$ are 0, if and only if Q is nilpotent ($\exists n > 0$ s.t. $Q^n = 0$). This is because Q can always be triangularized in $\mathbb{C}$: $Q = UTU^{-1}$ where T is upper-triangular with the eigenvalues of Q on the diagonal that is, only zeros if $\rho(Q) = 0$. This implies that $\exists n > 0$ s.t. $T^n = U^{-1}Q^n U = 0 \implies Q^n = 0$ hence Q is nilpotent. The reverse is obviously true: if Q is nilpotent then $\rho(Q) = 0$, (otherwise there would exist $\lambda \neq 0$, $v \neq 0$ and $n > 0$ s.t. $Q^n = 0$ and $Qv = \lambda v \implies Q^n v = \lambda^n v = 0$, which is absurd). By definition, matrix Q is nilpotent of order n if and only if the Markov Chain reaches an absorbing state in at most n steps (a.s.). In conclusion, $\rho(Q) = 0$ if and only if the option is almost surely bounded. This happens if and only if there is no cycle in the option (with probability 1, every non-absorbing state is visited at most once).

2. In the case where $\rho(Q) > 0$: it is clear that $\mathbb{E}\left[e^{\lambda(\tau-\overline{\tau})}\right]$ can not be bounded by a function of the form $\lambda \to e^{\frac{\sigma^2\lambda^2}{2}}$ for $\lambda \geq -\log\left(\rho(Q)\right)$ so τ is not sub-Gaussian (Definition 6.3). However, since $\rho(Q) < 1$ we can choose $0 < c_0 < -\log\left(\rho(Q)\right)$ and we have $\mathbb{E}\left[e^{\lambda(\tau-\overline{\tau})}\right] < +\infty$ for all $|\lambda| < c_0$, which implies that τ is sub-exponential (Definition

6.2).

In conclusion, either the option contains inner-loops (some states are visited several times with non-zero probability) in which case the distribution of τ is sub-Exponential but not sub-Gaussian, or it has no inner-loop in which case τ is bounded (and thus sub-Gaussian). There is no other alternative.

The distribution of rewards R is not as simple: the reward of an option is the sum of all micro-rewards obtained at every time step before the option ends, and every micro-reward earned at each time step can have a different distribution. The only constraint is that all micro-rewards should be (a.s.) bounded between 0 and $r_{\max}$. As a result, if τ is a.s. bounded (by let's say $t_{\max}$) then R is also a.s. bounded (by $r_{\max}t_{\max}$). But if τ is unbounded then R may still be bounded if for example, all micro-rewards are 0. If however all micro-rewards are equal to $r_{\max}$ then R has a discrete phase-type distribution just like τ. R can thus be unbounded (and even not sub-Gaussian). However, we will show that R is always sub-Exponential. Using the law of total expectations and the fact that $p\left(R \leq r_{\max}\tau\right) = 1$ we have:

$$\begin{aligned}
\forall \lambda > 0, \quad \mathbb{E}\left[e^{\lambda(R-\overline{R})}\right] = \sum_{k=1}^{\infty} \mathbb{E}\left[e^{\lambda(R-\overline{R})}|\tau = k\right] p(\tau = k) &\leq \sum_{k=1}^{\infty} \mathbb{E}\left[e^{\lambda(r_{\max}\tau-\overline{R})}|\tau = k\right] p(\tau = k) \\
&= \sum_{k=1}^{\infty} \mathbb{E}\left[e^{\lambda(r_{\max}k-\overline{R})}|\tau = k\right] p(\tau = k) \\
&= \sum_{k=1}^{\infty} e^{\lambda(r_{\max}k-\overline{R})} p(\tau = k) \\
&= e^{\lambda(r_{\max}-\overline{R})} e_s^{\intercal}\left[\sum_{k=0}^{\infty}\left(e^{\lambda r_{\max}}Q\right)^k\right] Ve
\end{aligned}$$

We can now conclude as we did for τ: let $0 < c_0 < -\frac{\log(\rho(Q))}{r_{\max}}$, for all $0 < \lambda < c_0$ the quantity $\mathbb{E}\left[e^{\lambda(R-\overline{R})}\right]$ is finite. Note that for $\lambda \leq 0$: $e^{\lambda R} \leq 1$ so $\mathbb{E}\left[e^{\lambda(R-\overline{R})}\right] < +\infty$. According to Def. 6.2, R is sub-Exponential.

D.2 Comparision of the MDP-sample complexity and the SMDP-sample complexity

Let M be a MDP and $\mathcal{O}$ a set of options on M. We denote by $M_{\mathcal{O}}$ the SMDP formed by M and $\mathcal{O}$. If an option is chosen at time step t, we denote by τ_t the (random) duration of that option. The set of all time steps is $\mathbb{N}$ ($t = 1, 2, ...$) and the set of time steps corresponding to decision steps (i.e., when an option is started) is denoted by $\mathcal{T}$. The set $\mathcal{T}$ is a random variable since it depends on the duration of the options. $\mathcal{T}$ is defined recursively: $\mathcal{T} = \{1, \tau_1 + 1, \tau_1 + 1 + \tau_{\tau_1+1}, , ...\} \subseteq \mathbb{N}$. The first option is taken at time $t_1 = 1$, the second option is taken at the end of the first option (i.e. at time $t_2 = \tau_1 + 1$, where τ_1 is a random variable), and so on. The i-th option is played at time step $t_i \in \mathcal{T}$ recursively defined as: $t_{i+1} = t_i + \tau_{t_i}$ and $t_1 = 1$.

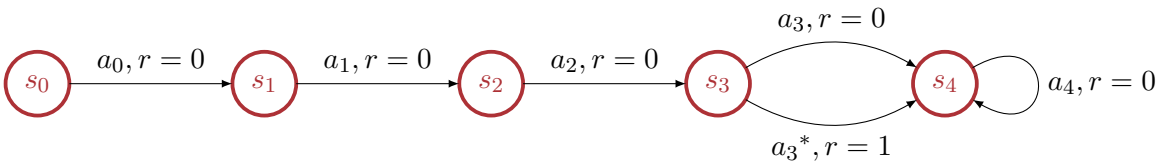


Figure D.1: MDP of counter-example 1.

For any learning algorithm $\mathfrak{A}$ on an MDP M, the MDP-sample complexity is defined as:

$$\sum_{t=0}^{+\infty} \mathbb{1}\Big\{v_\gamma^{\mathfrak{A}_t}(s_t) \leq v_\gamma^*(s_t) - \epsilon\Big\}. \tag{D.1}$$

Let's assume that algorithm $\mathfrak{A}$ is SMDP-RMAX (Brunskill and Li, 2014) applied to SMDP $M_\mathcal{O}$ formed by MDP M and option set $\mathcal{O}$. $\mathfrak{A}$ can indeed be seen as a learning algorithm on M (see Lem. 6.2) and so the sample complexity given in (D.1) is correctly defined. However, we can also choose to "ignore" what is happening within an option (we only look at the epochs, i.e. times $t \in \mathcal{T}$). Thus, we can also interpret $\mathfrak{A}$ as a learning algorithm on the SMDP $M_\mathcal{O}$. The corresponding SMDP-sample complexity is defined as (Brunskill and Li, 2014):

$$\sum_{t\in\mathcal{T}} \tau_t \mathbb{1}\Big\{v_\gamma^{\mathfrak{A}_t}(s_t) \leq v_\mathcal{O}^*(s_t) - \epsilon\Big\} = \sum_{i=0}^{+\infty} \tau_{t_i} \mathbb{1}\Big\{v_\gamma^{\mathfrak{A}_{t_i}}(s_{t_i}) \leq v_\mathcal{O}^*(s_{t_i}) - \epsilon\Big\}. \tag{D.2}$$

Brunskill and Li (2014) use the quantity given in equation (D.2) instead of the quantity given in equation (D.1) to derive the final bound on their algorithm (Theorem 3). The implicit assumption is that the following inequality holds:

$$\sum_{t=0}^{+\infty} \mathbb{1}\Big\{v_\gamma^{\mathfrak{A}_t}(s_t) \leq v_\gamma^*(s_t) - \epsilon\Big\} \leq \sum_{t\in\mathcal{T}} \tau_t \mathbb{1}\Big\{v_\gamma^{\mathfrak{A}_t}(s_t) \leq v_\mathcal{O}^*(s_t) - \epsilon\Big\} \tag{D.3}$$

Inequality (D.3) should hold with probability 1 (or at least with probability $1 - \delta$) for Theorem 3 to hold true. This requirement is never mentioned in the article. We give two counter-examples showing that this inequality will not hold in general (not even with high probability), even if we assume that the set of options is optimal, i.e. if $v_\mathcal{O}^* = v_\gamma^*$. The problem arises when the algorithm is ϵ-optimal at an epoch but there exists at least a step before the next epoch where the algorithm is not ϵ-optimal.

D.2.1 Counter-example 1

In this example we have: $\mathcal{S} = \{s_0, s_1, s_2, s_3, s_4\}$ and $\mathcal{A} = \{a_0, a_1, a_2, a_3, a_3{}^*, a_4\}$. We assume the MDP is fully deterministic: $p(s_1|s_0, a_0) = p(s_2|s_1, a_1) = p(s_3|s_2, a_2) = p(s_4|s_3, a_3) = p(s_4|s_3, a_3{}^*) = p(s_4|s_4, a_4) = 1$. The graph of the MDP is represented on Figure D.1. We define R as follows:

- $r(s_1|s_0,a_0) = r(s_2|s_1,a_1) = r(s_3|s_2,a_2) = r(s_4|s_3,a_3) = r(s_4|s_4,a_4) = 0$,
- $r(s_4|s_3,{a_3}^*) = 1$.

We define policy $\pi \in \Pi^{\text{SD}}$ by: $\pi(s_3) = a_3$ (we don't need to specify the actions taken in other states since there is only one possible action in those states). The optimal policy is such that: $\pi^*(s_3) = {a_3}^*$. Trivially we have:

- $v^*_\gamma(s_0) = \gamma^3, v^\pi_\gamma(s_0) = 0$,
- $v^*_\gamma(s_3) = 1, v^\pi_\gamma(s_3) = 0$.

Now if we set $1 \geq \epsilon > \gamma^3$, we have: $v^\pi_\gamma(s_0) > v^*_\gamma(s_0) - \epsilon$ and $v^\pi_\gamma(s_3) \leq v^*_\gamma(s_3) - \epsilon$. In other words, π is ϵ-optimal in s_0 but not in s_3.

Let's define two options, o and o^*, by:

- $\mathcal{I}_o = \mathcal{I}_{o^*} = \{s_0\}$,
- $\beta_o(s) = 0$ if $s \neq s_4$ and 1 otherwise, $\beta_{o^*} = \beta_o$,
- $\pi_o = \pi$ and $\pi_{o^*} = \pi^*$.

If we denote by $M_{\mathcal{O}}$ the SMDP formed by M and the set of options $\mathcal{O} = \{o, o^*, a_4\}$, we have that $v^*_{\mathcal{O}} = v^*_\gamma$ ($\mathcal{O}$ is an optimal set of options). Suppose we execute a SMDP-learning algorithm $\mathfrak{A}$ (e.g. SMDP-RMAX) which starts in s_0. The SMDP-sample complexity is always 0 because both o and o^* are ϵ-optimal in s_0, and a_4 is optimal in s_4. However, the MDP-sample complexity of $\mathfrak{A}$ is at least equal to 1 if option o is taken (and equals 0 when o^* is chosen instead). There is no reason that the algorithm should select o^* rather than o (the SMDP is initially unknown). So we can not guarantee that the event $\{o$ is taken in $s_0\}$ will happen with probability lower than δ. In this example, the SMDP-sample complexity is not upper bounding the MDP-sample complexity (not even with high probability).

It is true that γ is usually close to one (e.g. $\gamma = 0.9$) and thus, in the previous example, ϵ cannot be too small ($\epsilon > \gamma^3 \simeq 0.73$). But it is possible to consider longer options in a bigger MDP and apply the same kind of reasoning. We would then obtain $\epsilon \geq \gamma^k$ with γ^k sufficiently small when k is sufficiently big.

One might argue that our example is not relevant since the cumulated reward does not increase after s_4 is reached (i.e. after step $t = 4$) because we have an absorbing state. But it is possible to make some minor changes and assume for example that action a_4 leads back to state s_0 instead of looping on state s_4. We would then need to choose $1 \geq \epsilon > \frac{\gamma^3}{1-\gamma^5}$, or more generally $1 \geq \epsilon > \frac{\gamma^k}{1-\gamma^{k+2}}$ with a longer chain, because in that case:

- $v^*_\gamma(s_0) = \frac{\gamma^k}{1-\gamma^{k+2}},\ v^\pi_\gamma(s_0) = 0$,
- $v^*_\gamma(s_k) = \frac{1}{1-\gamma^{k+2}},\ v^\pi_\gamma(s_k) = 0$.

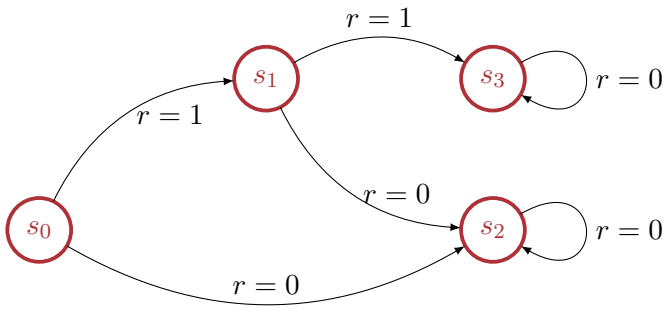


Figure D.2: Graph of the MDP of Example 2

Assuming $\mathfrak{A}$ starts with option o:

- $v_\gamma^{\mathfrak{A}_0}(s_0) \geq 0 > v_\gamma^*(s_0) - \epsilon$,
- $v_\gamma^{\mathfrak{A}_k}(s_k) \leq \dfrac{\gamma^{k+2}}{1-\gamma^{k+2}} \leq v_\gamma^*(s_k) - \epsilon$.

The SMDP-sample complexity is again 0 because in s_0 the algorithm is necessarily ϵ-optimal since $0 > v_\gamma^*(s_0) - \epsilon$. But the MDP sample complexity is at least 1 if $\mathfrak{A}_0(s_0) = o$. The same holds as long as o is chosen at least once in s_0 (not necessarily at $t_0 = 0$). Indeed, at the iteration where o is chosen in s_0, the value function in s_k is upper bounded by $\dfrac{\gamma^{k+2}}{1-\gamma^{k+2}} \leq v_\gamma^*(s_k) - \epsilon$.

D.2.2 Counter-example 2

In Example 1, we had deterministic transitions. We now consider the case where the transitions are random. To simplify the calculations, we assume that $\gamma = 1$. The graph of the MDP is represented on Figure D.2. In this example we have: $\mathcal{S} = \{s_0, s_1, s_2, s_3\}$ and $\mathcal{A} = \{a_0, {a_0}^*, a_1, {a_1}^*, a_2, a_3\}$. Let $1/2 > \epsilon > 0$. We define p as follows:

- action a_0: $p(s_1|s_0, a_0) = 1/3$, $p(s_2|s_0, a_0) = 2/3$,
- action a_0^*: $p(s_1|s_0, a_0^*) = \dfrac{1+\epsilon}{3+2\epsilon}$, $p(s_2|s_0, a_0^*) = \dfrac{2+\epsilon}{3+2\epsilon}$,
- action a_1: $p(s_2|s_1, a_1) = p(s_3|s_1, a_1) = 1/2$,
- action a_1^*: $p(s_2|s_1, a_1^*) = 1/2 - \epsilon$, $p(s_3|s_1, a_1^*) = 1/2 + \epsilon$.

We define r as follows:

- $r(s_1|s_0, a_0) = r(s_1|s_0, a_0^*) = r(s_3|s_1, a_1) = r(s_3|s_1, a_1^*) = 1$,
- $r(s_2|s_0, {a_0}^*) = r(s_2|s_0, a_0^*) = r(s_2|s_1, a_1) = r(s_2|s_1, {a_1}^*) = r(s_2|s_2, a_2) = r(s_3|s_3, a_3) = 0$.

Note that in this example all rewards depend only on the initial and final state of the transitions (it does not depend on the action taken). There are four deterministic policies:

- policy π_1: $\pi_1(s_0) = a_0$, $\pi_1(s_1) = a_1$,
- policy π_2: $\pi_2(s_0) = a_0^*$, $\pi_2(s_1) = a_1^*$,

- policy π_3: $\pi_3(s_0) = a_0^*$, $\pi_3(s_1) = a_1$,
- policy π_4: $\pi_4(s_0) = a_0$, $\pi_4(s_1) = a_1^*$.

The value functions associated to these policies are:

- $v_\gamma^{\pi_1}(s_0) = \frac{1}{2}$,
- $v_\gamma^{\pi_2}(s_0) = \frac{1+\epsilon}{2}$,
- $v_\gamma^{\pi_3}(s_0) = \frac{1+\epsilon}{2+4/3\epsilon}$,
- $v_\gamma^{\pi_4}(s_0) = \frac{1}{2} + \frac{1}{3}\epsilon$,
- $v_\gamma^{\pi_1}(s_1) = v_\gamma^{\pi_3}(s_1) = \frac{1}{2}$,
- $v_\gamma^{\pi_2}(s_1) = v_\gamma^{\pi_4}(s_1) = \frac{1}{2} + \epsilon$.

We deduce that: $\pi^* = \pi_2$, $v_\gamma^*(s_0) = \frac{1+\epsilon}{2}$ and $v_\gamma^*(s_1) = \frac{1}{2} + \epsilon$. Furthermore: $v_\gamma^{\pi_1}(s_0) > v_\gamma^*(s_0) - \epsilon$ and $v_\gamma^{\pi_1}(s_1) \leq v_\gamma^*(s_1) - \epsilon$. In other words, π_1 is ϵ-optimal in s_0 but not in s_1.

Let's define the following options o and o^*:

- $\mathcal{I}_o = \mathcal{I}_{o^*} = \{s_0\}$,
- $\beta_o(s_2) = \beta_o(s_3) = 1$ and 0 otherwise, $\beta_{o^*} = \beta_o$,
- $\pi_o = \pi_1$ and $\pi_{o^*} = \pi^* = \pi_2$.

If we denote by $M_\mathcal{O}$ the SMDP formed by M and the set of options $\mathcal{O} = \{o, o^*, a_2, a_3\}$, we have that $v_\mathcal{O}^* = v_\gamma^*$ ($\mathcal{O}$ is an optimal set of options). Suppose we execute a SMDP-learning algorithm $\mathfrak{A}$ which starts in s_0. The SMDP-sample complexity is always 0 because both o and o^* are ϵ-optimal in s_0, and a_2 and a_3 are optimal in s_2 and s_3 respectively. However, the MDP-sample complexity of $\mathfrak{A}$ is equal to 1 when option o is taken and s_1 is reached (when option o is chosen in s_0, s_1 is reached with probability 1/3). There is no reason that the algorithm should select o^* rather than o (the SMDP is initially unknown). So we can not guarantee that o will be chosen with probability lower than δ. So the expected MDP-sample complexity will be (strictly) positive. In this example, the SMDP-sample complexity is not upper bounding the MDP-sample complexity.

It should be possible to change Example 2 as we did with Example 1 to make sure that the cumulated reward keeps increasing (i.e. delete all absorbing states), but this is likely to require tedious calculations. The purpose of this example was only to show that the MDP-sample complexity can be higher than its SMDP analogue for other reasons than the presence of a discounting factor (namely, the presence of random transitions).

D.2.3 Conclusion

In the previous examples we have shown that in an MDP with options, the SMDP-sample complexity does not upperbound the MDP-sample complexity in general, even when the set of option is optimal. It is not difficult to show that the opposite is also true: the MDP-sample complexity does not upperbound the SMDP-sample complexity. Thus, without any additional assumptions on the options and/or the algorithm, it is not possible to use the SMDP-sample complexity to prove that options can be beneficial to learn a MDP.

Bibliography

Abbasi-Yadkori, Y. and Szepesvári, C. (2015). Bayesian optimal control of smoothly parameterized systems. In *UAI*, pages 1–11. AUAI Press. 42

Abbasi-Yadkori, Y. and Szepesvári, C. (2015). Bayesian optimal control of smoothly parameterized systems. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, pages 2–11, Arlington, Virginia, United States. AUAI Press. 93

Agrawal, S. and Jia, R. (2017). Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *NIPS*, pages 1184–1194. 42, 86, 87

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2007). Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory*, pages 150–165, Berlin, Heidelberg. Springer Berlin Heidelberg. 52

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.*, 410(19):1876–1902. 53

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (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. 92

Auer, P. and Ortner, R. (2007). Logarithmic online regret bounds for undiscounted reinforcement learning. In Schölkopf, B., Platt, J. C., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 49–56. MIT Press. 43

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*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272, International Convention Centre, Sydney, Australia. PMLR. 69, 70, 79, 86, 120, 121

Bacon, P.-L. and Precup, D. (2015). The option-critic architecture. In *NIPS'15 Deep Reinforcement Learning Workshop*. 157

Bartlett, P. L. and Tewari, A. (2009). REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *UAI*, pages 35–42. AUAI Press. 17, 41, 47, 48, 61, 64, 65, 92, 119

Bellemare, M. G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. (2016). Unifying count-based exploration and intrinsic motivation. In *NIPS*, pages 1471–1479. 120

Bellman, R. (1954). The theory of dynamic programming. *Bull. Amer. Math. Soc.*, 60(6):503–515. 13, 15, 17

Bertsekas, D. P. (1995). *Dynamic programming and optimal control. Vol II.* Number 2. Athena scientific Belmont, MA. 31, 34, 127

Bertsekas, D. P. (2007). *Dynamic Programming and Optimal Control, Vol. II.* Athena Scientific, 3rd edition. 19

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford. 52, 53

Bremaud, P. (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues.* Springer-Verlag Inc, Berlin; New York. 21, 183, 184

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). OpenAI Gym. *ArXiv e-prints*. 92

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym. *CoRR*, abs/1606.01540. 110

Brunskill, E. and Li, L. (2014). PAC-inspired Option Discovery in Lifelong Reinforcement Learning. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *JMLR Proceedings*, pages 316–324. JMLR.org. 158, 168, 170, 217

Burnetas, A. N. and Katehakis, M. N. (1997). Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255. 38, 39, 41, 42, 47

Castro, P. S. and Precup, D. (2012). Automatic construction of temporally extended actions for mdps using bisimulation metrics. In *Proceedings of the 9th European Conference on Recent Advances in Reinforcement Learning*, EWRL'11, pages 140–152, Berlin, Heidelberg. Springer-Verlag. 157

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA. 92

Şimşek, O. and Barto, A. G. (2004). Using relative novelty to identify useful temporal abstractions in reinforcement learning. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04. 157

Dai, F. Z. and Walter, M. R. (2019). Maximum expected hitting cost of a markov decision process and informativeness of rewards. *CoRR*, abs/1907.02114. 63

Dann, C. and Brunskill, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS 15, pages 2818–2826. MIT Press. 53, 56, 70

Dann, C., Lattimore, T., and Brunskill, E. (2017). Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *NIPS*, pages 5717–5727. 59

Dietterich, T. G. (2000). Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of Artificial Intelligence Research*, 13:227–303. 180, 198

Filippi, S., Cappé, O., and Garivier, A. (2010). Optimism in Reinforcement Learning and Kullback-Leibler Divergence. This work has been accepted and presented at ALLERTON 2010. 48

Freedman, D. A. (1975). On tail probabilities for martingales. *Ann. Probab.*, 3(1):100–118. 78

Fruit, R. and Lazaric, A. (2017). Exploration-Exploitation in MDPs with Options. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 576–584, Fort Lauderdale, FL, USA. PMLR. 158, 170, 179, 180, 195, 197

Fruit, R., Pirotta, M., and Lazaric, A. (2018a). Near optimal exploration-exploitation in non-communicating markov decision processes. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 2994–3004. Curran Associates, Inc. 92, 93

Fruit, R., Pirotta, M., Lazaric, A., and Brunskill, E. (2017). Regret minimization in mdps with options without prior knowledge. In *NIPS*, pages 3169–3179. 158, 166, 198

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. (2018b). Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1573–1581. 121, 127, 128, 129, 133, 136, 138, 147, 213

Garivier, A., Ménard, P., and Stoltz, G. (Sep. 2018). Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*. 41

Givan, R., Leach, S., and Dean, T. (2000). Bounded-parameter markov decision processes. *Artificial Intelligence*, 122(1):71 – 109. 35

Gopalan, A. and Mannor, S. (2015). Thompson sampling for learning parameterized markov decision processes. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 861–898. JMLR.org. 93

Grinstead, C. M. and Snell, J. L. (2003). *Introduction to Probability*. AMS, 2 edition. 21, 164

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600. 16, 29, 35, 40, 41, 45, 46, 47, 48, 49, 55, 57, 58, 62, 63, 66, 67, 70, 71, 75, 85, 117, 140, 150, 179, 207

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is q-learning provably efficient? *CoRR*, abs/1807.03765. 120, 121

Jong, N. K., Hester, T., and Stone, P. (2008). The utility of temporal abstraction in reinforcement learning. In *The Seventh International Joint Conference on Autonomous Agents and Multiagent Systems*. 158

Kakade, S., Wang, M., and Yang, L. F. (2018). Variance Reduction Methods for Sublinear Reinforcement Learning. *ArXiv e-prints*. 70, 79, 86

Kakade, S., Wang, M., and Yang, L. F. (2018). Variance reduction methods for sublinear reinforcement learning. *CoRR*, abs/1802.09184. 120, 121

Kaufmann, E., Cappe, O., and Garivier, A. (2012). On bayesian upper confidence bounds for bandit problems. In Lawrence, N. D. and Girolami, M., editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 592–600, La Palma, Canary Islands. PMLR. 15

Kirkland, S. J., Neumann, M., and Sze, N.-S. (2008). On optimal condition numbers for markov chains. *Numerische Mathematik*, 110(4):521–537. 43

Lai, T. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1):4–22. 14

Lakshmanan, K., Ortner, R., and Ryabko, D. (2015). Improved regret bounds for undiscounted continuous reinforcement learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 524–532, Lille, France. PMLR. 155

Lattimore, T. and Hutter, M. (2012). Pac bounds for discounted mdps. In *In Proc. 23rd International Conf. on Algorithmic Learning Theory (ALT'12), volume 7568 of LNAI*. Springer. 53, 70, 79

Lattimore, T. and Hutter, M. (2014). Near-optimal pac bounds for discounted mdps. *Theoretical Computer Science*, 558:125–143. 53, 70, 79, 86

Lattimore, T. and Szepesvári, C. (2018). Bandit algorithms. Pre-publication version. 54, 59

Levy, K. Y. and Shimkin, N. (2011). Unified inter and intra options learning using policy gradient methods. In Sanner, S. and Hutter, M., editors, *EWRL*, volume 7188 of *Lecture Notes in Computer Science*, pages 153–164. Springer. 157

Lewis, M., Ayhan, H., and D. Foley, R. (1999). Bias optimality in a queue with admission control. *Probability in the Engineering and Informational Sciences*, 13. 38

Lewis, M. E. and Puterman, M. L. (2002). *Bias Optimality*, chapter 2, pages 89–111. Springer US, Boston, MA. 31, 38

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., editors, *Advances in Neural Information Processing Systems 27*, page 1835–1843. Curran Associates, Inc. 79

Mann, T. A., Mankowitz, D. J., and Mannor, S. (2014). Time-regularized interrupting options (TRIO). In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1350–1358. JMLR.org. 157

Mann, T. A. and Mannor, S. (2014). Scaling up approximate value iteration with options: Better policies with fewer iterations. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 127–135. JMLR.org. 157, 158

Martin, J., Sasikumar, S. N., Everitt, T., and Hutter, M. (2017). Count-based exploration in feature space for reinforcement learning. *CoRR*, abs/1706.08090. 120

Maurer, A. and Pontil, M. (2009). Empirical bernstein bounds and sample-variance penalization. In *COLT*. 52

McGovern, A. and Barto, A. G. (2001). Automatic discovery of subgoals in reinforcement learning using diverse density. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 361–368. 157

Menache, I., Mannor, S., and Shimkin, N. (2002). Q-cut—dynamic discovery of sub-goals in reinforcement learning. In *Proceedings of the 13th European Conference on Machine Learning, Helsinki, Finland, August 19–23, 2002*, pages 295–306. Springer Berlin Heidelberg. 157

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015a). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529. 91

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015b). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533. 14

Moore, A. W. (1990). Efficient memory-based learning for robot control. Technical report. 91, 92

Munos, R. and Moore, A. (1999). Influence and variance of a markov chain: Application to adaptive discretization in optimal control. In *Proceedings: International Astronomical Union Transactions, v. 16B p*, pages 355–362. 79

Nielsen, B. F. (2012). Lecture notes on phase–type distributions for stochastic processes. 166

Ok, J., Proutière, A., and Tranos, D. (2018). Exploration in structured reinforcement learning. In *NeurIPS*, pages 8888–8896. 39, 40, 41, 43, 47

Ortner, R. (2008). Optimism in the face of uncertainty should be refutable. *Minds and Machines*, 18(4):521–526. 89, 90, 141

Ortner, R. (2010). Online regret bounds for markov decision processes with deterministic transitions. *Theor. Comput. Sci.*, 411(29-30):2684–2695. 45

Ortner, R. (2016). Some open problems for average reward mdps. In *European Workshop on Reinforcement Learning*. https://ewrl.files.wordpress.com/2016/12/ortner.pdf. 16

Ortner, R. (2018). Regret Bounds for Reinforcement Learning via Markov Chain Concentration. *arXiv e-prints*, page arXiv:1808.01813. 41, 48, 87

Ortner, R. and Ryabko, D. (2013). Online regret bounds for undiscounted continuous reinforcement learning. *CoRR*, abs/1302.2550. 155

Osband, I. and Roy, B. V. (2016). Posterior sampling for reinforcement learning without episodes. *CoRR*, abs/1608.02731. 93

Osband, I. and Roy, B. V. (2017). Why is posterior sampling better than optimism for reinforcement learning? In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 2701–2710. PMLR. 42, 87

Osband, I., Russo, D., and Roy, B. V. (2013). (more) efficient reinforcement learning via posterior sampling. In *NIPS*, pages 3003–3011. 42, 93

Osband, I. and Van Roy, B. (2016). On Lower Bounds for Regret in Reinforcement Learning. *arXiv e-prints*, page arXiv:1608.02732. 41

Ostrovski, G., Bellemare, M. G., van den Oord, A., and Munos, R. (2017). Count-based exploration with neural density models. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 2721–2730. PMLR. 120

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. (2017a). Learning unknown markov decision processes: A thompson sampling approach. In *NIPS*, pages 1333–1342. 42, 206

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. (2017b). Learning unknown markov decision processes: A thompson sampling approach. In *Advances in Neural Information Processing Systems 30*, pages 1333–1342. Curran Associates, Inc. 93

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA. 19, 21, 22, 23, 24, 26, 27, 28, 29, 30, 32, 33, 34, 36, 40, 47, 58, 62, 93, 94, 95, 101, 122, 127, 132, 134, 135, 136, 162, 209, 210, 212, 213

Qian, J., Fruit, R., Pirotta, M., and Lazaric, A. (2018a). Concentration inequalities for multinoulli random variables. https://rlgammazero.github.io/docs/JFPL2018notesPSRL.pdf. 87

Qian, J., Fruit, R., Pirotta, M., and Lazaric, A. (2018b). Exploration bonus for regret minimization in undiscounted discrete and continuous markov decision processes. *CoRR*, abs/1812.04363. 121, 148, 149, 155

Sairamesh, M. and Ravindran, B. (2012). Options with exceptions. In *Proceedings of the 9th European Conference on Recent Advances in Reinforcement Learning*, EWRL'11, pages 165–176, Berlin, Heidelberg. Springer-Verlag. 157

Schweitzer, P. J. (1985). On undiscounted markovian decision processes with compact action spaces. *RAIRO - Operations Research - Recherche Opérationnelle*, 19(1):71–86. 27, 161

Schweitzer, P. J. and Federgruen, A. (1979). Geometric convergence of value-iteration in multichain markov decision problems. *Advances in Applied Probability*, 11(1):188–217. 57

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489. 14

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., and Graepel, T. (2017). Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *arXiv e-prints*, page arXiv:1712.01815. 14

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, 550:354–. 14

Sorg, J. and Singh, S. P. (2010). Linear Options. In *AAMAS*, pages 31–38. 157

Stolle, M. and Precup, D. (2002). Learning options in reinforcement learning. In *SARA*, volume 2371 of *Lecture Notes in Computer Science*, pages 212–223. Springer. 157

Strehl, A. L. and Littman, M. L. (2008a). An analysis of model-based interval estimation for markov decision processes. *J. Comput. Syst. Sci.*, 74:1309–1331. 46

Strehl, A. L. and Littman, M. L. (2008b). An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331. 120

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. Adaptive computation and machine learning. MIT Press, second edition. 116

Sutton, R. S., Precup, D., and Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181 – 211. 157, 159, 166, 197

Talebi, M. S. and Maillard, O. (2018a). Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *ALT*, volume 83 of *Proceedings of Machine Learning Research*, pages 770–805. PMLR. 79

Talebi, M. S. and Maillard, O.-A. (2018b). Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In Janoos, F., Mohri, M., and Sridharan, K., editors,

Proceedings of Algorithmic Learning Theory, volume 83 of *Proceedings of Machine Learning Research*, pages 770–805. PMLR. 48

Tang, H., Houthooft, R., Foote, D., Stooke, A., Chen, X., Duan, Y., Schulman, J., Turck, F. D., and Abbeel, P. (2017). #exploration: A study of count-based exploration for deep reinforcement learning. In *NIPS*, pages 2750–2759. 120

Tessler, C., Givony, S., Zahavy, T., Mankowitz, D. J., and Mannor, S. (2016). A deep hierarchical approach to lifelong learning in minecraft. *CoRR*, abs/1604.07255. 157

Tewari, A. and Bartlett, P. L. (2007a). Bounded parameter markov decision processes with average reward criterion. In Bshouty, N. H. and Gentile, C., editors, *Learning Theory*, pages 263–277, Berlin, Heidelberg. Springer Berlin Heidelberg. 35

Tewari, A. and Bartlett, P. L. (2007b). Optimistic linear programming gives logarithmic regret for irreducible mdps. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pages 1505–1512, USA. Curran Associates Inc. 43

Theocharous, G., Wen, Z., Abbasi-Yadkori, Y., and Vlassis, N. (2017). Posterior sampling for large scale reinforcement learning. *CoRR*, abs/1711.07979. 93

Thompson, W. R. (1933a). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294. 14

Thompson, W. R. (1933b). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294. 42

Tossou, A., Basu, D., and Dimitrakakis, C. (2019). Near-optimal Optimistic Reinforcement Learning using Empirical Bernstein Inequalities. *arXiv e-prints*, page arXiv:1905.12425. 41, 48, 87

Von Neumann, J. and Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton University Press. 13, 22

Wainwright, M. (2015). *Course on Mathematical Statistics*, chapter 2: Basic tail and concentration bounds. University of California at Berkeley, Department of Statistics. 167, 168, 177

Weissman, T., Ordentlich, E., Seroussi, G., Verdú, S., and Weinberger, M. J. (2003). Inequalities for the l 1 deviation of the empirical distribution. 52, 53