Adaptive Pure Exploration in Markov Decision Processes and Bandits

Aymen Al Marjani


THESE

en vue de l'obtention du grade de Docteur, délivré par

l'ECOLE NORMALE SUPERIEURE DE LYON

Ecole Doctorale N° 512
École Doctorale en Informatique, Mathématiques de Lyon (InfoMaths)

Discipline : Mathématiques

Soutenue publiquement le 06/12/2023, par :

Aymen AL MARJANI

Adaptive Pure Exploration in Markov Decision Processes and Bandits

Exploration Pure Adaptative dans les Processus de Décision Markoviens et les bandits

---

Devant le jury composé de :

NOWAK, Robert	Professeur	Univ. of Wisconsin-Madison	Rapporteur
GAST, Nicolas	Chargé de recherche	INRIA	Rapporteur
GAUJAL, Bruno	Directeur de recherche	INRIA	Examinateur
JONSSON, Anders	Professeur	Univ. Pompeu Fabra	Examinateur
PIKE-BURKE, Ciara	Maitresse de Conférences	Imperial College London	Examinatrice
GARIVIER, Aurélien	Professeur des universités	CNRS, ENS de Lyon	Directeur de thèse
KAUFMANN, Emilie	Chargée de recherche	CNRS, Univ. de Lille	Co-directrice

إلى والدي العزيزين

To my beloved parents

Acknowledgments - Remerciements

L'aventure mathématique que relate ce manuscrit a commencé lors de mon stage de master, suite à ma découverte de l'article *"Optimal Best Arm Identification with Fixed Confidence"*, rédigé par ceux qui allaient devenir mes directeurs de thèse. Je me souviens qu'en lisant la formule du temps charactéristique pour la première fois, j'avais l'impression d'apprendre une sorte de loi fondamentale de l'exploration, à l'image du deuxième principe de la thermodynamique. En effet, je trouvais -et trouve toujours- fascinant le fait que même un oracle qui connaît le bandit multi-bras auquel il a affaire, a besoin d'échantillons pour garantir qu'il retourne le meilleur bras. Merci à tous les deux pour cette source d'inspiration, ainsi que pour votre disponibilité et votre bonne humeur tout au long des trois années de thèse qui s'en suivirent. Aurélien, merci de m'avoir transmis le sens de la persévérance dans la recherche (En témoigne le long parcours du papier all-epsilon :)). Merci aussi pour tout ce temps consacré à lire ensemble le papier du Simulator et à démystifier sa technique de preuve (c'est pas sorcier après tout !). Enfin, ce fut toujous un régal de suivre tes exposés sur des sujets mathématiques assez variés, allant de la "differential privacy" jusqu'à l'estimateur de Good-Turing en passant par les bornes minimax de tests statistiques. Émilie, merci pour ton encadrement infaillible: que ce soit pour nos très fructueuses séances de brainstorming algorithmique sur le tableau (CovGame en témoigne :)), la relecture (extrêmement méticuleuse !) de mes preuves ou encore ton feedback toujours pertinent sur chacun de mes exposés. Merci aussi pour ton acceuil chaleureux à chaquefois que je te visitais à SCOOL. C'est ainsi que j'ai découvert que ta maîtrise des bandits dépasse le monde de la recherche et va jusqu'au jeu de Bang !

Before and during my PhD, I had the chance to collaborate with some amazing researchers to whom I am much obliged: Achraf, Alexandre, Andrea, Deb, Marc and Tomáš. Alexandre, merci pour ton accueil chaleureux à Stockholm, pour m'avoir introduit au monde de la recherche et encouragé à poursuivre une thèse en ML. Andrea, I have always admired your impressive academic writing skills and your ability to take a step back, look at the bigger picture then formulate a more relevant problem to solve or a more ambitious result to aim for. Thanks for teaching me these and for our joyful collaboration experience. I want to thank Achraf, Debabrota and Marc with whom I have much enjoyed wandering across NeurIPS halls and New Orleans at night. Cheers to that and to the interesting discussions we had at SCOOL on differential privacy and the many open questions related to it (How much (ϵ, δ)-budget have I spent so far?). Tomáš, it was a pleasure to discuss continuous BAI and other ambitious ideas with you. I hope they get to see the light one day!

Nicolas and Robert, I am very honored that you have accepted to review this manuscript and I warmly thank you for your careful reading. Anders, Bruno and Ciara, thank you for taking interest in this work and agreeing to serve on my thesis committee.

Je remercie les secrétaires Jessica, Magalie et Virginia pour leur bienveillance et leur serviabilité à toute épreuve.

Merci aux jeunes chercheurs de l'équipe d'Aurélien avec lesquels j'ai eu le plaisir de faire des Stats au tableau ou simplement de discuter de RL, bandits, graphes et autres:

Alexandre, Antoine, Élise, Hugues, Mehrasa, Pierre et Tomáš. Le "A" de l'UMPA et la joie de la recherche mathématique s'éteindraient sans vous.

Merci à tous les membres de l'UMPA avec qui j'ai partagé des moments agréables, que ce soit autour d'un déjeuner, une pause café, une partie de badminton ou encore la préparation d'un TD: Alexandre, Antoine, Basile, Céline, Charlie, Corentin, David, Denis, Élise, Héloïse, Hugues, Jules, Juliana, Léo, Micael, Mohamed, Grégory, Paul, Raphaël D, Raphaël R, Riccardo, Ronan, Matthieu, Thomas, Valentine, Vanessa, Vianney, et William.

Merci également aux membres de SCOOL qui m'ont toujours bien accueilli aussi bien dans leur séminaire que dans leurs parties de pétanque: Achraf, Adrienne, Alena, Deb, Dorian, Hector, Marc, Matheus, Odalric, Omar, Philippe, Reda, Rémy, Timothée et Tuan.

La thèse, c'est aussi quelques moments de galère et il faut une bonne compagnie pour y survivre.

A ce titre, je tiens à remercier le fidèle Arslan (I trust that your French is good enough to understand these words by now :)) ainsi que mes colocs du 22 Branly : Hakime, Jokim, Mélitine, Pierre-Etienne et Tarsila. La légende raconte qu'il suffisait d'un conte magique autour d'un (faux) feu de cheminée pour se rendre compte que "Tout est relatif !".

« قم للمعلم وفِّه التبجيلا، كاد المعلم أن يكون رسولا» . أود هنا أن أشكر معلمِيَّ و أساتذتي في كل مراحل الدراسة و خاصة مدرسي الرياضيات الذين حببوا إلي هذه المادة و شجعوني على التعمق فيها: المعلمة بوطوالة، الأستاذ السليمي، الأستاذ بن داوود، الأستاذ بناني، الأستاذ أسلالو و الأستاذ الطيبي.

A big salute to my friends scattered across the globe. Badr, Habib and Hamza: Cheers to the wildest trips (P.S: Making jokes while at border control was maybe not a good idea, but if you read this thesis you will see why we need more samples to gain certainty) and the most intellectually vibrant chat group when we're not in the same place. My visits to London, Paris and the US would have had no charm if not for the company of Ahmed-Taha, my brother Ali (who eats almost all the moroccan biscuits that Mom sends for both of us), Ayoub (the NY saint), Habib, Hakime, Ismail (the best London guide), Mohenned, Mossaab (the hospitable parisian who has been "visiting Lyon soon" for the past three years), Naoufal (the Côte-Azur connoisseur) and Oussama (the good old days roommate). During early Covid days, I had the chance to meet an exotic quartet of PhD students in Stockholm who encouraged me to pursue a PhD: Abderrahman, Anass, Othmane and Yassir. I want to thank you for your hospitality every time that I was passing by and to remind you to "Look at the Vikings' rock !" whenever you struggle to answer some difficult question. See you all some time very soon!

و في مِسك الختام، كل الشكر و العرفان لعائلتي الكبيرة و بالخصوص لأسرتي الصغيرة التي طالما ساندتني: أمي و أبي، أخي و أختي (المشاكسة)، جدتاي و جداي (رحمهما الله). إلى أمي و أبي: ما كنت لأبلغ هذه المرتبة دون دعمكما الدائم لي و حرصكما على تميزي في جميع المستويات الدراسية. أنا مدين لكما بكل شيء. عسى إهدائي هذه الرسالة لكما يرد و لو يسيراً من كل ذلك المعروف.

Aymen Al-Marjani,
Lyon, December 2023.

Contents

Résumé

Cette thèse s'intéresse aux problèmes d'exploration pure dans les Processus de Décision Markoviens (PDM) et les Bandits Multi-Bras. Ces problèmes ont surtout été étudiés dans une optique "pire-des-cas". L'objet de cette thèse est d'aller au-delà de ce cadre pessimiste en approfondissant notre compréhension de la complexité "spécifique à l'instance", c'est-à-dire du nombre d'observations dont un algorithme *adaptatif* aurait besoin pour accomplir une tâche d'exploration pure dans un PDM qui n'est pas nécessairement difficile.

Premièrement, nous étudions le problème d'identification de la meilleure politique (en anglais "Best Policy Identification" ou BPI) dans un PDM. En s'inspirant de travaux existants dans le cas particulier des bandits, nous démontrons une borne inférieure sur la complexité des algorithmes de BPI dans un PDM escompté. Ensuite nous proposons un algorithme inspiré par cette borne et qui explore les paires d'état-action du PDM proportionellement aux fréquences optimales dictées par la borne. Nous démontrons que cet algorithme est, à un facteur 2 près, asymptotiquement optimal.

Dans un deuxième temps, nous développons une approche d'exploration plus directe qui permet de collecter n'importe quel nombre souhaité d'observations depuis n'importe quelles paires d'état-action dans un PDM épisodique, tout en utilisant un nombre minimal d'épisodes. Nous verrons que pour un bon choix du nombre d'observations, une telle stratégie peut être employée pour résoudre le problème de BPI mais aussi celui de l'exploration sans récompense ("Reward-Free Exploration" en anglais). Ceci donne lieu à des algorithmes admettant des bornes plus fines sur leur complexité, qui dépendent notamment du PDM que l'on souhaite résoudre.

Finalement, à travers le problème d'identification de l'ensemble des bras ε-optimaux dans un bandit multi-bras, nous explorons une méthode alternative pour prouver des bornes inférieures dans les problèmes d'exploration pure. Nous illustrons certains cas où les bornes obtenues ainsi sont plus fines que celles prouvées via la méthode classique.

Mots Clés. Processus de Décision Markoviens · Identification de la meilleure politique · Exploration sans récompense · Apprentissage par Renforcement · Exploration pure

Abstract

This thesis studies pure exploration problems in Markov Decision Processes (MDP) and Multi-Armed Bandits. These problems have mainly been studied in a "worst-case" perspective. Our aim is to go beyond this pessimistic framework by deepening our understanding of the "problem-dependent" sample complexity, i.e., of the number of observations that an *adaptive* algorithm would need to accomplish a pure exploration task in an MDP that is not necessarily difficult.

First, we study the problem of "Best Policy Identification" (BPI) in a infinite-horizon discounted MDP. Drawing inspiration from existing work in the particular case of bandits, we derive a lower bound on the sample complexity of fixed-confidence BPI algorithms. Then we propose Navigate-and-Stop, an algorithm that explores the state-action pairs of the MDP proportionally to the optimal frequencies dictated by the bound. We prove that this algorithm is, within a factor of 2, asymptotically optimal.

In a second part, we develop a more direct exploration approach which allows to collect any desired number of observations from any state-action pairs in an episodic MDP, while using a minimal number of episodes. We will see that for a good choice of the number of observations, such a strategy can be used to solve the problem of BPI but also that of Reward-Free Exploration (RFE). This leads to algorithms that enjoy tighter bounds on their sample complexity, which depend in particular on the MDP that the algorithm is facing.

Finally, through the problem of All-ε-Best-Arms-Identification in a multi-armed bandit, we explore an alternative method to prove lower bounds on the sample complexity for pure exploration problems. Notably, we illustrate certain cases where the bounds obtained in this way are tighter than those proven via the classical method.

Keywords. Reinforcement Learning · Markov Decision Processes · Best Policy Identification · Reward-Free Exploration · Pure Exploration

List of Publications

List of publications in international conferences with proceedings

Below is the list of publications that I co-authored during my PhD.

- Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. **Navigating to the Best Policy in Markov Decision Processes**. In *Advances in Neural Information Processing Systems* (NeurIPS), 34, 2021.
- Aymen Al Marjani, Andrea Tirinzoni, and Emilie Kaufmann. **Active Coverage for PAC Reinforcement Learning**. In *Proceedings of the 36th Conference On Learning Theory* (COLT), 2023.
- Aymen Al Marjani, Tomáš Kocák, Aurélien Garivier. **On the Complexity of All ε-Best Arms Identification**. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (ECMLPKDD), 2022.
- Andrea Tirinzoni, Aymen Al Marjani, and Emilie Kaufmann. **Near Instance-Optimal PAC Reinforcement Learning for Deterministic MDPs**. In *Advances in Neural Information Processing Systems* (NeurIPS), 2022.
- Andrea Tirinzoni, Aymen Al Marjani, and Emilie Kaufmann. **Optimistic PAC Reinforcement Learning: the Instance-Dependent View**. In *Algorithmic Learning Theory* (ALT), 2023.
- Achraf Azize, Marc Jourdan, Aymen Al Marjani and Debabrota Basu. **On the Complexity of Differentially Private Best-Arm Identification with Fixed Confidence**. In *European Workshop on Reinforcement Learning* (EWRL), 2023.

This manuscript is based upon the first four articles in the list above. Chapter 1 introduces the settings that we study in this thesis and presents a few selected contributions, including one from the NeurIPS 2022 paper. Chapter 2 is based upon the NeurIPS 2021 paper and studies exact Best Policy Identification in the online setting, thereby generalizing results we had for the generative-model setting in the ICML 2021 paper. Chapters 3 and 4 are both extracted from the COLT 2023 paper. The former investigates the problem of active coverage with application to Reward-Free Exploration in episodic MDPs, while the latter explores the problem of ε-Best Policy Identification in the same setting. Finally, Chapter 5 presents some results for the problem of All-ε Best Arms Identification in multi-armed bandits, which come from the ECMLPKDD 2022 paper.

1. Introduction

Contents

1.1 Preamble

"*Human life is one long decision tree.*" (Sterelny, 2007). Based on the previous truism, one could argue that Humankind's eternal tragedy lies in the fact that we are forever doomed to learn the best decisions again, in nodes of this tree that our ancestors already encountered. From walking to writing a thesis on to playing chess, there are numerous tasks that no amount of transmitted knowledge alone can help us master. Instead, we must learn those through practice, by *trial-and-error*.

Reinforcement Learning (RL) (Sutton & Barto, 2018) offers a paradigm for learning a task by framing it as a *sequence of state-dependent decisions* that maximizes some notion of long-term utility. For the toddler learning to walk, RL reduces the task to answering the question: "Which muscle should I move next and in which direction when my body is in its current position?". The utility would then be maintaining equilibrium through several steps. Alternatively, for the Ph.D. student trying to write a thesis, the question is rather: "Which idea should I present next and to which level of detail must I do so, given the current state of my manuscript?". Here, a possible definition of the utility would be receiving the least amount of corrections to make from your supervisors.

The underlying mathematical model for studying RL is the framework of *Markov Decision Processes* (MDPs) (Puterman, 1994). Formally, an *agent* interacts sequentially with some *unknown environment* starting from an initial *state*. At every time step, the agent must select an *action* to play among a set of available actions. She then receives a *reward* and the *new state* of the environment, both of which depend on the previous state and the action that she played. Agents can act in various ways, characterized by their *policy*, i.e. the function that determines which action the agent would play given her current state and the information that she collected from the past rounds of interaction. The quality of an agent's policy is measured through the expected sum of rewards received across a given *time horizon*, which can eventually be infinite. For example, the toddler would obtain a reward for each step made without falling and the horizon would be infinite. In the chess apprentice's case, a reward could be given when she wins the game, but also when she captures an adversary's piece or traps her in a fork situation. The horizon can then be set as the largest number of moves ever recorded in a chess game.

1.2 (The Need for) Pure Exploration in RL

RL algorithms start with zero knowledge of the environment and aim to learn a near-optimal policy through repeated interactions with it. A central question that arises then is:

How to evaluate the learning trajectory of a given algorithm?

The most common approach in theoretical RL literature is to compare the rewards gathered by the algorithm with those of a mighty agent who knew an optimal policy from the very beginning. This leads to two distinct but similar performance criteria: *regret minimization* (Lai & Robbins, 1985) and *PAC-MDP* (Kakade, 2003). As its name indicates, regret minimization penalizes the amount of mistakes, represented by the rewards that the algorithm missed during the learning process when it acts in a sub-optimal way. On the other hand, the PAC-MDP criterion counts the number of time steps where the algorithm plays according to an ε-sub-optimal policy, for some $\varepsilon > 0$. Both regret minimization and PAC-MDP algorithms face an *exploration-exploitation dilemma*. Indeed, they must balance the need to explore the environment to learn more about how it behaves with the need to exploit the knowledge gained so far and act following the policy that appears to be optimal given this knowledge.

However, the exploration-exploitation paradigm does not capture all the possible situations that one might encounter in the real world. Indeed, in many applications, we are interested in learning some property of the unknown environment *using the least amount of interactions* with it, *regardless of the rewards missed while learning.* We refer to this learning framework as *pure exploration.* The following examples illustrate some use cases where we might instead want to perform pure exploration:

- **A/B tests:** Consider an E-business company running an A/B test experiment in order to decide which among several possible versions of their website generates more revenue or increases their user-engagement metrics. A/B tests are often modeled through the framework of *sequential hypothesis tests*: The practitioner performs a statistical test where they split the incoming traffic between a *control* version and one or more *treatment* versions. Whenever the p-values of the collected data are conclusive about the identity of the best version, the practitioner may decide to stop the experiment (Johari et al., 2022). In addition, as explained in (Kohavi et al., 2013), A/B test practitioners also need to be able to quickly detect whether some treatment version performs very badly and abort the experiment. If they fail to do so, their website might witness "user abandonment", i.e., frustrated users will lose interest and never return back, and the company will incur costs in millions of dollars. This "early stopping" component is typical in pure exploration algorithms, where a *stopping rule* decides whether we have collected enough evidence to cease exploration and return a good answer, see Section 1.4.4. In contrast, algorithms for regret minimization or PAC-MDP either (i) assume a fixed time-bugdet for the experiment and quantify the losses made by the algorithm during that time period or (ii) prove theoretical upper bounds on the number of mistakes made by the algorithm but without providing a method to know *when the policies played by the algorithm have become good enough.* A/B tests pose an exploration challenge as well since practitioners continuously monitor the experiment data to adjust the proportion of traffic allocated to each version (Johari et al., 2022; Russac et al., 2021)
- **Iterative environment design:** A crucial problem for economists is that of market design, i.e. which rules and incentives should we implement in order to get a certain desired behavior by economic agents (companies and households). For instance, what is the best way to reduce airlines carbon emissions? Is it through a direct carbon tax? If so, should we tax flight tickets or the plane constructors? Or should we perhaps subsidize other means of transportation to shift the collective behavior of consumers? Recent works use an RL approach to answer this question (Zheng et al., 2020; Johanson et al., 2022). Imagine that you are tasked with building a simulator where a legislator could input their guess for an adequate reward function and get to observe how the market would evolve in such conditions. The agents within the simulator may need to relearn a near-optimal policy several times, as many as it takes for the legislator to ensure that the proposed reward will induce the desired behavior. Now since the reward that we seek to maximize at each round is only temporary, the mistakes made by the RL agent while learning are not relevant *per se.* What matters most to the simulator's user is the ability to identify the optimal policies for a given reward as fast as possible.
  A similar problem also arises in video game design, where designers need to ensure that pathological strategies, for instance running straight to the opponent's goalkeeper in a soccer game, can not win. Here it is rather the environment's dynamics (i.e. what is the game's next state when the player chooses to play a certain action in the current situation) that need to be tuned carefully to deliver a good gameplay experience. For a given choice of game dynamics and a set of undesirable policies,

we need a RL agent that can check with high confidence whether some undesirable policy is near-optimal. Again, the emphasis in this case is on the speed by which one can implement this iterative design strategy. In other words, we want to minimize the number of games that the RL agent needs to complete this task with enough certainty about its final answer. Notably, the losses incurred by our RL agent in its training phase hold no particular interest to the game designer. All of these are *pure exploration* problems, where we want to gather some information about the environment with high confidence.

1.3 Markov Decision Processes

In order to formalize the different pure exploration problems studied in this thesis, we first need to define Markov Decision Processes (MDPs) and recall some classical results in Dynamic Programming (Puterman, 1994).

1.3.1 Infinite horizon MDPs with discount

To define an MDP, one needs a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a collection of *reward distributions* $q : \mathcal{S} \times \mathcal{A} \to \mathcal{P}([0,1])$ and a Markov *transition kernel* $p : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$, where $\mathcal{P}(\mathcal{X})$ denotes the set of probability distributions over the set $\mathcal{X}$. An infinite horizon MDP is then defined as the tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, p, q)$. The interaction of an agent with $\mathcal{M}$ takes place at discrete time steps, $t \in \mathbb{N}_{\geq 1}$. First, an initial state of the environment $s_1 \in \mathcal{S}$ is drawn from some initial distribution $\mu \in \mathcal{P}(\mathcal{S})$. At every time step t, the agent observes the state of the environment s_t and plays an action $a_t \in \mathcal{A}$ of her choice. She then observes a reward $R_t(s_t, a_t)$ and the new state of the environment s_{t+1} respectively drawn from $q(.|s_t, a_t)$ and $p(.|s_t, a_t)$. This interaction carries on indefinitely, yielding a trajectory $\big(s_t, a_t, R_t(s_t, a_t)\big)_{t\geq 1}$. Except when we derive information-theoretic lower bounds, we will most often forget about the reward distribution and use only its mean $r(s,a) := \mathbb{E}_{q(.|s,a)}[R(s,a)]$. The mapping $(s,a) \mapsto r(s,a)$ is called the *reward function*.

Assumption 1.1 Throughout this thesis, we always assume that $\mathcal{S}$ and $\mathcal{A}$ are finite. We use S and A to denote their respective cardinals. We say then that the MDP is *tabular*.

We define the history of observations up to time t as $\mathcal{H}_t := (s_1, a_1, R_1, \ldots, R_{t-1}, s_t)$ and denote by $\mathcal{B}_t$ the set of all possible histories at that time. Further, we let $\mathcal{B} := \cup_{t\geq 1}\mathcal{B}_t$ be the set of all possible histories. Then we can characterize any agent by her policy, denoted by π, which is a mapping from $\mathcal{B}$ to $\mathcal{P}(\mathcal{A})$ that determines which action the agent will play based on the history of her previous observations. Two special classes of policies are the set of stationary Markovian policies $\Pi_{\mathrm{S}} = \{\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})\}$[1] and stationary deterministic Markovian policies $\Pi_{\mathrm{D}} = \{\pi : \mathcal{S} \to \mathcal{A}\}$. Markovian policies only look at the current state to compute a distribution over actions, from which the next action will be sampled. Deterministic Markovian policies have the additional property that they output a single action instead of a distribution.

The RL objective As explained before, RL tries to solve tasks by reducing them to finding policies that maximize some long-term utility. In the case of infinite horizon discounted MDPs, the utility of a policy π is defined as

$$\mathcal{U}^\pi := \sum_{t=1}^{\infty} \gamma^{t-1} R_t(s_t, a_t) \ ,$$

where $\gamma \in [0,1)$ is a pre-specified *discount factor*. Larger values of γ indicate that we value future rewards (almost) as much as the immediate reward of step $t = 1$, while smaller values

[1] The $_S$ in Π_{S} stands for "stochastic".

indicate a strong preference for the present. Note that $\mathcal{U}^\pi$ is a random variable whose value depends on the stochastic trajectory. Therefore, to optimize the utility one needs a metric that summarizes its distribution when the agent executes some policy π. In this thesis, we focus on the classical RL setting where the objective is set as the expectation of the utility,

$$\begin{aligned} V^\pi(s) &:= \mathbb{E}_{\mathcal{M},\pi}\left[\sum_{t=1}^{\infty}\gamma^{t-1}R_t(s_t,a_t)\middle|\ s_1=s\right], \\ &= \mathbb{E}_{\mathcal{M},\pi}\left[\sum_{t=1}^{\infty}\gamma^{t-1}r(s_t,a_t)\middle|\ s_1=s\right]. \end{aligned} \tag{1.1}$$

The expectation above is taken over the randomness of the trajectory[2] $(s_t,a_t)_{t\geq 1}$ that results from the interaction of the policy π with the MDP $\mathcal{M}$, i.e, when $a_t\sim\pi(s_t)$ and $s_{t+1}\sim p(.|s_t,a_t)$ for all $t\geq 1$. The mapping $s\mapsto V^\pi(s)$ is called the *value function* of policy π. A policy $\pi^\star$ is said to be optimal if it maximizes the value function at every state

$$\forall s\in\mathcal{S},\quad V^{\pi^\star}(s)=\max_{\pi:\mathcal{B}\to\mathcal{P}(\mathcal{A})}V^\pi(s) \tag{1.2}$$

Theorem 5.5.3 in (Puterman, 1994) proves that for every history-dependent policy π there exists a Markovian policy $\widetilde{\pi}$ such that $V^\pi(s)=V^{\widetilde{\pi}}(s)$ for all states s. Therefore, if there exists an optimal policy in $\mathcal{M}$, it is sufficient to search for it among Markovian policies.

Bellman equations and value gaps A central object in the analysis of policy values is the *action-value function*

$$Q^\pi(s,a) := \mathbb{E}_{\mathcal{M},\pi}\left[\sum_{t=1}^{\infty}\gamma^{t-1}R_t(s_t,a_t)\middle|\ s_1=s,a_1=a\right], \tag{1.3}$$

which quantifies, for every state-action pair (s,a) and policy π, the total reward that the agent would receive when she starts in state s, plays action a, then commits to playing actions $a_t\sim\pi(.|\mathcal{H}_t)$ in later time steps $t\geq 2$. The optimal action-value function is simply defined as

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A},\quad Q^\star(s,a):=\max_{\pi:\mathcal{B}\to\mathcal{P}(\mathcal{A})}Q^\pi(s,a). \tag{1.4}$$

Both the action-value function Q^π of a deterministic Markovian policy π and the optimal action-value function $Q^\star$ satisfy the *Bellman equations*

$$Q^\pi(s,a)=r(s,a)+\gamma\sum_{s'\in\mathcal{S}}p(s'|s,a)Q^\pi(s',\pi(s')), \tag{1.5}$$

$$Q^\star(s,a)=r(s,a)+\gamma\sum_{s'\in\mathcal{S}}p(s'|s,a)\max_{a'\in\mathcal{A}}Q^\star(s',a'), \tag{1.6}$$

for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. One way to measure how suboptimal it is to play action a at state s is through its value gap

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A},\quad \Delta(s,a):=V^\star(s)-Q^\star(s,a). \tag{1.7}$$

Indeed, actions with $\Delta(s,a)=0$ are optimal, while a large value gap indicates that the agent loses considerable reward when she plays action a at state s even if she commits to executing an optimal policy later.

[2] For a rigorous construction of the underlying probability space and stochastic process, see Chapter 2 in (Puterman, 1994).

1.3.2 Non-stationary finite-horizon MDPs

In this thesis, we also study pure exploration within the framework of *episodic MDPs*. Similar to its infinite-horizon counterpart, a non-stationary episodic MDP is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h\in[H-1]}, \{q_h\}_{h\in[H]}, s_1)$. Here H denotes the *horizon* while the *transition kernel* p_h and the *reward distributions* q_h may now also depend on the *step* $h \in [H]$. The interaction of an RL agent with a finite-horizon MDP is structured through *episodes* $t \in \mathbb{N}_{\geq 1}$, where each episode consists of H steps. At the beginning of every episode t, the environment is at the initial state $s_1^t := s_1$. At each step $h \in [H-1]$, the agent observes the current state s_h^t and plays an action a_h^t. She then observes an immediate reward $R_h^t(s_h^t, a_h^t)$ and a next state s_{h+1}^t respectively drawn from $q_h(.|s_h^t, a_h^t)$ and $p_h(.|s_h^t, a_h^t)$. In the last step $h = H$, after playing an action a_H^t, the agent only observes a reward $R_H^t(s_H^t, a_H^t)$ before the current episode terminates and a new one begins. As with discounted MDPs, we will often make use of the *reward function* $r : (h, s, a) \mapsto r_h(s, a)$ where $r_h(s, a) := \mathbb{E}_{q_h(.|s,a)}[R_h(s, a)]$.

Remark 1.1 The assumption of a fixed initial state s_1 is without loss of generality. Indeed, suppose that the initial state of $\mathcal{M}$ was drawn from some distribution $\mu \in \mathcal{P}(\mathcal{S})$. Then any RL problem on $\mathcal{M}$ can be solved on an "augmented" MDP $\mathcal{M}'$ where we add a step $h = 0$ and a fictional initial state s_0 such that[a]

$$\forall s_1 \in \mathcal{S},\ \forall a \in \mathcal{A},\ p_0'(s_1|s_0, a) = \mu(s_1) \text{ and } q(.|s_0, a) = \delta_0$$

We leave the transition kernels and reward distributions at steps $h \geq 1$ unchanged. The new MDP $\mathcal{M}'$ now has a fixed intial state and the total reward collected by any policy is the same in $\mathcal{M}$ and $\mathcal{M}'$. Only the horizon has changed, as $H' = H + 1$. ■

[a] δ_0 denotes the dirac distribution located at 0

The history of past observations is now defined for every episode as $\mathcal{H}_1 = (s_1)$ and $\mathcal{H}_t = (s_1^1, a_1^1, R_1^1, \ldots, s_H^{t-1}, R_H^{t-1})$ for $t \geq 2$. Similar to the infinite-horizon case, we let $\mathcal{B}_t$ and $\mathcal{B} := \cup_{t\geq 1}\mathcal{B}_t$ respectively denote the set of possible histories at the beginning of episode t and the set of all possible histories. Markovian policies become mappings from a state-step pair to action distributions $\Pi_S := \{\pi : [H] \times \mathcal{S} \to \mathcal{P}(\mathcal{A})\}$. For a Markovian policy π, we denote by $\pi_h(a|s)$ the probability that an agent executing π plays action a when the environment state is s at step h. Finally, a Markovian policy is deterministic if outputs a Dirac distribution over a single action. We let $\Pi_D = \{\pi : [H] \times \mathcal{S} \to \mathcal{A}\}$ denote the set of deterministic Markovian policies.

The RL objective In the episodic setting, the goal is to maximize the expected sum of rewards over an episode[3]

$$V_1^\pi := \mathbb{E}_{\mathcal{M},\pi}\left[\sum_{h=1}^{H} R_h(s_h, a_h) \,\middle|\, s_1\right]. \tag{1.8}$$

[3] When it is clear from context, we drop the dependence on the initial state from the value function since the latter is fixed.

V_1^π is called the *value function at the root*. For analysis purposes, it is convenient to define the step-wise value function and the step-wise action-value function

$$\forall (h,s) \in [H] \times \mathcal{S}, \quad V_h^\pi(s) := \mathbb{E}_{\mathcal{M},\pi}\left[\sum_{\ell=h}^{H} R_\ell(s_\ell, a_\ell) \,\middle|\, s_h = s\right], \\ = \mathbb{E}_{\mathcal{M},\pi}\left[\sum_{\ell=h}^{H} r_\ell(s_\ell, a_\ell) \,\middle|\, s_h = s\right]. \tag{1.9}$$

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}, \quad Q_h^\pi(s,a) := \mathbb{E}_{\mathcal{M},\pi}\left[\sum_{\ell=h}^{H} R_\ell(s_\ell, a_\ell) \,\middle|\, s_h = s, a_h = a\right], \\ = \mathbb{E}_{\mathcal{M},\pi}\left[\sum_{\ell=h}^{H} r_\ell(s_\ell, a_\ell) \,\middle|\, s_h = s, a_h = a\right]. \tag{1.10}$$

A policy $\pi^\star$ is *Bellman optimal* if

$$\forall (h,s) \in [H] \times \mathcal{S}, \quad V_h^{\pi^\star}(s) = \max_{\pi : \mathcal{B} \to \mathcal{P}(\mathcal{A})} V_h^\pi(s). \tag{1.11}$$

In this thesis, we will mainly investigate a more relaxed notion of optimality. Namely, we say that a policy $\pi^\star$ is optimal if

$$V_1^{\pi^\star} = \max_{\pi : \mathcal{B} \to \mathcal{P}(\mathcal{A})} V_1^\pi. \tag{1.12}$$

In other words, a policy is optimal if it yields the best value at the initial state s_1.

Backward Induction For episodic MDPs, Proposition 4.4.3 together with Theorem 4.5.1 in (Puterman, 1994) guarantee that

- There always exists a deterministic Markovian policy $\pi^\star \in \Pi_{\mathrm{D}}$ that is optimal,
- $\pi^\star$ can be computed by the *backward induction* algorithm, also referred to as *dynamic programming*. Its pseudo-code is presented below.

Algorithm 1 Backward Induction

1: **Input:** Transition kernel p, reward function r.
2: Initialize optimal action-value function $Q_{H+1}^\star(s,a) \leftarrow 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$
3: Initialize "artificial" transitions $p_H(s'|s,a) \leftarrow \mathbb{1}(s' = s)$ for all (s,a)
4: **for** $h = H, H-1, \dots, 1$ **do**
5: Compute action-value of step h:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}, \; Q_h^\star(s,a) \leftarrow r_h(s,a) + \sum_{s' \in \mathcal{S}} p_h(s'|s,a) \max_{a \in \mathcal{A}} Q_{h+1}^\star(s',a') \tag{1.13}$$

6: Compute optimal policy at step h:

$$\forall s \in \mathcal{S}, \; \pi_h^\star(s) \leftarrow \arg\max_{a \in \mathcal{A}} Q_h^\star(s,a) \tag{1.14}$$

7: **end for**

1.3.3 Finite-armed bandits

In Chapter 5, we will study a particular case of tabular MDPs, namely the *multi-armed bandit* (MAB) model (Lattimore & Szepesvari, 2019). A finite MAB is defined by collection of reward distributions $\nu := (\nu_a)_{a \in [K]}$ called *arms*, where $K \in \mathbb{N}_{\geq 1}$. The agent interacts

with the bandit at discrete time steps $t \in \mathbb{N}_{\geq 1}$. At every time step t, the agent *pulls* an arm $A_t \in [K]$ and observes a reward $R_t \sim \nu_{A_t}$. The samples from different arms at different time steps are independent. In other words, for any sequence of time steps $(t_1, \dots, t_N)$ and sequence of actions $(a_1, \dots, a_N)$, the reward vector $(R_{t_1}, \dots, R_{t_N})$ is a sample from $\nu_{a_1} \otimes \dots \otimes \nu_{a_N}$ conditionally on the event $(A_{t_1} = a_1, \dots, A_{t_N} = a_N)$. At any time step $t \geq 1$, the history of observation is defined by $\mathcal{H}_t := (A_u, R_u)_{1 \leq u \leq t}$.

We will be interested in the mean-rewards of the arms, denoted by $(\mu_a)_{a \in [K]}$. We denote by $a^\star \in \arg\max_{a \in [K]} \mu_a$ an arm with the largest mean, with ties broken arbitrarily. Finally, $\mu^\star := \mu_{a^\star}$.

Remark 1.2 The finite-armed bandit is a special case of tabular episodic MDPs, where $S = 1, H = 1, A = K$ and $q(.|s_1, a) := \nu_a$. ■

1.3.4 Sampling models and sampling rules

There exist mainly two sampling models that define how RL algorithms can collect observations in some MDP $\mathcal{M}$.

Online model In general, RL algorithms must interact with $\mathcal{M}$ according to the same protocol that deployed RL agents follow, see Sections 1.3.1 and 1.3.2. Given a history of observations $\mathcal{H}_t$, the sampling rule of an algorithm $\mathbb{A}$ determines which policy $\mathbb{A}$ will execute in the next step to explore $\mathcal{M}$. Formally, in the discounted setting the sampling rule is the mapping

$$\begin{aligned} \text{SMP}: \quad & \mathcal{B} \to \Pi_{\text{S}} \\ & (s_1, a_1, R_1, \dots, R_t, s_{t+1}) \mapsto \pi^{t+1} \end{aligned} \tag{1.15}$$

where π^{t+1} is the policy used to select an action in the $(t+1)$-th time step. Similarly, in the episodic framework, it is defined as

$$\begin{aligned} \text{SMP}: \quad & \mathcal{B} \to \Pi_{\text{S}} \\ & \left(\left(s_h^e, a_h^e, R_h^e \right)_{h \in [H]} \right)_{1 \leq e \leq t} \mapsto \pi^{t+1} \end{aligned} \tag{1.16}$$

where π^{t+1} is the policy executed by $\mathbb{A}$ in episode $t+1$.

Generative model In some cases, the algorithms that we design to learn good policies may have more degrees of freedom in the training phase than when they are finally deployed. For instance, we might have access to a simulator, often called a *generative model*, that enables us to query observations from any state-action pair (s, a) even if s is not the current state of the environment (Chapter 2 in Kakade, 2003). More precisely, we think of a generative model as a random sampler that takes as input a pair (s, a) and returns an independent sample $(R, s') \sim q(.|s, a) \otimes p(.|s, a)$. We denote by

$$(R, s') \leftarrow \text{GenerativeModel}(s, a) \tag{1.17}$$

the act of sampling a reward and a transition from the state-action pair (s, a) using the generative model. Under this model, the sampling rule of an algorithm is a mapping from histories to distributions over states and actions, that determines which state-action pair we will query next

$$\begin{aligned} \text{SMP}: \quad & \mathcal{B} \to \mathcal{P}(\mathcal{S} \times \mathcal{A}) \\ & (s_u, a_u, R_u)_{1 \leq u \leq t} \mapsto (s_{t+1}, a_{t+1}) \end{aligned} \tag{1.18}$$

1.4 Pure Exploration Problems

Now we present the pure exploration problems studied in this thesis.

1.4.1 Best Policy Identification

Exact Best Policy Identification A special problem of particular interest is that of exact Best Policy Identification (BPI), where we want to find an optimal policy as fast as possible. Assuming that there is a unique optimal policy $\pi^\star$, we wish to design an algorithm that will interact with $\mathcal{M}$ until it has gathered enough observations to return an estimate $\widehat{\pi}$ that is *certified to be correct with high probability*. We investigate this problem in the setting of discounted MDPs.

Assumption 1.2 Let $\mathfrak{M}_{\star,1}$ be the class of infinite-horizon discounted MDPs with a unique optimal policy. We assume that $\mathcal{M} \in \mathfrak{M}_{\star,1}$ and we denote its optimal policy by $\pi^\star(\mathcal{M})$. In other words,

$$\left\{ \pi \in \Pi_{\mathrm{S}} : \forall s \in \mathcal{S},\ V^\pi(s;\mathcal{M}) = \max_{\pi' \in \Pi_{\mathrm{S}}} V^{\pi'}(s;\mathcal{M}) \right\} = \{\pi^\star(\mathcal{M})\ \}, \tag{1.19}$$

where we indexed the value functions by $\mathcal{M}$ to emphasize their dependency on the MDP.

Definition 1.1 An algorithm $\mathbb{A}$ for exact Best Policy Identification interacts with the MDP for a possibly random number of steps τ and returns an estimate of the best policy $\widehat{\pi} \in \Pi_{\mathrm{D}}$. Given a risk $\delta \in (0,1)$, we say that $\mathbb{A}$ is δ-PAC (or δ-correct) for BPI on the class $\mathfrak{M}_{\star,1}$ if

$$\forall \mathcal{M} \in \mathfrak{M}_{\star,1}, \quad \mathbb{P}_{\mathcal{M},\mathbb{A}}\big(\tau < +\infty, \widehat{\pi} = \pi^\star(\mathcal{M})\big) \geq 1-\delta, \tag{1.20}$$

where $\mathbb{P}_{\mathcal{M},\mathbb{A}}$ denotes the distribution of observations when $\mathbb{A}$ interacts with $\mathcal{M}$.

ε-Best Policy Identification (ε-BPI) In the ε-BPI problem we require the algorithm to find a policy whose value is, with high probability, within a range of ε from the optimal value. We will present results for approximate BPI in the setting of episodic MDPs.

Definition 1.2 An algorithm $\mathbb{A}$ for ε-Best Policy Identification interacts with the MDP for a possibly random number of steps τ and returns an estimated policy $\widehat{\pi} \in \Pi_{\mathrm{D}}$. Given a precision $\varepsilon \geq 0$ and a risk $\delta \in (0,1)$, we say that $\mathbb{A}$ is (ε,δ)-PAC (or (ε,δ)-correct) for ε-BPI on some class of MDPs $\mathfrak{M}$ if

$$\forall \mathcal{M} \in \mathfrak{M}, \quad \mathbb{P}_{\mathcal{M},\mathbb{A}}\big(\tau < +\infty,\ V_{1,\mathcal{M}}^{\widehat{\pi}} \geq V_{1,\mathcal{M}}^{\star} - \varepsilon\big) \geq 1-\delta. \tag{1.21}$$

1.4.2 Reward-Free Exploration (RFE)

Imagine that you have access to some dynamical system, represented by a transition kernel $p : [H] \times \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$, where you can play actions and observe the (possibly stochastic) evolution of the system's state following each action. You are tasked with learning the dynamics and delivering an estimate $\widehat{p}$. This estimate will then be used by a planning agent to maximize some utility given by their own mean-reward function, which is yet undisclosed to you at present. Naturally, the planner expects your estimate to be sufficiently accurate so that they never lose more than ε in value when they plan using $\widehat{p}$ instead of the correct model p. How would you explore this system and how would you decide if you have gathered enough data to satisfy the previous requirement? This is the topic of reward-free exploration, which we will study in the setting of episodic MDPs.

We denote by $\widehat{\pi}_r$ an optimal policy in the MDP whose transition kernel is $\widehat{p}$ and the

mean-reward function is r. We also let $V_1^{\pi}(s_1; r)$ be the value function of policy π under the true transition model p when the mean reward function is r.

Definition 1.3 An algorithm $\mathbb{A}$ for reward-free exploration interacts with a dynamical system $(\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h\in[H]}, s_1)$ for a possibly random number of steps τ and returns an estimate of the transition kernel $\widehat{p}$. Given a precision $\varepsilon > 0$ and a risk $\delta \in (0,1)$, we say that $\mathbb{A}$ is (ε, δ)-PAC (or (ε, δ)-correct) for RFE on some class of transitions kernels $\mathfrak{P}$ if for all $p \in \mathfrak{P}$,

$$\mathbb{P}_{p,\mathbb{A}}\big(\tau < +\infty,\ \forall r \in [0,1]^{SAH} : V_1^{\widehat{\pi}_r}(s_1; r) \geq V_1^{\star}(s_1; r) - \varepsilon\big) \geq 1 - \delta, \tag{1.22}$$

where $\mathbb{P}_{p,\mathbb{A}}$ denotes the distribution of observations when $\mathbb{A}$ interacts with p. In this case, we say that $\widehat{p}$ *is an* ε*-good transition kernel.*

1.4.3 All ε-best arms Identification (All-ε-BAI)

Last but not least, we will also investigate the problem of All ε-best arms Identification (All-ε-BAI) in a multi-armed bandit. We consider Gaussian MABs with unit variance of the form $\nu = \big(\mathcal{N}(\mu_a, 1)\big)_{a\in[K]}$, where $\mathcal{N}(\theta, \sigma^2)$ denotes the Gaussian distribution of mean θ and standard-deviation σ. For simplicity, we abuse notation and refer to a MAB ν by its vector of mean-rewards $\mu = (\mu_a)_{a\in[K]}$. The goal in All-ε-BAI is to identify the set of "good" arms $G_\varepsilon(\mu) := \{a \in [K] :\ \mu_a > \mu^\star - \varepsilon\}$ with high probability, where $\varepsilon > 0$ is a pre-determined precision parameter and $\mu^\star := \max_{b\in[K]} \mu_b$.

Definition 1.4 An algorithm $\mathbb{A}$ for All ε-Best Arms Identification interacts with a MAB μ for a possibly random number of steps τ and returns an estimated set $\widehat{G}$. Given a risk $\delta \in (0,1)$ and a precision $\varepsilon > 0$, we say that $\mathbb{A}$ is (ε, δ)-PAC (or (ε, δ)-correct) for All-ε-BAI if

$$\forall \mu \in \mathbb{R}^K,\ \mathbb{P}_{\mathbb{A},\mu}\big(\tau < +\infty,\ \widehat{G} = G_\varepsilon(\mu)\big) \geq 1 - \delta. \tag{1.23}$$

Assumption 1.3 For the All-ε-BAI problem to be solvable with finite sample complexity, we assume that there is no arm a such that $\mu_a = \mu^\star - \varepsilon$.

All-ε-BAI was initially proposed by (Mason et al., 2020) as an alternative objective to two other pure-exploration problems in the multi-armed bandit literature, namely the TOP$-k$ arms selection (Kalyanakrishnan & Stone, 2010) and the THRESHOLD bandits (Locatelli et al., 2016). The former aims to find the k arms with the highest means, while the latter seeks to identify all arms with means larger than a given threshold s. As argued by (Mason et al., 2020), finding all the ε-optimal arms is a more robust objective than the TOP-K and THRESHOLD problems, which require some prior knowledge of the distributions in order to return a relevant set of solutions. Take for example drug discovery applications, where the goal is to perform an initial selection of potential drugs through *in vitro* essays before conducting more expensive clinical trials: setting the number of arms k too high or the threshold s too low may result into poorly performing solutions. Conversely, if we set k to a small number or the threshold s too high we might miss promising drugs that will prove to be more efficient under careful examination. The All-ε-BAI objective circumvents these issues by requiring to return all the drugs whose efficiency lies within a certain range from the best.

1.4.4 General structure of pure exploration algorithms

Besides the sampling rule described in Section 1.3.4, pure exploration algorithms have two additional components:

- **Stopping rule:** The stopping rule determines when an algorithm has gathered enough observations to return a good answer, either an (ε-)optimal policy for BPI or an ε-good transition kernel for RFE, with the desired level of confidence. Concretely, the stopping rule is a sequence of random variables, denoted $(\mathrm{STP}_t)_{t\geq 1}$, with values in the set $\{\mathrm{True}, \mathrm{False}\}$. This sequence is measurable with respect to the filtration generated by the sigma algebras of histories $(\sigma(\mathcal{H}_t))_{t\geq 1}$.
- **Recommendation rule:** The recommendation rule determines the final answer of the algorithm. It is a sequence of random variables, denoted $(\mathrm{REC}_t)_{t\geq 1}$, measurable with respect to the filtration generated by the sigma algebras of histories $(\sigma(\mathcal{H}_t))_{t\geq 1}$. For BPI (resp. RFE), REC_t takes values in the set of deterministic Markovian policies Π_{D} (resp. the set of all probability kernels $\mathcal{P}(\mathcal{S})^{SAH}$). For All-$\varepsilon$-BAI it is with values in $2^{[K]}$, the power set of $[K]$.

Below are general templates for pure exploration algorithms in an episodic MDP in the online setting and in a multi-armed bandit.[4]

Algorithm 2 Pure exploration protocol in episodic MDPs

Input: precision ε, risk $\delta \in (0,1)$.
Initialize history $\mathcal{H}_0 \leftarrow (s_1)$
for $t = 1, 2, \ldots$ **do**
  $\pi^t \leftarrow$ **SMP**$(\mathcal{H}_{t-1})$ // SAMPLING RULE
  for $h = 1, 2, \ldots, H$ **do**
    Play $a^t \sim \pi_h^t(s_h^t)$ and observe reward R_h^t (only for BPI) and next state s_{h+1}^t
  end for
  Update history with the last trajectory $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{(s_h^t, a_h^t, R_h^t)\}_{1\leq h\leq H}$.
  if **STP**$_t$**:** // STOPPING RULE
    Stop and return **REC**$_t$ // RECOMMENDATION RULE
  end if
end for

Algorithm 3 Pure exploration protocol in bandits

Input: precision ε, risk $\delta \in (0,1)$.
Initialize history $\mathcal{H}_0 \leftarrow ()$
for $t = 1, 2, \ldots$ **do**
  $a_t \leftarrow$ **SMP**$(\mathcal{H}_{t-1})$ // SAMPLING RULE
  Play a^t and observe reward R_t
  Update history with the last observation $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{(a_t, R_t)\}$.
  if **STP**$_t$**:** // STOPPING RULE
    Stop and return **REC**$_t$ // RECOMMENDATION RULE
  end if
end for

1.5 The Sample Complexity of Pure Exploration

Notation For a pair of functions with real values f and g, we shall use $f = \mathcal{O}(g)$ (resp. $f = \Omega(g)$) if there exists a universal constant $c > 0$ such that $f \leq c \cdot g$ (resp. $f \geq c \cdot g$).

[4] By convention, we set $s_{H+1}^t := s_1$.

The expression $\text{poly}(X_1, \dots, X_m)$ will refer to any polynomial function in the variales $X_1, \dots, X_m$. We write $f = \widetilde{\mathcal{O}}(g)$ if $f \leq \text{poly}(\log(Y_1), \dots, \log(Y_m)) \cdot g$, where $Y_1, \dots, Y_m$ are other parameters that we shall specify after every $\widetilde{\mathcal{O}}$ statement. We denote the infinity norm of f by $\|f\|_\infty := \sup_{x \in \text{Dom}(f)} |f(x)|$ where $\text{Dom}(f)$ is the domain of f.

Performance criteria When we design a pure exploration algorithm, we try to achieve *sample efficiency* by minimizing the number of observations needed from the environment to complete the pure exploration task. Concretely, we seek to minimize the stopping time of our algorithm,

$$\tau := \inf \big\{ t \geq 1 : \ \text{STP}_t = \text{True} \big\}. \tag{1.24}$$

We note that τ is a random variable whose value depends on the stochastic process resulting from the interaction of an algorithm $\mathbb{A}$ with the MDP $\mathcal{M}$. Therefore, bounds on τ in the literature feature either an expectation or a $(1-\delta)$-quantile, both of which we refer to as the *sample complexity*. Indeed for infinite horizon MDPs and episodic MDPs, τ respectively corresponds to the number of collected samples and the number of played episodes before the algorithm stops. In the latter case, since each episode consists of H samples of the form (R_h, s_{h+1}), τ can be easily linked to the number of samples N through the equation $N = \tau H$.

Remark 1.3 There is a ubiquitous discrepancy in theoretical RL literature between lower and upper bounds on the stopping time. To the best of our knowledge, all the existing lower bound results feature an expectation, i.e., they state that $\mathbb{E}[\tau]$ is always larger than some quantity LB that depends on the problem being considered. On the other hand, with a few exceptions, most upper bounds feature a $(1-\delta)$-quantile, i.e., they state that with probability at least $1-\delta$ we must have $\tau \leq \text{UB}$ for some quantity UB. For this reason, we allow the sample complexity definition above to refer to both measures. This mismatch between lower and upper bounds is also reflected in the definition of minimax optimality below. ■

1.5.1 Minimax complexity and optimality

An interesting quantity in any statistical learning problem is the *minimax rate* which quantifies the performance of algorithms in the worst-case (Tsybakov, 2008). In pure exploration problems, it has the following definition.

Definition 1.5 For some class of MDPs of interest $\mathfrak{M}$, we define the minimax rate as

$$\text{Minimax}(\mathfrak{M}) := \inf_{\mathbb{A}} \sup_{\mathcal{M} \in \mathfrak{M}} \mathbb{E}_{\mathbb{A},\mathcal{M}}[\tau], \quad \text{s.t} : \mathbb{A} \text{ is } (\varepsilon, \delta) - \text{PAC}. \tag{1.25}$$

We will say that an (ε, δ)-PAC algorithm $\mathbb{A}$ is *minimax optimal* on $\mathfrak{M}$ if

$$\forall \mathcal{M} \in \mathfrak{M}, \ \mathbb{P}_{\mathbb{A},\mathcal{M}}\Big(\tau = \widetilde{\mathcal{O}}\big(\text{Minimax}(\mathfrak{M})\big)\Big) \geq 1 - \delta, \tag{1.26}$$

where $\widetilde{\mathcal{O}}$ hides poly-logarithmic factors in $S, A, (1-\gamma), 1/\varepsilon, \log(1/\delta)$.

Since its introduction by (Fiechter, 1994), BPI has mostly been investigated from a minimax perspective. For infinite-horizon MDPs with a discount factor γ, (Azar et al., 2013) showed that $\Omega\big(\frac{SA\log(1/\delta)}{(1-\gamma)^3\varepsilon^2}\big)$ samples are necessary to produce an estimate $\widehat{Q}$ of the action-value function such that $\mathbb{P}_{\mathbb{A},\mathcal{M}}\left(\left\|\widehat{Q} - Q\right\|_\infty \leq \varepsilon\right) \geq 1 - \delta$ using a generative model. Although this lower bound is for algorithms with a different objective (approximating the

optimal Q-function up to ε), there seems to be a consensus in theoretical RL literature that it should also be a valid lower bound for ε-BPI. Hence, a wide variety of works have proposed ε-BPI algorithms that seek to match this bound (Even-Dar et al., 2006; Azar et al., 2013; Sidford et al., 2018a; Agarwal et al., 2020; Li et al., 2020; Kozuno et al., 2022) when a generative model is available. Perhaps the confusion about the lower bound originates from Lemma D.1 in (Sidford et al., 2018a), which states that $\Omega\big(\frac{SA\log(1/\delta)}{(1-\gamma)^3\varepsilon^2}\big)$ samples are also necessary to identify an ε-optimal policy. However, we believe that their proof is false.[5] Therefore, we formulate the following question.

Open question 1.1 Prove that any algorithm that outputs an ε-optimal policy in discounted MDPs with probability larger than $1-\delta$ needs at least $\Omega\big(\frac{SA\log(1/\delta)}{(1-\gamma)^3\varepsilon^2}\big)$ samples.

The picture is more clear for ε-BPI in finite-horizon MDPS. (Dann & Brunskill, 2015) proved that any PAC RL agent must play at least $\Omega(SAH^2\log(1/\delta)/\varepsilon^2)$ episodes to identify an ε-optimal policy in the *worst-case*. Their lower bound was derived under the assumption of time-homogeneous rewards and transitions, i.e. $p_h(.|s,a) = p(.|s,a)$ and $r_h(s,a) = r(s,a)$ for all $h \in [H]$, while a lower bound of $\Omega(SAH^3\log(1/\delta)/\varepsilon^2)$ episodes was later derived by (Domingues et al., 2021) for the time-inhomogeneous case. BPI in the episodic setting was investigated by several works (Dann & Brunskill, 2015; Dann et al., 2019; Kaufmann et al., 2021; Ménard et al., 2021), all of which managed to propose algorithms with polynomial sample complexity. Notably, (Ménard et al., 2021) managed to match the minimax bound for all regimes of ε and δ.

1.5.2 The case for instance-dependent pure exploration

BPI and ε-BPI In order to derive the minimax lower bounds of the previous section, one needs to design very specific hard MDPs. For instance, Figure 1.1 shows the hard MDP class used in (Domingues et al., 2021) to prove the $\Omega(SAH^3\log(1/\delta)/\varepsilon^2)$ bound. In this example, the agent starts at state s_w and can only collect non-zero reward if it reaches the goal state s_g at some step $h \geq \overline{H}+2$, where $\overline{H}$ is a parameter of the MDP. To do that, she has to keep playing the same action a_w exactly $\overline{H}$ times then play a different action at step $h = \overline{H}+1$ to reach an intermediate state s_1. From there, she has to carefully pick the action that has $1/2+\varepsilon'$ probability of making her reach s_g, where $\varepsilon' > \varepsilon$. Playing any other action only yields a chance of $1/2$ to reach s_g and the corresponding policy would not be ε-optimal in that case. A few comments are in order about this construction. First of all, *real-world problems are rarely this difficult*. In particular, the fact that all actions in s_1 have zero reward and are only different by ε' in their transition probabilities makes the problem somewhat hopeless, specifically designed to mislead the learning algorithm. Second, establishing that some algorithm $\mathbb{A}$ *is minimax optimal only reveals that* $\mathbb{A}$ *performs well for this class of worst-case MDPs*. However, it does not indicate whether the algorithm *adapts* to the hardness of the MDP that it faces, i.e., whether the optimal policy of a very easy MDP would be learned very quickly. Indeed, the minimax bounds do not make a distinction between episodic (resp. discounted) MDPs of the same size (S,A,H) (resp. (S,A,γ)). Finally, *in some settings the focus on minimax optimality leads to naive exploration strategies*. For instance, it is known that sampling state-action pairs uniformly is enough to achieve minimax optimality for BPI with a generative model in discounted MDPs (Azar et al., 2013; Sidford et al., 2018a; Agarwal et al., 2020). This uniform sampling is the opposite of what one might expect from any reasonable learning algorithm, that is, gradually focusing its exploration efforts on regions where the reward is higher. This has motivated a recent line

[5] Indeed, their proof makes use of the High-Precision-MDP-Solver from (Sidford et al., 2018b). But they mistakenly state that the sample complexity of that algorithm is $\widetilde{\mathcal{O}}(\frac{S}{(1-\gamma)^3\varepsilon^2})$ instead of $\widetilde{\mathcal{O}}(\frac{SA}{(1-\gamma)^3\varepsilon^2})$.

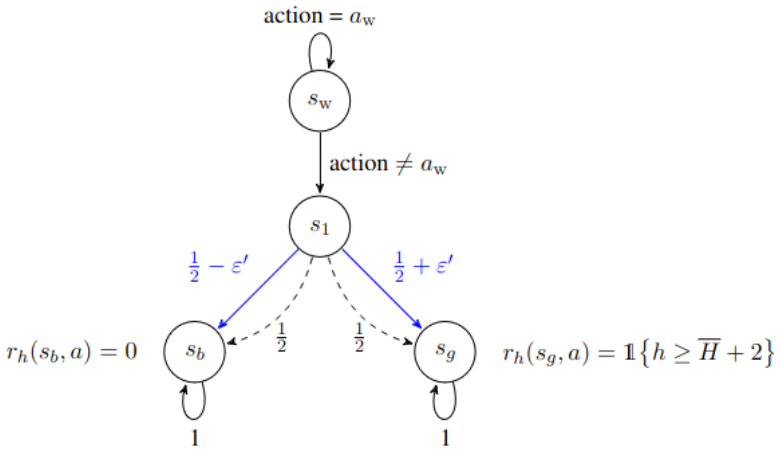


Figure 1.1: class of hard MDPs

of works that focused on designing adaptive algorithms with *instance-dependent* guarantees, i.e., a sample complexity based on properties of the MDP such as sub-optimality gaps. The first algorithm of this kind is BESPOKE (Zanette et al., 2019), which was proposed for discounted MDPs with a generative model. Notably, BESPOKE can adapt to the MDP through a more intelligent sampling scheme than that of minimax algorithms. It solves an optimization problem to compute an optimal vector of samples $(n_{sa})_{s\in\mathcal{S},a\in\mathcal{A}}$ such that querying n_{sa} samples from each (s,a) will halve the uncertainty *only on the value of policies whose current empirical value is above a certain threshold.* By progressively focusing exploration on state-action pairs visited by high-value policies, BESPOKE finds an ε-optimal policy using at most $\widetilde{\mathcal{O}}(\sum_{s,a}\mathcal{C}(s,a,\varepsilon)\log(1/\delta))$ samples where[6]

$$\mathcal{C}(s,a,\varepsilon) := \min\left(\frac{1}{(1-\gamma)^3\varepsilon^2}, \frac{\mathrm{Var}[R(s,a)] + \gamma^2\mathrm{Var}_{s'\sim p(.|s,a)}[V^\star(s')]}{\max(\Delta(s,a),(1-\gamma)\varepsilon)^2} + \frac{1}{(1-\gamma)\max(\Delta(s,a),(1-\gamma)\varepsilon)}\right),$$

$\Delta(s,a) = V^\star(s) - Q^\star(s,a)$ is the value gap of state-action pair (s,a), and Var denotes the variance operator. Two notable features of this result are that the sample complexity of BESPOKE (i) scales as $\mathcal{O}(SA\log(1/\delta)/(1-\gamma)^3\varepsilon^2)$ in the worst-case, which is the conjectured minimax lower bound for this setting (Azar et al., 2013); (ii) it can be significantly smaller than minimax whenever the MDP is such that playing different actions yields very different total rewards, i.e., when the value gaps $(\Delta(s,a))_{s\in\mathcal{S},a\in\mathcal{A}}$ are large compared to ε. Taking inspiration from BESPOKE, we will present in Chapter 2 MDP-NaS, an algorithm for exact BPI in the online setting that builds upon this idea of adapting the sampling strategy to the MDP. We sketch some ideas and results that led to the design of MDP-Nas in Section 1.6.1.

The problem of achieving instance-dependent complexity for ε-BPI in episodic MDPs also attracted some recent attention from the theoretical RL community. The first algorithm with such guarantees is MOCA (Wagenmaker et al., 2022a). Its sample complexity is upper bounded by $\mathcal{C}(\mathcal{M},\varepsilon)\log(1/\delta)$, where $\mathcal{C}(\mathcal{M},\varepsilon)$ is a functional of the MDP that depends on the gaps $(\Delta_h(s,a))_{s,a}$. (Wagenmaker et al., 2022a) show that their complexity is never worse than the minimax lower bound by more than an extra H^2 and $\log^2(1/\delta)$ factors. Therefore, in MDPs where $H \ll SA$ and regimes where δ is not too small, MOCA can improve upon the worst-case lower bound. MOCA is based on coupling a clever exploration strategy with *state-action eliminations*, i.e. using confidence intervals on $Q_h^\star(s,a)$ to detect

[6] $\widetilde{\mathcal{O}}$ hides logarithmic factors in $S, A, 1/(1-\gamma), 1/\varepsilon$ and the gaps $(\Delta(s,a))_{s,a}$.

whether $a \neq \pi_h^\star(s)$ and discarding (h, s, a) in that case. In Chapter 4, we will present PRINCIPLE, an algorithm for ε-BPI with instance-dependent guarantees based on an alternative technique of *policy eliminations*. Instead of looking at state-action pairs, policy eliminations use confidence bounds on the values of policies $(V_1^\pi)_{\pi \in \Pi_S}$ to detect if π is suboptimal. When that is the case, we adjust the sampling rule to cease exploration of the regions visited by π.

RFE Beyond BPI, one may wonder whether it is possible to design adaptive algorithms for RFE and what an instance-dependent complexity might look like for this problem. RFE was introduced by (Jin et al., 2020) who proved that at least $\Omega\big(\frac{S^2AH^3}{\varepsilon^2}\big)$ episodes are necessary to solve the problem in a minimax sense. Later on, (Kaufmann et al., 2021) noted that any (ε, δ)-PAC algorithm for RFE is also (ε, δ)-PAC for BPI, since it can plan ε-optimally for any reward function. This implies that the minimax lower bound of $\Omega(SAH^3 \log(1/\delta)/\varepsilon^2)$ episodes holds also for RFE. Together, these two results yield a minimax rate of $\Omega\big(\frac{H^3SA\log(1/\delta)+S^2AH^3}{\varepsilon^2}\big)$ episodes. This rate was matched by (Ménard et al., 2021) and (Zhang et al., 2021b).

(Wu et al., 2022) showed that there is hope for improving upon this worst-case bound, provided that one introduces additional assumptions about the reward functions used with $\widehat{p}$ for planning at test time. Assuming that there exists a parameter $\rho > 0$ such that the test reward functions induce a minimum value gap $\Delta_{\min}(\mathcal{M})$ larger than ρ, they designed an RFE algorithm with a sample complexity of

$$\widetilde{\mathcal{O}}\left(\frac{H^3SA}{\rho\varepsilon} + \frac{H^4S^2A}{\varepsilon}\right)$$

episodes. Therefore, if we choose ε to be small enough w.r.t ρ and $1/H$, the bound of (Wu et al., 2022) will be smaller than the minimax rate. Beyond such a restricted setting, adaptivity to the MDP in the vanilla version of RFE seems to be a hopeless problem at first glance. Indeed, without further assumptions, the test reward can be chosen adversarially and so one might think that vanilla RFE is a worst-case problem by definition. One of the major contributions of this thesis, which the author of these lines is most proud of, is to show that one can still adapt to the transition kernel of the MDP and achieve a complexity that is smaller than the minimax rate in some regimes.

Contribution 1.1 In Chapter 3, we will present an RFE algorithm named Proportional Coverage Exploration (PCE). With probability $1 - \delta$, the sample complexity of PCE is upper bounded by

$$\widetilde{\mathcal{O}}\left(\mathcal{C}(\mathcal{M}, \varepsilon, \delta) + \frac{\mathrm{poly}(S, A, H)}{\varepsilon}\right),$$

where the functional $\mathcal{C}(\mathcal{M}, \varepsilon, \delta)$ satisfies the following properties

1. For all MDPs,
$$\mathcal{C}(\mathcal{M}, \varepsilon, \delta) \leq \frac{SAH^4 \log(1/\delta) + S^2AH^5}{\varepsilon^2},$$
2. For a class of *"ergodic" MDPs*
$$\mathcal{C}(\mathcal{M}, \varepsilon, \delta) \leq \frac{S^\alpha AH^4 \log(1/\delta) + S^{1+\alpha}AH^5}{\varepsilon^2},$$
where α is a parameter in $(0, 1)$,

3. If the MDP is a "hidden" contextual bandits, i.e., when $p_h(.|s,a) = p_h(.|s)$ for all (h,s,a),

$$\mathcal{C}(\mathcal{M},\varepsilon,\delta) \leq \frac{AH^3\log(1/\delta) + SAH^5}{\varepsilon^2}.$$

We see that up to an additional H^2 factor, the sample complexity of PCE is never worse than the minimax rate in the small ε regime. Furthermore, it has a reduced dependence on the number of states in benign cases such as 2. and 3.

1.6 Overview of Contributions

This section contains some selected contributions from this thesis. We start by deriving a lower bound which we will later use to design an algorithm *à la Track-and-Stop* for the BPI problem in discounted MDPs, see Section 1.6.1. Then we discuss in Section 1.6.2 some limitations of the lower bounds derived using the KL-contraction (also known as the data-processing inequality). We further show through the example of All-ε-BAI how *the simulator technique* can be leveraged to prove tighter bounds in some regimes. Lastly, we present some *coverage* methods that seek to collect observations from an episodic MDP in an efficient manner.

1.6.1 Lower-bound-inspired algorithm for BPI

1.6.1.1 A recipe for optimality from the bandit literature

To study the problem of BPI, we draw inspiration from related work on the special case of *Best Arm Identification* (BAI) in a multi-armed bandit. Assuming that there is a unique optimal arm $a^\star$, the goal in BAI is to identify $a^\star$ with a probability of error smaller than δ, where $\delta \in (0,1)$ is a pre-specified risk. When the arms distributions come from a *single-parameter exponential family*[7] (SPEF), (Garivier & Kaufmann, 2016) propose an instance-dependent lower bound on the sample complexity of any δ-correct BAI algorithm, along with a strategy that matches it. A few notations are due before introducing their results.

Notation The Kullback-Leibler divergence between two distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$\mathrm{KL}(\mathbb{P},\mathbb{Q}) := \begin{cases} \mathbb{E}_{X\sim\mathbb{P}}[\log(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(X))] & \text{if } \mathbb{P} \ll \mathbb{Q} \\ +\infty \quad \text{Otherwise} \end{cases}$$

where $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}$ denotes the Radon-Nikodym derivative of $\mathbb{P}$ with respect to $\mathbb{Q}$. Distributions belonging to the same SPEF can be fully characterized by their means. Therefore, we simply refer to a bandit $(\nu_a)_{a\in[K]}$ by its vector of means $\mu := (\mu_a)_{a\in[K]}$. We use $\mathrm{d}(x,y)$ as a shorthand for the Kullback-Leibler divergence between the distributions, belonging to the SPEF we consider, whose means are x and y respectively. We refer to the set of possible means of a bandit model by Θ, where $\Theta \subset \mathbb{R}^K$. $\mathrm{Alt}(\mu) := \{\lambda \in \Theta : \ a^\star(\lambda) \neq a^\star(\mu)\}$ is the set of alternative bandit models, i.e., bandits with a different optimal arm. For $a \in [K]$, $N_a(t) := \sum_{u=1}^t \mathbb{1}(a_u = a)$ will denote the number of pulls of arm a after t steps of interaction between the algorithm and μ. Finally, we let $\Sigma_K := \{\omega \in \mathbb{R}_+^K, \ \sum_{a\in[K]}\omega_a = 1\}$ denote the simplex of dimension $(K-1)$.

Proposition 1.1 (Theorem 1, (Garivier & Kaufmann, 2016)) The stopping time of any

[7] e.g. Bernoulli, Exponential or Gaussians with a known variance.

δ-correct BAI algorithm $\mathbb{A}$ interacting with the bandit μ is lower bounded as

$$\mathbb{E}_{\mu,\mathbb{A}}[\tau] \geq T^\star(\mu)\log(1/2.4\delta), \text{ where } \quad T^\star(\mu) := \left(\sup_{\omega\in\Sigma_K}\inf_{\lambda\in \mathrm{Alt}(\mu)}\sum_{a\in[K]}\omega_a \mathrm{d}(\mu_a,\lambda_a)\right)^{-1}. \tag{1.27}$$

We briefly recall their proof here, as it will be useful to contrast with another proof method that will be presented in Section 1.6.2.

Proof. We consider an alternative bandit $\lambda \in \mathrm{Alt}\,(\mu)$ and let $\mathrm{kl}(p,q)$ denote the Kullback-Leibler divergence between Bernoulli distributions of respective means p and q. Thanks to Lemma 1 from (Kaufmann et al., 2016), we have that

$$\sum_{a\in[K]} \mathbb{E}_{\mu,\mathbb{A}}[N_a(\tau)]\mathrm{d}(\nu_a,\lambda_a) \geq \mathrm{kl}(\mathbb{P}_{\mu,\mathbb{A}}(\mathcal{E}),\mathbb{P}_{\lambda,\mathbb{A}}(\mathcal{E}))\ , \tag{1.28}$$

for any event $\mathcal{E}$ that is measurable w.r.t the filtration generated by the observations of the algorithm until it stops $(a_1,R_1,\ldots,a_\tau,R_\tau)$. The idea is to come up with an event $\mathcal{E}$ whose probability varies significantly between the two bandit problems. We choose $\mathcal{E} := \big(\widehat{a}=a^\star(\mu)\big)$, where $\widehat{a}$ is the arm answered by the algorithm. Since $\mathbb{A}$ is δ-correct, it holds that $\mathbb{P}_{\mu,\mathbb{A}}(\mathcal{E}) \geq 1-\delta$ while $\mathbb{P}_{\lambda,\mathbb{A}}(\mathcal{E}) \leq \delta$. Using the monotonicity properties of $(x,y)\mapsto \mathrm{kl}(x,y)$, we get that

$$\mathrm{kl}(\mathbb{P}_{\mu,\mathbb{A}}(\mathcal{E}),\mathbb{P}_{\lambda,\mathbb{A}}(\mathcal{E})) \geq \mathrm{kl}(1-\delta,\delta). \tag{1.29}$$

Therefore, since (1.28) and (1.29) hold for any alternative model λ, we have that

$$\begin{aligned}
\mathrm{kl}(1-\delta,\delta) &\leq \inf_{\lambda\in\mathrm{Alt}(\mu)} \mathrm{kl}(\mathbb{P}_{\mu,\mathbb{A}}(\mathcal{E}),\mathbb{P}_{\lambda,\mathbb{A}}(\mathcal{E})) \\
&\leq \inf_{\lambda\in\mathrm{Alt}(\mu)} \sum_{a\in[K]} \mathbb{E}_{\mu,\mathbb{A}}[N_a(\tau)]\mathrm{d}(\nu_a,\lambda_a) \\
&= \mathbb{E}_{\mu,\mathbb{A}}[\tau] \inf_{\lambda\in\mathrm{Alt}(\mu)} \sum_{a\in[K]} \frac{\mathbb{E}_{\mu,\mathbb{A}}[N_a(\tau)]}{\mathbb{E}_{\mu,\mathbb{A}}[\tau]}\mathrm{d}(\nu_a,\lambda_a) \\
&\leq \mathbb{E}_{\mu,\mathbb{A}}[\tau] \sup_{\omega\in\Sigma_K}\inf_{\lambda\in\mathrm{Alt}(\mu)} \sum_{a\in[K]} \omega_a \mathrm{d}(\mu_a,\lambda_a),
\end{aligned} \tag{1.30}$$

where we used that the vector of proportions $\omega := \big(\frac{\mathbb{E}_{\mu,\mathbb{A}}[N_a(\tau)]}{\mathbb{E}_{\mu,\mathbb{A}}[\tau]}\big)_{a\in[K]}$ belongs to the simplex Σ_K. The proof is concluded by noting that $\log(1/2.4\delta) \leq \mathrm{kl}(1-\delta,\delta)$. ■

Observe that the bound of Proposition 1.1 is problem-specific, since it depends on the bandit μ that the algorithm is facing. The authors then propose the Track-and-Stop algorithm which is *asymptotically optimal*, i.e., it satisfies

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\nu,\mathbb{A}}[\tau]}{\log(1/\delta)} \leq T^\star(\mu).$$

The intuition behind Track-and-Stop is that the solution to the optimization program in (1.27) defines a vector of "ideal" frequencies $\omega^\star(\mu) := (\omega_a^\star(\mu))_{a\in[K]}$ according to which every arm must be pulled. This can be seen directly in the last part of the proof above. However, this vector is initially unknown to the algorithm, so the sampling rule of Track-and-Stop is based on *tracking* the optimal vector computed for the empirical bandit $\big(\omega^\star(\widehat{\mu}_a(t))\big)_{a\in[K]}$. The tracking is coupled with a forced exploration component, e.g. pulling

any under-sampled arm a such that $N_a(t) \leq \sqrt{t}$, which ensures consistency of the mean estimator $(\widehat{\mu}_a(t))_{a\in[K]}$. Since Track-and-Stop, several asymptotically optimal algorithms with improved computational cost were later proposed (Degenne et al., 2019a; Jedra & Proutiere, 2020; Wang et al., 2021). These algorithms remove the need to solve (1.27) at every iteration either by using lazy updates (Jedra & Proutiere, 2020), sub-gradient ascent methods (Wang et al., 2021) or online learning algorithms (Degenne et al., 2019a). However, a common property of these algorithms is that they all seek, one way or another, to achieve the following "golden" property

$$\forall a \in [K], \quad \frac{N_a(t)}{t} \xrightarrow[t\to\infty]{a.s.} \omega_a^\star(\mu). \tag{1.31}$$

We shall now detail another contribution which consists of deriving an analogue of the lower bound in (1.27) for the BPI problem, then designing a sampling rule which satisfies the counterpart of the optimality recipe (1.31) in MDPs.

1.6.1.2 A problem-dependent lower bound for BPI

Notation We consider the setting of infinite-horizon discounted MDPs, see Section 1.3.1. The set of alternative MDPs is $\text{Alt}(\mathcal{M}) := \{\mathcal{M}' \in \mathfrak{M}_{\star,1} : \pi^\star(\mathcal{M}') \neq \pi^\star(\mathcal{M})\}$. Let us define the Kullback-Leibler divergence between MDPs $\mathcal{M}$ and $\mathcal{M}'$ at some state-action pair (s,a) by $\text{KL}_{\mathcal{M}|\mathcal{M}'}(s,a) := \text{KL}(q_{\mathcal{M}}(.|s,a), q_{\mathcal{M}'}(.|s,a)) + \text{KL}(p_{\mathcal{M}}(.|s,a), p_{\mathcal{M}'}(.|s,a))$ where KL was defined above. We use $\Sigma = \{\omega \in \mathbb{R}_+^{SA} : \sum_{i=1}^{SA} \omega_i = 1\}$ to denote the simplex of dimension $SA-1$. $\Omega(\mathcal{M}) := \{\omega \in \Sigma : \forall s \in \mathcal{S}, \sum_{a\in\mathcal{A}} \omega_{sa} = \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} p(s|s',a')\omega_{s'a'}\}$ refers to the set of weight vectors that satisfy the *navigation constraints*, otherwise knwon as the *mass-balance equations*. Lastly, $N_{sa}(t) := \sum_{u=1}^{t-1} \mathbb{1}(s_t = s, a_t = a)$ denotes the number of visits to state-action pair (s,a) up to the t-th step of interaction with the MDP.

Contribution 1.2 Our second contribution is an asymptotic lower bound on the sample complexity of BPI algorithms. We actually derive a lower bound that holds for all $\delta > 0$ in the proof. However, the limit bound when δ goes to zero is more interesting as it suggests ideas for designing asymptotically efficient algorithms.

Theorem 1.1 — (Proposition 2, Al-Marjani et al., 2021). The sample complexity of any δ-PAC BPI algorithm $\mathbb{A}$ satisfies,

$$\forall \mathcal{M} \in \mathfrak{M}_{\star,1}, \quad \liminf_{\delta\to 0} \frac{\mathbb{E}_{\mathbb{A},\mathcal{M}}[\tau]}{\log(1/\delta)} \geq T^\star(\mathcal{M}),$$
$$\text{where} \quad T^\star(\mathcal{M}) := \left(\sup_{\omega\in\Omega(\mathcal{M})} \inf_{\mathcal{M}'\in\text{Alt}(\mathcal{M})} \sum_{s,a} \omega_{sa}\text{KL}_{\mathcal{M}|\mathcal{M}'}(s,a)\right)^{-1}. \tag{1.32}$$

Theorem 1.1 gives a fundamental limit to the sample complexity of any BPI algorithm $\mathbb{A}$ that is δ-correct. To better understand the expression of $T^\star(\mathcal{M})$, imagine a zero-sum two-player game between $\mathbb{A}$ and *nature*. $\mathbb{A}$ plays an *allocation vector* in the set $\Omega(\mathcal{M})$ which defines the proportion of time ω_{sa} that $\mathbb{A}$ wants to spend exploring each state-action pair (s,a). Nature then chooses an alternative MDP $\mathcal{M}'$ such that $\mathbb{A}$ will have a hard time distinguishing $\mathcal{M}$ from $\mathcal{M}'$ while using ω as an exploration strategy. In information-theoretic terms, Nature achieves this goal by minimizing the Kullback-Leibler divergence between the distribution of observations under $\mathcal{M}$ and its counterpart under $\mathcal{M}'$. We will prove in (1.39) that this KL divergence is exactly the objective of the optimization program in (1.32). Now, in order to figure out which policy among $\pi^\star(\mathcal{M})$ and $\pi^\star(\mathcal{M}')$ is the correct answer, $\mathbb{A}$ has to distinguish which MDP is actually generating the observations, i.e. it has to maximize this KL divergence. The value of

the resulting max-min optimization program defines the *easiness of learning the optimal policy in* $\mathcal{M}$: the larger the optimal KL divergence is, the easier it is for $\mathbb{A}$ to separate $\mathcal{M}$ from all $\mathcal{M}' \in \mathrm{Alt}(\mathcal{M})$. It is only natural then that taking the inverse of this value gives a lower bound on the sample complexity of BPI.

Remark 1.4 In contrast with the BAI lower bound (1.27) where the allocation vector could take any value within the simplex, here the algorithm can only play a vector that satisfies the navigation constraints. To see why, suppose that there exists a state s that can only be accessed from another state $s^- \in \mathcal{S}$, i.e $\forall (s', a') \in \mathcal{S} \times \mathcal{A},\ p(s|s', a') = \mathbb{1}(s' = s^-)\, p(s|s^-, a')$. In that case, we expect, for any algorithm, a positive correlation between the number of visits to s and to s^-. The navigation constraints capture these dependencies that arise between the number of visits to different states because of the structure of the transition kernel p. ■

The proof of Theorem 1.1 can be decomposed into three steps:

1. First, we show that the vector of expected visits at the stopping time $(\mathbb{E}[N_{sa}(\tau)])_{s\in\mathcal{S},a\in\mathcal{A}}$ satisfies an information-theoretic constraint, see Lemma 1.1. It captures the fact that there is a *minimal number of samples* that any BPI algorithm must collect to distinguish $\mathcal{M}$ from alternative MDPs $\mathcal{M}' \in \mathrm{Alt}(\mathcal{M})$.
2. Second, in Lemma 1.2 we prove the navigation constraints described above.
3. Finally, using the fact that $\mathbb{E}[\tau] = \sum_{s,a} \mathbb{E}[N_{sa}(\tau)]$, we write the corresponding optimization program that bounds the sample complexity.

Lemma 1.1 For all $\mathcal{M}' \in \mathrm{Alt}(\mathcal{M})$, it holds that

$$\sum_{s\in\mathcal{S},a\in\mathcal{A}} \mathbb{E}_{\mathcal{M}}[N_{sa}(\tau)]\mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a) \geq \mathrm{kl}(\delta, 1-\delta);$$

Proof. To simplify the analysis, we will abuse the notation of section 1.4 and write $\mathrm{SMP}(a_t|s_1,\dots,R_{t-1},s_t)$ for the probability that the sampling rule of $\mathbb{A}$ plays action a_t after observing the history $(s_1,\dots,R_{t-1},s_t)$. Similarly $\mathrm{STP}(s_1,\dots,R_{t-1},s_t)$ (resp. $\mathrm{REC}(\pi|s_1,\dots,R_{t-1},s_t)$) will denote the probability that $\mathbb{A}$ decides to stop exploration (resp. recommends the policy π) after observing $\big((s_u,a_u,R_u)_{1\leq u\leq t-1}, s_t\big)$. We recall that μ denotes the distribution of the initial state. Let $\mathbb{P}_{\mathcal{M},\mathbb{A}}$ denote the distributions of the stopping time, trajectories and recommendation when $\mathbb{A}$ interacts with $\mathcal{M}$. $\mathbb{E}_{\mathcal{M},\mathbb{A}}$ will refer to the corresponding expectation operator. Concretely, for any integer $t \geq 1$, sequence $\big((s_u,a_u,x_u)_{1\leq u\leq t-1}, s_t\big) \in (\mathcal{S}\times\mathcal{A}\times[0,1])^{t-1}\times\mathcal{S}$ and policy $\pi \in \Pi_{\mathrm{D}}$,

$$\mathbb{P}_{\mathcal{M},\mathbb{A}}\Big(\tau = t, (S_u,A_u,R_u)_{1\leq u\leq t-1} = (s_u,a_u,x_u)_{1\leq u\leq t-1}, S_t = s_t\ \widehat{\pi} = \pi\Big) := \mu(s_1)$$
$$\times\left[\prod_{u=1}^{t-1} \mathrm{SMP}(a_u|s_1,\dots,s_u) q_{\mathcal{M}}(x_u|s_u,a_u) p_{\mathcal{M}}(s_{u+1}|s_u,a_u)[1-\mathrm{STP}((s_k,a_k,x_k)_{1\leq k\leq u-1},s_u)]\right]$$
$$\times \mathrm{STP}((s_k,a_k,x_k)_{1\leq k\leq t-1},s_t)\mathrm{REC}(\pi|(s_u,a_u,R_u)_{1\leq u\leq t-1},s_t), \tag{1.33}$$

is the probability that the algorithm starts at s_1, plays a sequence of actions that generates the trajectory $(s_u,a_u,x_u)_{1\leq u\leq t}$, stops at the t-th step and returns the policy π. We use $\mathcal{F}$ to denote the filtration generated by the sigma-algebra of the trajectory until $\mathbb{A}$ stops $\sigma\big((s_u,a_u,x_u)_{1\leq u\leq\tau-1}, s_\tau, \widehat{\pi}\big)$. Finally, $\mathrm{kl}(x,y)$ denotes the Kullback-Leibler divergence between Bernoulli distributions of parameters x and y.

The proof starts by fixing an alternative MDP $\mathcal{M}' \in \mathrm{Alt}(\mathcal{M})$. By the *KL-contraction* (Lemma 1 from (Garivier et al., 2019)), for any $\mathcal{F}$-measurable variable Z with values in

$[0,1]$ it holds that

$$\mathrm{KL}(\mathbb{P}_{\mathcal{M},\mathbb{A}}, \mathbb{P}_{\mathcal{M}',\mathbb{A}}) \geq \mathrm{kl}(\mathbb{E}_{\mathcal{M},\mathbb{A}}[Z], \mathbb{E}_{\mathcal{M}',\mathbb{A}}[Z]). \tag{1.34}$$

We take $Z := \mathbb{1}(\widehat{\pi} = \pi^\star(\mathcal{M}))$. By δ-correctness of $\mathbb{A}$, we have that $\mathbb{E}_{\mathcal{M},\mathbb{A}}[Z] = \mathbb{P}_{\mathcal{M},\mathbb{A}}(\widehat{\pi} = \pi^\star(\mathcal{M})) \geq 1-\delta$ while $\mathbb{E}_{\mathcal{M}',\mathbb{A}}[Z] = \mathbb{P}_{\mathcal{M}',\mathbb{A}}(\widehat{\pi} = \pi^\star(\mathcal{M})) \leq \mathbb{P}_{\mathcal{M}',\mathbb{A}}(\widehat{\pi} \neq \pi^\star(\mathcal{M}')) \leq \delta$. Therefore, using the monotonicity properties of $(x,y) \mapsto \mathrm{kl}(x,y)$ we have that

$$\mathrm{kl}(\mathbb{E}_{\mathcal{M},\mathbb{A}}[Z], \mathbb{E}_{\mathcal{M}',\mathbb{A}}[Z]) \geq \mathrm{kl}(1-\delta, \delta). \tag{1.35}$$

On the other hand, by definition of the KL divergence we have that

$$\mathrm{KL}(\mathbb{P}_{\mathcal{M},\mathbb{A}}, \mathbb{P}_{\mathcal{M}',\mathbb{A}}) = \mathbb{E}_{\mathcal{M},\mathbb{A}}\left[\log\left(\frac{\mathrm{d}\mathbb{P}_{\mathcal{M},\mathbb{A}}(\mathcal{O}_\tau)}{\mathrm{d}\mathbb{P}_{\mathcal{M}',\mathbb{A}}(\mathcal{O}_\tau)}\right)\right], \tag{1.36}$$

where $\mathcal{O}_\tau := \big(t, (s_u, a_u, x_u)_{1\leq u\leq \tau-1}, s_\tau, \pi\big)$ is a stream of possible observations. Now we study the log-likelihood ratio of observations under $\mathcal{M}$ and $\mathcal{M}'$. For any

$$\begin{aligned}
L(\mathcal{O}_\tau) &:= \log\left(\frac{\mathrm{d}\mathbb{P}_{\mathcal{M},\mathbb{A}}(\mathcal{O}_\tau)}{\mathrm{d}\mathbb{P}_{\mathcal{M}',\mathbb{A}}(\mathcal{O}_\tau)}\right) \\
&\overset{(a)}{=} \log\left(\prod_{u=1}^{\tau-1} \frac{q_{\mathcal{M}}(x_u|s_u,a_u)p_{\mathcal{M}}(s_u|s_u,a_u)}{q_{\mathcal{M}'}(x_u|s_u,a_u)p_{\mathcal{M}'}(s_u|s_u,a_u)}\right) \\
&= \sum_{s\in\mathcal{S},a\in\mathcal{A}} \underbrace{\sum_{u=1}^{\tau-1} \mathbb{1}(s_u = s, a_u = a)\left(\log\left(\frac{q_{\mathcal{M}}(x_u|s,a)}{q_{\mathcal{M}'}(x_u|s,a)}\right) + \log\left(\frac{p_{\mathcal{M}}(s_{u+1}|s,a)}{p_{\mathcal{M}'}(s_{u+1}|s,a)}\right)\right)}_{:=L_{sa}(\tau)},
\end{aligned} \tag{1.37}$$

where in (a) we simplified by the probabilities of sampling, stopping and recommendation rules in (1.33) which do not depend on the MDP as they are a property of the algorithm. Next we study $L_{sa}(\tau)$ for a given pair (s,a). We introduce the random variables Y_k and Z_k as the next state and the collected reward after the k-th time (s,a) has been visited. We can re-write $L_{sa}(\tau)$ as:

$$L_{sa}(\tau) = \sum_{k=1}^{N_{sa}(\tau)} \left(\log\frac{p_{\mathcal{M}}(Y_k|s,a)}{p_{\mathcal{M}'}(Y_k|s,a)} + \log\frac{q_{\mathcal{M}}(Z_k|s,a)}{q_{\mathcal{M}'}(Z_k|s,a)}\right)$$

Observe that $\xi_k := \log\frac{p_{\mathcal{M}}(Y_k|s,a)}{p_{\mathcal{M}'}(Y_k|s,a)} + \log\frac{q_{\mathcal{M}}(Z_k|s,a)}{q_{\mathcal{M}'}(Z_k|s,a)}$ and $\mathbb{1}_{\{N_{sa}(\tau)>k-1\}}$ are independent, because under the event $\{N_{sa}(\tau) \leq k-1\}$, Y_k and Z_k have not been observed yet. Further notice that $\mathbb{E}_{\mathcal{M}}[\xi_k] = \mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a)$. We deduce that

$$\begin{aligned}
\mathbb{E}_{\mathcal{M}}[L_{sa}(\tau)] &= \mathbb{E}_{\mathcal{M}}\left[\sum_{k=1}^{\infty} \xi_k \mathbb{1}_{\{N_{sa}(\tau)>k-1\}}\right] \\
&= \sum_{k=1}^{\infty} \mathbb{P}_{\mathcal{M}}[N_{sa}(\tau) > k-1]\mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a) \\
&= \mathbb{E}_{\mathcal{M}}[N_{sa}(\tau)]\mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a).
\end{aligned} \tag{1.38}$$

Summing over all pairs (s,a) and plugging this back into (1.36) yields that

$$\mathrm{KL}(\mathbb{P}_{\mathcal{M},\mathbb{A}}, \mathbb{P}_{\mathcal{M}',\mathbb{A}}) = \sum_{s\in\mathcal{S},a\in\mathcal{A}} \mathbb{E}_{\mathcal{M}}[N_{sa}(\tau)]\mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a). \tag{1.39}$$

Combining (1.34) and (1.39) completes the proof. ■

The second ingredient in proving the lower bound is the navigation constraints, which are specific to the MDP setting and are otherwise absent in BAI.

Lemma 1.2 For any algorithm $\mathbb{A}$, and for all states $s \in \mathcal{S}$, we have

$$\left| \sum_{a \in \mathcal{A}} \mathbb{E}_{\mathcal{M},\mathbb{A}}[N_{sa}(\tau)] - \sum_{s',a'} p_{\mathcal{M}}(s|s',a') \mathbb{E}_{\mathcal{M},\mathbb{A}}[N_{s'a'}(\tau)] \right| \leq 1. \tag{1.40}$$

Proof. For convenience we define for all states s, $N_s(t) := \sum_{a \in \mathcal{A}} N_{sa}(t)$. For any $s \in \mathcal{S}$, by looking at the last state-action that was played before each visit to s, we have that

$$N_s(\tau) = \mathbb{1}(s_1 = s) + \sum_{s',a'} \sum_{u=1}^{N_{\tau-1}(s',a')} \mathbb{1}\big(W_{s'a'}(u) = s\big),$$

where $W_{s'a'}(u)$ denotes the next state observed after the u-th time (s',a') has been visited. Now fix (s',a') and let us introduce $G_{s'a'}(t) = \sum_{u=1}^{N_{t-1}(s',a')} \mathbb{1}\big(W_{s'a'}(u) = s\big)$. Observe that the events $\big(W_{s'a'}(u) = s\big)$ and $(N_{t-1}(s',a') > u-1)$ are independent. Furthermore, for any u, $\mathbb{E}_{\mathcal{M},\mathbb{A}}\big[\mathbb{1}\big(W_{s'a'}(u) = s\big)\big] = p_{\mathcal{M}}(s|s',a')$. Hence, by Wald's lemma

$$\mathbb{E}_{\mathcal{M},\mathbb{A}}[G_{s'a'}(\tau)] = p_{\mathcal{M}}(s|s',a') \mathbb{E}_{\mathcal{M},\mathbb{A}}[N_{\tau-1}(s',a')].$$

By plugging this in the first equality and taking the expectation we get

$$\mathbb{E}_{\mathcal{M},\mathbb{A}}[N_\tau(s)] = \mathbb{P}_{\mathcal{M}}[\mathbb{1}(s_1 = s)] + \sum_{s',a'} p_{\mathcal{M}}(s|s',a') \mathbb{E}_{\mathcal{M},\mathbb{A}}[N_{\tau-1}(s',a')]. \tag{1.41}$$

From the above equality, the lemma is proved by just observing that $\mathbb{P}_{\mathcal{M}}[\mathbb{1}(s_1 = s)] \leq 1$, $\mathbb{E}_{\mathcal{M},\mathbb{A}}[N_{\tau-1}(s',a')] \leq \mathbb{E}_{\mathcal{M},\mathbb{A}}[N_\tau(s',a')]$ for any (s',a'), and $\mathbb{E}_{\mathcal{M},\mathbb{A}}[N_s(\tau)] \leq \mathbb{E}_{\mathcal{M},\mathbb{A}}[N_{\tau-1}(s)]+1$ for any s. ■

In the final step in the proof of Theorem 1.1, we wrap-up the constraints from Lemmas 1.2 and 1.1 to get that

$$\mathbb{E}_{\mathcal{M},\mathbb{A}}[\tau] \geq \inf_{\substack{(n_{sa})_{s\in\mathcal{S},a\in\mathcal{A}} \text{ s.t.:} \\ \forall \mathcal{M}' \in \mathrm{Alt}(\mathcal{M}),\ \sum_{s,a} n_{sa} \mathrm{KL}_{\mathcal{M}'|\mathcal{M}}(s,a) \geq \mathrm{kl}(1-\delta,\delta), \\ \forall s \in \mathcal{S},\ \left| \sum_a n_{sa} - \sum_{s',a'} p(s|s',a') n_{s'a'} \right| \leq 1}} \sum_{s\in\mathcal{S},a\in\mathcal{A}} n_{sa}.$$

We define $m_{sa} := n_{sa}/\log(1/\delta)$. Dividing by $\log(1/\delta)$ and taking the $\liminf$ when $\delta \to 0$ we get

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_{\mathcal{M},\mathbb{A}}[\tau]}{\log(1/\delta)} \geq \inf_{\substack{(m_{sa})_{s\in\mathcal{S},a\in\mathcal{A}} \text{ s.t.:} \\ \forall \mathcal{M}' \in \mathrm{Alt}(\mathcal{M}),\ \sum_{s,a} m_{sa} \mathrm{KL}_{\mathcal{M}'|\mathcal{M}}(s,a) \geq 1, \\ \forall s \in \mathcal{S},\ \sum_a m_{sa} = \sum_{s',a'} p(s|s',a') m_{s'a'}}} \sum_{s\in\mathcal{S},a\in\mathcal{A}} m_{sa},$$

where we used that $\mathrm{kl}(1-\delta,\delta) \underset{\delta\to 0}{\sim} \log(1/\delta)$. Next, we define the vector $\omega \in \mathbb{R}^{SA}$ such that $\omega_{sa} = m_{sa}/\sum_{s',a'} m_{s'a'}$. One can check that $\omega \in \Omega(\mathcal{M})$ and that

$$\begin{aligned} \sum_{s\in\mathcal{S},a\in\mathcal{A}} m_{sa} &= \left(\frac{\sum_{s\in\mathcal{S},a\in\mathcal{A}} m_{sa} \mathrm{KL}_{\mathcal{M}'|\mathcal{M}}(s,a)}{\sum_{s\in\mathcal{S},a\in\mathcal{A}} m_{sa}} \right)^{-1} \sum_{s\in\mathcal{S},a\in\mathcal{A}} m_{sa} \mathrm{KL}_{\mathcal{M}'|\mathcal{M}}(s,a) \\ &\geq \left(\sum_{s\in\mathcal{S},a\in\mathcal{A}} \omega_{sa} \mathrm{KL}_{\mathcal{M}'|\mathcal{M}}(s,a) \right)^{-1}. \end{aligned}$$

From here, one can easily show that the value of the optimization program above is larger than the characteristic time $T^\star(\mathcal{M})$ defined in (1.32).

1.6.1.3 Converging to a target allocation vector: From tracking to navigation

In this section, we will assume that there is a unique optimal solution $\omega^\star(\mathcal{M})$ to (1.32) [8]. The question that naturally arises then is how to achieve the counterpart of the optimality recipe from section 1.6.1.1 in MDPs. In other words, given a mapping $\mathcal{M} \mapsto \omega^\star(\mathcal{M})$, we want to design a sampling rule that satisfies

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}, \quad \frac{N_{sa}(t)}{t} \underset{t\to\infty}{\longrightarrow} \omega^\star_{sa}(\mathcal{M}). \tag{1.42}$$

We recall that for MAB problems, the *C-tracking* rule from (Garivier & Kaufmann, 2016) is enough for this purpose. Concretely, given a sequence of optimal allocations for empirical bandits $(\omega^\star(\widehat{\mu}(u)))_{1\leq u\leq t}$, C-tracking pulls the arm defined by

$$a_{t+1} = \begin{cases} \arg\min_{a\in[K]} N_a(t) & \text{if } \min_{a\in[K]} N_a(t) \leq \sqrt{t} - K/2, \\ \arg\min_{a\in[K]} N_a(t) - \sum_{u=1}^t \omega^\star_a(\widehat{\mu}(u)) & \text{otherwise.} \end{cases} \tag{1.43}$$

The first case above corresponds to the *forced exploration* component, which guarantees that $\widehat{\mu}(t) \underset{t\to\infty}{\longrightarrow} \mu$ (and by continuity $\omega^\star(\widehat{\mu}(t)) \underset{t\to\infty}{\longrightarrow} \omega^\star(\mu)$) almost surely. The second case is the tracking component that enables finite-time control of the difference $|N_a(t) - \sum_{u=1}^t \omega^\star_a(\widehat{\mu}(u))|$. To our knowledge, (Degenne et al., 2020) proved the tightest upper bound on this quantity, of order $\mathcal{O}(\log(K))$ at every time step $t \geq 1$. Still in the MAB framework, a remarkably simpler solution in terms of analysis consists in sampling the $(t+1)$-th arm according to $\omega^\star(\widehat{\mu}(t))$, i.e. $a_{t+1} \sim \text{Multinomial}\big(\omega^\star(\widehat{\mu}(t))\big)$. Indeed, (Tirinzoni et al., 2020) showed that this is sufficient to achieve asymptotic optimality in a regret minimization problem. The idea is that under this sampling rule, for every arm $a \in [K]$, the sequence $\big(M_a(t) := N_a(t) - \sum_{u=1}^t \omega^\star_a(\widehat{\mu}(u))\big)_{t\geq 1}$ becomes a martingale of bounded differences w.r.t the filtration generated by the history of observations. Standard martingale concentration results then guarantee that each term M_t is upper bounded by $\widetilde{\mathcal{O}}(\sqrt{t})$ with high probability.

Unfortunately, none of the approaches mentioned above can be transferred straightforwardly to the MDP setting. This is because the convergence (1.42) that we seek must hold over both *states* and actions. Alas, the algorithms can only choose actions to play and observe the next state s_{t+1}, which is the outcome of sampling from the transition kernel $p_{\mathcal{M}}(.|s_t, a_t)$. In other words, we seek to enforce the proportion of time spent exploring sate-action pairs but *we do not have direct control over the state of the environment*. This is the challenge of *navigation* and our third contribution will be to propose a sampling rule, named *C-Navigation*[9] that solves it.

The idea behind C-Navigation is to use the mixing properties of Markov chains to achieve (1.42). Observe that each Markovian policy $\pi \in \Pi_S$ induces a Markov chain on the set of state-action pairs $\mathcal{S} \times \mathcal{A}$ whose transition kernel is given by

$$P_\pi(s', a'|s, a) = \pi(a'|s') p_{\mathcal{M}}(s'|s, a).$$

Furthermore, under mild conditions on the MDP, it can be shown (for instance Theorem 8.8.2 in (Puterman, 1994)) that any vector $\omega \in \Omega(\mathcal{M})$ is the unique stationary distribution of the Markov chain induced by the policy π_ω, where

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}, \ \pi_\omega(a|s) := \begin{cases} \dfrac{\omega_{sa}}{\sum_{b\in\mathcal{A}} \omega_{sb}} & \text{if } \sum_{b\in\mathcal{A}} \omega_{sb} > 0, \\ 1/A & \text{otherwise.} \end{cases} \tag{1.44}$$

[8] While this may not hold in general, we have derived in (Al-Marjani & Proutiere, 2021) a proxy objective which admits a unique maximizer. We elaborate more on this in Chapter 2.

[9] *C*-Navigation stands for cumulative navigation.

Now imagine that we have access to an oracle that gives the value of $\omega^\star(\mathcal{M})$. In this case, we can simply compute the corresponding policy $\pi_{\omega^\star}$ through (1.44) and use it as a sampling rule, i.e., play $a_t \sim \pi_{\omega^\star}(.|s_t)$ at every step $t \geq 1$. Indeed, the Ergodic theorem (see for example Theorem 4.16 in (Levin et al., 2006)) would then guarantee that

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}, \quad \frac{N_{sa}(t)}{t} = \frac{\sum_{u=1}^{t-1} \mathbb{1}\left(s_u = s, a_u = a\right)}{t} \underset{t\to\infty}{\longrightarrow} \omega^\star_{sa}(\mathcal{M}), \tag{1.45}$$

almost surely. We refer to this procedure as the *mixing of the Markov chain* $P_{\pi_{\omega^\star}}$. However, since the algorithm does not know $\omega^\star(\mathcal{M})$, we will apply the mixing procedure to $\omega^\star(\widehat{\mathcal{M}}_t)$, where $\widehat{\mathcal{M}}_t$ is the empirical MDP built using the maximum-likelihood estimates of $(p_{\mathcal{M}}, r_{\mathcal{M}})$. As in C-tracking, we add a forced exploration component to ensure the consistency of our estimates.

Contribution 1.3 In Chapter 2 we will present Navigate-and-Stop, a BPI algorithm based on the C-Navigation sampling rule and prove that it is asymptotically optimal up to a multiplicative factor of 2.

Letting $\pi_{\omega^\star}(\mathcal{M})$ denote the policy extracted from $\omega^\star(\mathcal{M})$ through (1.44), C-Navigation plays an action $a_t \sim \pi_t(.|s_t)$ where

$$\pi_t(a|s) = \frac{\varepsilon_t}{A} + (1-\varepsilon_t)\frac{\sum_{u=1}^{t} \pi_{\omega^\star}(\widehat{\mathcal{M}}_u)(a|s)}{t}, \tag{1.46}$$

and $(\varepsilon_t)_{t\geq 1}$ is a decreasing sequence of *mixture parameters* that converges to zero. In particular, we shall discuss how $(\varepsilon_t)_{t\geq 1}$ must be tuned as a function of the underlying MDP in order to ensure that $\widehat{\mathcal{M}}_t \underset{t\to\infty}{\longrightarrow} \mathcal{M}$ almost surely. Intuitively, this guarantees that $\pi_t \approx \pi_{\omega^\star}(\mathcal{M})$ for t large enough. Hence, we can use an ergodic theorem, albeit for a non-homogeneous Markov Chain in this case, to prove that C-Navigation satisfies (1.42).

1.6.2 Bandit lower bounds beyond the KL contraction: The simulator technique

1.6.2.1 Limits of the classical KL-contraction-based lower bounds

The standard method to derive problem-dependent lower bounds for pure exploration problems follows the proof schemes of Proposition 1.1 and Lemma 1.1[10]. Lower bounds of this style are tight at least in the asymptotic regime $\delta \to 0$, where a wide variety of algorithms are able to match them (Garivier & Kaufmann, 2016; Degenne et al., 2019a; Jedra & Proutiere, 2020; Wang et al., 2021). However, as explained in (Simchowitz et al., 2017), there are scenarios where this result becomes loose in the moderate δ regime. The example given in that paper is for the BAI problem when the class of possible models is restricted to Gaussian bandits with means in the simplex, i.e. $\nu = \big(\mathcal{N}(\mu_a, 1)\big)_{a\in[K]}$ such that $\mu \in \{\mu \in \mathbb{R}^K_{\geq 0},\ \sum_{a\in[K]} \mu_a = 1\}$. If the ground truth bandit μ is such that $\mu_1 = 0.9$ then arm 1 is the best arm. In this case, one can show that the characteristic time $T^\star(\mu)$ defined in (1.27) is less than one. Indeed, one can show that there exists an algorithm such that $\mathbb{E}_{\nu,\mathbb{A}}[\tau] = \mathcal{O}_{\delta\to 0}\big(\log(1/\delta)\big)$. Hence, (1.27) becomes

$$\mathbb{E}_{\theta_1,\mathbb{A}}[\tau] \geq c \log(1/2.4\delta), \tag{1.47}$$

where $c \in (0,1]$. Intuitively, since arms are constrained in the simplex, any alternative instance λ must have $\lambda_1 < 0.5$ so that by Pinsker's inequality $d(\mu_1,\lambda_1) \geq \frac{(0.9-0.5)^2}{2} \geq \Omega(1)$.

[10] This method is efficient only for pure exploration problems with a single correct answer. Problems with multiple correct answers require a more involved analysis, see (Degenne & Koolen, 2019) and (Garivier & Kaufmann, 2021)

Therefore, an algorithm that focuses its sampling effort on arm 1, meaning that $\frac{N_1(t)}{t} \underset{t\to\infty}{\longrightarrow} 1$, is able to distinguish between μ and any $\lambda \in \mathrm{Alt}(\mu)$. But the bound in (1.47) exhibits no dependence at all on the number of arms K! This is at odds with what one might expect from any δ-correct BAI algorithm, that is to sample each arm a few number of times, which would result into $\tau = \Omega(K)$. So why is this linear dependency on K absent from Proposition 1.1 in this setting?

The answer is that there is oracle knowledge embedded in the proof above. More precisely, the proof takes the point of view of an oracle that *already knows the correct answer for μ and its set of alternative instances and only seeks to confirm its beliefs*. In contrast, *any algorithm starts with zero prior knowledge on the ground truth model which generates the observations*, so there must be some sample complexity cost associated with learning that $\mathrm{Alt}(\mu)$ only contains instances such that $d(\mu_1, \lambda_1) = \Omega(1)$. This cost is not captured by the KL-contraction proof scheme, hence we need new methods to derive a more refined lower bound. One such method is the simulator technique which was proposed in (Simchowitz et al., 2017) for BAI. Below, we illustrate their proof method for the BAI problem.

1.6.2.2 Illustrating the simulator technique for Best Arm Identification

Notation: Before stating their result, let us introduce some notations. We denote by $\mathbf{S}_K$ the group of permutations over $[K]$. For a bandit instance $\nu = (\nu_1, \dots, \nu_K)$ we define the *permuted instance* $\pi(\nu) = (\nu_{\pi(1)}, \dots, \nu_{\pi(K)})$. $\mathbf{S}_K(\nu) = \{\pi(\nu),\ \pi \in \mathbf{S}_K\}$ refers to the set of all permuted instances of ν. We will write $\pi \sim \mathbf{S}_K$ to indicate that a permutation is drawn uniformly at random from $\mathbf{S}_K$. Finally, for two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ defined over the same probability space $(\Omega, \mathcal{F})$, $\mathrm{TV}(\mathbb{P}, \mathbb{Q}) := \sup_{\mathcal{E}\subset\Omega} |\mathbb{P}(\mathcal{E}) - \mathbb{Q}(\mathcal{E})|$ is the total-variation distance between $\mathbb{P}$ and $\mathbb{Q}$.

Definition 1.6 An algorithm $\mathbb{A}$ is said to be symmetric if it satisfies for any permutation π, any integer $n \geq 1$ and any sequence of actions $A_1, \dots, A_n$,

$$\mathbb{P}_{\mathbb{A},\nu}\big((a_1, \dots, a_n) = (A_1, \dots, A_n)\big) = \mathbb{P}_{\mathbb{A},\pi(\nu)}\big((a_1, \dots, a_n) = (\pi(A_1), \dots, \pi(A_n))\big).$$

In other words, $\mathbb{A}$ is symmetric if it is indifferent to the order of the arms and acts only based on the underlying distributions. (Simchowitz et al., 2017) showed that for any algorithm $\mathbb{A}$, one can easily build a symmetrized version $\mathbb{A}_{sym}$ such that for any bandit instance ν, $\mathbb{E}_{\pi\sim\mathbf{S}_K}\mathbb{E}_{\mathbb{A},\pi(\nu)}[\tau_\delta] = \mathbb{E}_{\mathbb{A}_{sym},\nu}[\tau_\delta]$. This will be important for the proofs to come, as we only need to consider symmetric algorithms.

Proposition 1.2 — (Theorem 3, Simchowitz et al., 2017). Fix $\delta \leq 1/4$. Any BAI algorithm $\mathbb{A}$ that is δ-correct over the class of Gaussian bandits with means in the simplex satisfies

$$\mathbb{E}_{\pi\sim\mathbf{S}_K}\mathbb{E}_{\mathbb{A},\pi(\nu)}[\tau_\delta] \geq \sum_{a\in[K]\setminus\{a\}} \frac{1-4\delta}{8(\mu^\star - \mu_a)^2}.$$

Thus, for the price of weaker dependence on the risk ($1-4\delta$ instead of $\log(1/\delta)$), the simulator method manages to prove that BAI algorithms must pay a linear cost in terms of the number of arms, even when their means are constrained to be within the simplex.

Proof. We restrict our attention to symmetric algorithms. Throughout the proof it will be useful to represent bandit instances using the *random table model* (Lattimore & Szepesvari, 2019): ν can be defined as a collection of random variables $(X_{a,t})_{a\in[K],t\geq 1}$ where $X_{a,t}$ represents the reward received when playing arm a for the t-th time. Therefore it is enough to specify the law of each $X_{a,t}$ to define ν.

The first step of the proof is to consider permutations where we only swap the best arm with another suboptimal arm: $\pi(a^\star) = a, \pi(a) = a^\star, \pi(b) = b\ \forall b \in [K] \setminus \{a^\star, a\}$, where $a \neq a^\star$. We define the non-stationary bandit instances $\widetilde{\nu}$ and $\widetilde{\pi}$ such that

$\widetilde{\nu}$:

Arm	First n rewards	Next rewards
$a^\star$	$\sim \mathcal{N}(\mu_{a^\star}, 1)$	$\sim \mathcal{N}(\mu_a, 1)$
a	$\sim \mathcal{N}(\mu_a, 1)$	$\sim \mathcal{N}(\mu_{a^\star}, 1)$
$k \in [K] \setminus \{a^\star, a\}$	$\sim \mathcal{N}(\mu_k, 1)$	$\sim \mathcal{N}(\mu_k, 1)$

and

$\widetilde{\pi}$:

Arm	First n rewards	Next rewards
$a^\star$	$\sim \mathcal{N}(\mu_a, 1)$	$\sim \mathcal{N}(\mu_{a^\star}, 1)$
a	$\sim \mathcal{N}(\mu_{a^\star}, 1)$	$\sim \mathcal{N}(\mu_a, 1)$
$k \in [K] \setminus \{a^\star, a\}$	$\sim \mathcal{N}(\mu_k, 1)$	$\sim \mathcal{N}(\mu_k, 1)$

$\widetilde{\nu}$ and $\widetilde{\pi}$ will only serve as intermediate steps in our change-of-measure argument. In particular, we do not require that the algorithm return a good answer on any of them. Let $\mathbb{P}_\lambda$ denote the law of all relevant random variables (rewards, actions played, stopping times..) when running algorithm $\mathbb{A}$ on instance λ and define the event $\mathcal{E} = (N_a(\tau) \leq n)$. Observe that $\mathbb{P}_\nu(\mathcal{E} \cap .) = \mathbb{P}_{\widetilde{\nu}}(\mathcal{E} \cap .)$, since under $\mathcal{E}$ algorithm $\mathbb{A}$ observes the same distribution of rewards. Thus using Bayes' Theorem one can write

$$\begin{aligned}
\mathrm{TV}(\mathbb{P}_{\widetilde{\nu}}, \mathbb{P}_\nu) &= \mathrm{TV}\big(\mathbb{P}_{\widetilde{\nu}}(\mathcal{E}) \times \mathbb{P}_{\widetilde{\nu}}(.|\mathcal{E}) + \mathbb{P}_{\widetilde{\nu}}(\mathcal{E}^c) \times \mathbb{P}_{\widetilde{\nu}}(.|\mathcal{E}^c),\ \mathbb{P}_\nu(\mathcal{E}) \times \mathbb{P}_\nu(.|\mathcal{E}) + \mathbb{P}_\nu(\mathcal{E}^c) \times \mathbb{P}_\nu(.|\mathcal{E}^c)\big) \\
&\overset{(a)}{=} \mathrm{TV}\big(\mathbb{P}_\nu(\mathcal{E}) \times \mathbb{P}_{\widetilde{\nu}}(.|\mathcal{E}) + \mathbb{P}_\nu(\mathcal{E}^c) \times \mathbb{P}_{\widetilde{\nu}}(.|\mathcal{E}^c),\ \mathbb{P}_\nu(\mathcal{E}) \times \mathbb{P}_\nu(.|\mathcal{E}) + \mathbb{P}_\nu(E^c) \times \mathbb{P}_\nu(.|\mathcal{E}^c)\big) \\
&\overset{(b)}{\leq} \mathbb{P}_\nu(\mathcal{E})\mathrm{TV}\big(\mathbb{P}_{\widetilde{\nu}}(.|\mathcal{E}),\ \mathbb{P}_\nu(.|\mathcal{E})\big) + \mathbb{P}_\nu(\mathcal{E}^c)\mathrm{TV}\big(\mathbb{P}_{\widetilde{\nu}}(.|\mathcal{E}^c),\ \mathbb{P}_\nu(.|\mathcal{E}^c)\big) \\
&\overset{(c)}{\leq} \mathbb{P}_\nu(\mathcal{E}^c) = \mathbb{P}_\nu\left(N_a(\tau) > n\right), \qquad (1.48)
\end{aligned}$$

where (a) is because $\mathbb{P}_\nu(\mathcal{E}) = \mathbb{P}_{\widetilde{\nu}}(\mathcal{E})$ hence $\mathbb{P}_\nu(\mathcal{E}^c) = \mathbb{P}_{\widetilde{\nu}}(\mathcal{E}^c)$ also, (b) is by the joint convexity of the TV distance and (c) is because $\mathbb{P}_\nu(E \cap .) = \mathbb{P}_{\widetilde{\nu}}(E \cap .)$ implies that $\mathbb{P}_\nu(.|\mathcal{E}) = \mathbb{P}_{\widetilde{\nu}}(.|\mathcal{E})$. Similarly, by considering event $\mathcal{E}' = (N_{a^\star}(\tau) \leq n)$ one can show that

$$\mathrm{TV}(\mathbb{P}_{\pi(\nu)}, \mathbb{P}_{\widetilde{\pi}}) \leq \mathbb{P}_{\pi(\nu)}\left(N_{a^\star}(\tau) > n\right). \qquad (1.49)$$

Using the above, one can write

$$\begin{aligned}
1 - 2\delta &\overset{(a)}{\leq} \mathbb{P}_\nu\left(\widehat{a} = a^\star\right) - \mathbb{P}_{\pi(\nu)}\left(\widehat{a} = a^\star\right) \\
&\overset{(b)}{\leq} \mathrm{TV}(\mathbb{P}_\nu, \mathbb{P}_{\pi(\nu)}) \\
&\leq \mathrm{TV}(\mathbb{P}_{\pi(\nu)}, \mathbb{P}_{\widetilde{\pi}}) + \mathrm{TV}(\mathbb{P}_{\widetilde{\pi}}, \mathbb{P}_{\widetilde{\nu}}) + \mathrm{TV}(\mathbb{P}_{\widetilde{\nu}}, \mathbb{P}_\nu) \\
&\overset{(c)}{\leq} \mathbb{P}_{\pi(\nu)}(N_{a^\star}(\tau) > n) + \mathbb{P}_\nu(N_a(\tau) > n) + \sqrt{\frac{\mathrm{KL}(\mathbb{P}_{\widetilde{\pi}}, \mathbb{P}_{\widetilde{\nu}})}{2}} \\
&\overset{(d)}{\leq} 2\mathbb{P}_\nu(N_a(\tau) > n) + \sqrt{\frac{\mathrm{KL}(\mathbb{P}_{\widetilde{\pi}}, \mathbb{P}_{\widetilde{\nu}})}{2}}, \qquad (1.50)
\end{aligned}$$

where (a) uses the δ-correctness of $\mathbb{A}$, (b) uses the definition of the total-variation distance, (c) comes from combining (1.48) and (1.49) and using Pinsker's inequality and (d) is because the symmetry of the algorithm implies that $\mathbb{P}_{\pi(\nu)}(N_{a^\star}(\tau) > n) = \mathbb{P}_\nu(N_a(\tau) > n)$. Now denote by $\widetilde{\pi}_i(t)$ (resp. $\widetilde{\nu}_i(t)$) the distribution of the t-th column corresponding to arm i

within the table of $\widetilde{\pi}$ (resp. $\widetilde{\nu}$). Observe that by an analogue of (1.39) for non-stationary bandits, we can write

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{\widetilde{\pi}}, \mathbb{P}_{\widetilde{\nu}}) &= \sum_{i\in[K]} \mathbb{E}_{\widetilde{\pi},\mathbb{A}}\left[\sum_{t=1}^{\tau} \mathbb{1}\left(A_t = i\right) \mathrm{KL}(\widetilde{\pi}_i(t), \widetilde{\nu}_i(t))\right] \\
&\overset{(1)}{=} \sum_{i\in\{a^\star,a\}} \mathbb{E}_{\widetilde{\pi},\mathbb{A}}\left[\sum_{t=1}^{\tau} \mathbb{1}\left(A_t = i\right) \mathrm{KL}(\widetilde{\pi}_i(t), \widetilde{\nu}_i(t))\right] \\
&\overset{(2)}{\leq} n\Big(\mathrm{KL}(\mathcal{N}(\mu_a, 1), \mathcal{N}(\mu_{a^\star}, 1)) + \mathrm{KL}(\mathcal{N}(\mu_{a^\star}, 1), \mathcal{N}(\mu_a, 1))\Big) \\
&= n(\mu^\star - \mu_a)^2
\end{aligned}$$

where (1) is because $\widetilde{\nu}$ and $\widetilde{\pi}$ only differ in the distributions of arms a and $a^\star$, (2) is because this difference only holds for the distributions of the first n rewards. Therefore, the inequality above simplifies to

$$1 - 2\delta - (\mu^\star - \mu_a)\sqrt{n/2} \leq 2\mathbb{P}_{\nu,\mathbb{A}}(N_a(\tau) > n).$$

By setting $n = 1/2(\mu^\star - \mu_a)^2$, we get that $\mathbb{P}_{\nu,\mathbb{A}}\big(N_a(\tau) > 1/2(\mu^\star - \mu_a)^2\big) \geq 1/4 - \delta$. Applying Markov's inequality implies that

$$\frac{1 - 4\delta}{8(\mu^\star - \mu_a)^2} \leq \mathbb{E}_{\nu,\mathbb{A}}[N_a(\tau)].$$

The proof is concluded by summing over all sub-optimal arms. ■

The simulator technique was used by (Mason et al., 2020) for the problem of All ε-Best Arms Identification (All-ε-BAI) in multi-armed bandits, by leveraging a reduction from All-ε-BAI to BAI. In our fourth contribution, we generalize the lower bound of (Mason et al., 2020) and simplify its proof. Notably, our proof demonstrates how the simulator technique can be used in MAB pure exploration problems without the need to perform a reduction to BAI.

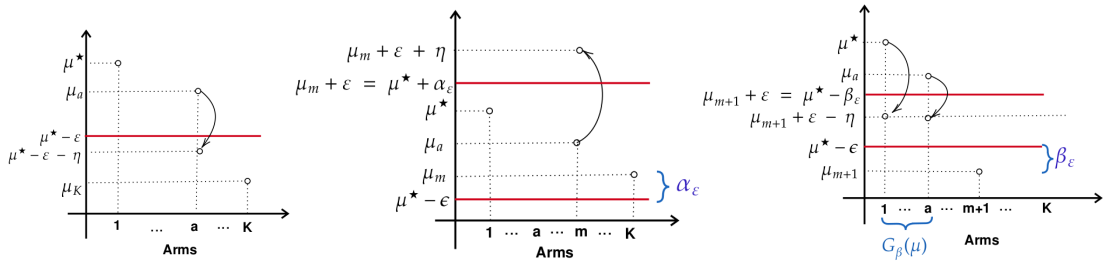
1.6.2.3 Basic analysis of All-ε-BAI

One can not fully grasp the added value of the simulator technique without a brief overview of what the KL-contraction method can achieve for the All-ε-BAI problem. For this purpose, we define the set of alternative bandits $\mathrm{Alt}(\mu) = \{\lambda \in \mathbb{R}^k : \ G_\varepsilon(\lambda) \neq G_\varepsilon(\mu)\}$. Further, define the upper and lower margins

$$\alpha_\varepsilon := \min_{a\in G_\varepsilon(\mu)} \mu_a - \mu^\star + \varepsilon \quad \text{and } \beta_\varepsilon := \min_{b\notin G_\varepsilon(\mu)} \mu^\star - \varepsilon - \mu_b. \tag{1.51}$$

For the simplicity of the presentation, we assume that the arms are ordered decreasingly $\mu^\star = \mu_1 \geq \mu_2 \geq \dots \mu_K$. We let $m := \arg\min_{a\in G_\varepsilon(\mu)} \mu_a - \mu^\star + \varepsilon$ with ties broken in favor of the largest index. Arm m is the arm with the lowest mean among good arms. Since the arms are in decreasing order, arm $m+1$ is necessarily the arm with the largest mean among bad arms and we have $m + 1 = \arg\min_{b\notin G_\varepsilon(\mu)} \mu^\star - \varepsilon - \mu_b$ with ties broken in favor of the smallest index. Let us explore ways to construct alternative instances λ by starting from μ and changing the mean reward of a single arm:

1. **switching the status of an arm:** Fix $\eta > 0$. For any good arm $a \in G_\varepsilon(\mu) \setminus a^\star$, we can lower its mean reward by defining λ such that $\lambda_a = \mu^\star - \varepsilon - \eta$ and $\lambda_b = \mu_b$ for all $b \neq a$, see Figure 1.2a. Note that the new instance satisfies $a \notin G_\varepsilon(\lambda)$. Alternatively, for a bad arm $a \notin G_\varepsilon(\mu)$ we increase its mean reward by letting $\lambda_a = \mu^\star - \varepsilon + \eta$ and



(a) Changing the status of an arm (b) Making arm m ε-suboptimal (c) Making arm $m+1$ ε-optimal

Figure 1.2: Possible changes-of-measure

$\lambda_b = \mu_b$ for all $b \neq a$. One can check that $a \in G_\varepsilon(\lambda)$. In both cases, we have come up with an instance λ such that $G_\varepsilon(\lambda) \neq G_\varepsilon(\mu)$ and

$$\begin{aligned}\sum_{i\in[K]} \mathbb{E}_{\mu,\mathbb{A}}[N_i(\tau)]\mathrm{KL}(\mu_i,\lambda_i) &= \mathbb{E}_{\mu,\mathbb{A}}[N_a(\tau)]\mathrm{KL}(\mu_a,\lambda_a)\\ &= \mathbb{E}_{\mu,\mathbb{A}}[N_a(\tau)]\frac{(\mu_a-\lambda_a)^2}{2}\\ &= \mathbb{E}_{\mu,\mathbb{A}}[N_a(\tau)]\frac{(|\mu_a-\mu^\star+\varepsilon|+\eta)^2}{2},\end{aligned}$$

where we have used the identity $\mathrm{KL}(\mathcal{N}(x,\sigma^2),\mathcal{N}(y,\sigma^2)) = (x-y)^2/2\sigma^2$. Now applying Lemma 1 from (Kaufmann et al., 2016), we have that

$$\begin{aligned}\mathbb{E}_{\mu,\mathbb{A}}[N_a(\tau)]\frac{(|\mu_a-\mu^\star+\varepsilon|+\eta)^2}{2} &\geq \mathrm{kl}\big(\mathbb{P}_{\mathbb{A},\mu}\big(\widehat{G}=G_\varepsilon(\mu)\big),\mathbb{P}_{\mathbb{A},\lambda}\big(\widehat{G}=G_\varepsilon(\mu)\big)\big)\\ &\geq \mathrm{kl}(1-\delta,\delta) \geq \log(1/2.4\delta),\end{aligned}$$

where in the second inequality we used the δ-correctness of $\mathbb{A}$ to establish that $\mathbb{P}_{\mathbb{A},\mu}\big(\widehat{G}=G_\varepsilon(\mu)\big) \geq 1-\delta$ and $\mathbb{P}_{\mathbb{A},\lambda}\big(\widehat{G}=G_\varepsilon(\mu)\big) \leq \mathbb{P}_{\mathbb{A},\lambda}\big(\widehat{G}\neq G_\varepsilon(\lambda)\big) \leq \delta$. Since the inequality above holds for all $\eta > 0$, we take the limit $\eta \to 0$ we get that

$$\forall a \in [|2,K|], \quad \mathbb{E}_{\mu,\mathbb{A}}[N_a(\tau)] \geq \frac{2\log(1/2.4\delta)}{(\mu_a-\mu^\star+\varepsilon)^2}. \tag{1.52}$$

2. **Making arm m bad:** Another way to build alternative instances is by increasing the maximum mean reward so that arm m is no longer within the set of good arms. Concretely, we fix $\eta > 0$, $a \neq m$ then define $\lambda_a = \mu^\star + \alpha_\varepsilon + \eta$ and $\lambda_b = \mu_b$ for all $b \neq a$, see Figure 1.2b. We now have

$$\begin{aligned}\lambda_a &= \mu^\star + \alpha_\varepsilon + \eta\\ &= \mu^\star + (\mu_m - \mu^\star + \varepsilon) + \eta\\ &= \mu_m + \varepsilon + \eta = \lambda_m + \varepsilon + \eta > \lambda_m + \varepsilon,\end{aligned}$$

where the second equality is by definition or m. Thus $m \notin G_\varepsilon(\lambda)$. Proceeding as above, we get that

$$\forall a \in [K] \setminus m, \quad \mathbb{E}_{\mu,\mathbb{A}}[N_a(\tau)] \geq \frac{2\log(1/2.4\delta)}{(\mu_a-\mu^\star-\alpha_\varepsilon)^2}. \tag{1.53}$$

3. **Making arm $m+1$ good:** Finally, we can also decrease the maximum mean reward so that arm $m+1$ becomes a good arm. While increasing the value of $\mu^\star$ can be done by focusing on the mean reward of a single arm, decreasing $\mu^\star$ might require changing the mean reward of more than one arm. Concretely, for all arms a in $G_{\beta_\varepsilon}(\mu) \setminus \{m+1\}$ we define $\lambda_a = \mu^\star - \beta_\varepsilon - \eta$ for some fixed $\eta > 0$. We leave the mean rewards of other arms unchanged, see Figure 1.2c. We thus have for all $a \in G_{\beta_\varepsilon}(\mu) \setminus \{m+1\}$

$$\begin{aligned} \lambda_{m+1} &= \mu_{m+1} \\ &= \mu^\star - \varepsilon - \beta_\varepsilon \\ &= \lambda_a - \varepsilon + \eta > \lambda_a - \varepsilon, \end{aligned} \tag{1.54}$$

where the second inequality is because arm $m+1$ achieves the argmin in the definition of β_ε. On the other hand, for arms in $[K] \setminus (G_{\beta_\varepsilon}(\mu) \cup \{m+1\})$ we have

$$\begin{aligned} \lambda_{m+1} &= \mu^\star - \varepsilon - \beta_\varepsilon \\ &\geq \mu_a - \varepsilon \\ &= \lambda_a - \varepsilon. \end{aligned}$$

Therefore $m+1 \in G_\varepsilon(\lambda)$. Applying Lemma 1 from (Kaufmann et al., 2016) and letting η go to zero we get that

$$\sum_{a \in G_{\beta_\varepsilon}(\mu) \setminus \{m+1\}} \mathbb{E}_{\mu,\mathbb{A}}[N_a(\tau)] \frac{(\mu_a - \mu^\star + \beta_\varepsilon)^2}{2} \geq \log(1/2.4\delta).$$

Since $\mu^\star - \beta_\varepsilon \leq \mu_a \leq \mu^\star$ for all $a \in G_{\beta_\varepsilon}(\mu)$, $(\mu_a - \mu^\star + \beta_\varepsilon)^2 \leq \beta_\varepsilon^2$ and the inequality becomes

$$\sum_{a \in G_{\beta_\varepsilon}(\mu) \setminus \{m+1\}} \mathbb{E}_{\mu,\mathbb{A}}[N_a(\tau)] \geq \frac{2\log(1/2.4\delta)}{\beta_\varepsilon^2}. \tag{1.55}$$

The lower bounds in (1.52) and (1.53) reveal that (ε, δ)-PAC algorithms must pay a minimum cost, in terms of samples, *for every arm but one*. However, (1.55) *only establishes a sample complexity cost for a special subset of arms*, namely those that are within β_ε margin from the optimal mean reward. (1.55) reflects the requirement that we should estimate the mean reward of *at least one arm in* $G_{\beta_\varepsilon}(\mu)$ up to β_ε precision. If we fail to do so, we might severely underestimate the value of $\mu^\star$ and wrongfully declare that $m+1$ is a good arm. While this cost is specific to $G_{\beta_\varepsilon}(\mu)$ (underestimating arms outside of this set does not change our answer about arm $m+1$), *the proof above does not take into account the fact that algorithms have to sample all arms a certain amount of time before learning which belong to* $G_{\beta_\varepsilon}(\mu)$. Here is yet another example of "oracle knowledge" within the proof, which affects the tightness of the resulting lower bound. This is where our contribution comes into the picture.

1.6.2.4 The simulator technique for bandit problems with many "special" arms

Here we use the same notation as Section 1.6.2.2.

> **Contribution 1.4** Our fourth contribution is a problem-dependent lower bound, averaged over all the possible permutations of the bandit ν.

Theorem 1.2 — (Theorem 3, Al Marjani et al., 2022). Fix $\delta \leq 1/10$ and $\varepsilon > 0$. Consider an instance ν such that there exists at least one bad arm: $G_\varepsilon(\mu) \neq [K]$. Then any (ε, δ)-PAC All-ε-BAI algorithm has an average sample complexity over all permuted instances satisfying

$$\mathbb{E}_{\pi\sim \mathbf{S}_K}\mathbb{E}_{\pi(\nu),\mathbb{A}}[\tau] \geq \frac{2\log(1/2.4\delta)}{\beta_\varepsilon^2} + \frac{1}{12|G_{\beta_\varepsilon}(\mu)|^3}\sum_{b\in[K]\setminus G_{\beta_\varepsilon}(\mu)}\frac{1}{(\mu^\star - \mu_b + \beta_\varepsilon)^2},$$

While we have previously assumed that the mean-rewards of arms are ordered decreasingly w.r.t their index, this is only done for the purpose of the analysis. In practice, the arms of a bandit may come in any arbitrary order. The averaging over permutations eliminates any artificial "luck" that an algorithm might have on ν just because it assumes a particular order of arms.

Remark 1.5 In the special case where $|G_{2\beta_\varepsilon}(\mu)| = 1$, then $|G_{\beta_\varepsilon}(\mu)| = 1$ also (since $\{1\} \subset G_{\beta_\varepsilon}(\mu) \subset G_{2\beta_\varepsilon}(\mu)$) and we recover the result of Theorem 4.1 of (Mason et al., 2020). The lower bound above informs us that we must pay a linear cost in K, *even when there are several arms close to the best arm*, provided that their cardinal does not scale with the total number of arms, i.e. $|G_{\beta_\varepsilon}| = \mathcal{O}(1)$.
Furthermore, we shall present in Chapter 5 a bandit instance where the lower bound obtained through KL-contraction can be arbitrarily smaller than the lower bound of Theorem 1.2. ∎

Proof of Theorem 1.2

We restrict our attention to symmetric algorithms and use the random table model from Section 1.6.2.2 to represent bandits. The first step of the proof is to show that no arm can be played significantly less than the arms in $G_{\beta_\varepsilon}(\mu)$. This is the purpose of the lemma below, which helps us avoid algorithmic reductions of the All-ε-BAI problem to BAI or β-isolated tests as was done in (Mason et al., 2020).

Lemma 1.3 For all arms $b \in [K] \setminus G_{\beta_\varepsilon}(\mu)$ and all integers $n \geq 1$,

$$\frac{1}{|G_{\beta_\varepsilon}(\mu)|}\sum_{a\in G_{\beta_\varepsilon}(\mu)}\mathbb{P}_{\nu,\mathbb{A}}\left(N_a(\tau) > n\right) - (\mu^\star - \mu_b)\sqrt{n/2} \leq 3\mathbb{P}_{\nu,\mathbb{A}}(N_b(\tau) > n).$$

Proof. Fix $a \in G_{\beta_\varepsilon}(\mu)$ and $n \geq 1$. Let π be the permutation that swaps arms a and b, i.e. $\pi(a) = b, \pi(b) = a$ and $\pi(k) = k$ for $k \in [K] \setminus \{a, b\}$. We define the non-stationary bandit instances $\widetilde{\nu}$ and $\widetilde{\pi}$ such that

$\widetilde{\nu}$:

Arm	First n rewards	Next rewards
a	$\sim \mathcal{N}(\mu_a, 1)$	$\sim \mathcal{N}(\mu_a, 1)$
b	$\sim \mathcal{N}(\mu_b, 1)$	$\sim \mathcal{N}(\mu_a, 1)$
$k \in [K] \setminus \{a, b\}$	$\sim \mathcal{N}(\mu_k, 1)$	$\sim \mathcal{N}(\mu_k, 1)$

and

$\widetilde{\pi}$:

Arm	First n rewards	Next rewards
a	$\sim \mathcal{N}(\mu_b, 1)$	$\sim \mathcal{N}(\mu_a, 1)$
b	$\sim \mathcal{N}(\mu_a, 1)$	$\sim \mathcal{N}(\mu_a, 1)$
$k \in [K] \setminus \{a, b\}$	$\sim \mathcal{N}(\mu_k, 1)$	$\sim \mathcal{N}(\mu_k, 1)$

Again, $\widetilde{\nu}$ and $\widetilde{\pi}$ will only serve as intermediate steps in our change-of-measure argument. In particular, we do not require that the algorithm return a good answer on any of them. Let $\mathbb{P}_\lambda$ denote the law of all relevant random variables (rewards, actions played, stopping times..) when running algorithm $\mathbb{A}$ on instance λ and define the event $E = (N_b(\tau) \leq n)$. Observe that $\mathbb{P}_\nu(E \cap .) = \mathbb{P}_{\widetilde{\nu}}(E \cap .)$, since under E algorithm $\mathbb{A}$ observes the same distribution of rewards. Following the same steps that lead to (1.48) we have that

$$\mathrm{TV}(\mathbb{P}_{\widetilde{\nu}}, \mathbb{P}_\nu) \leq \mathbb{P}_\nu(E^c) = \mathbb{P}_\nu(N_b(\tau) > n), \tag{1.56}$$

Similarly, by considering event $E' = (N_a(\tau) \leq n)$ it holds that

$$\mathrm{TV}(\mathbb{P}_{\pi(\nu)}, \mathbb{P}_{\widetilde{\pi}}) \leq \mathbb{P}_{\pi(\nu)}(N_a(\tau) > n). \tag{1.57}$$

Using the above, one can write

$$\begin{aligned}
\mathbb{P}_\nu(N_a(\tau) > n) - \mathbb{P}_{\pi(\nu)}(N_a(\tau) > n) &\overset{(a)}{\leq} \mathrm{TV}(\mathbb{P}_\nu, \mathbb{P}_{\pi(\nu)}) \\
&\leq \mathrm{TV}(\mathbb{P}_{\pi(\nu)}, \mathbb{P}_{\widetilde{\pi}}) + \mathrm{TV}(\mathbb{P}_{\widetilde{\pi}}, \mathbb{P}_{\widetilde{\nu}}) + \mathrm{TV}(\mathbb{P}_{\widetilde{\nu}}, \mathbb{P}_\nu) \\
&\overset{(b)}{\leq} \mathbb{P}_{\pi(\nu)}(N_a(\tau) > n) + \mathbb{P}_\nu(N_b(\tau) > n) + \sqrt{\frac{\mathrm{KL}(\mathbb{P}_{\widetilde{\pi}}, \mathbb{P}_{\widetilde{\nu}})}{2}} \\
&\overset{(c)}{\leq} 2\mathbb{P}_\nu(N_b(\tau) > n) + \sqrt{\frac{\mathrm{KL}(\mathbb{P}_{\widetilde{\pi}}, \mathbb{P}_{\widetilde{\nu}})}{2}},
\end{aligned} \tag{1.58}$$

where (a) is thanks to the definition of the total variation, (b) comes from combining (1.56) and (1.57) and using Pinsker's inequality and (c) is because the symmetry of the algorithm implies that $\mathbb{P}_{\pi(\nu)}(N_a(\tau) > n) = \mathbb{P}_\nu(N_b(\tau) > n)$. Now denote by $\widetilde{\pi}_i(t)$ (resp. $\widetilde{\nu}_i(t)$) the distribution of the t-th column corresponding to arm i within the table of $\widetilde{\pi}$ (resp. $\widetilde{\nu}$). By an analogue of (1.39) for non-stationary instances, we can write

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{\widetilde{\pi}}, \mathbb{P}_{\widetilde{\nu}}) &= \sum_{i \in [K]} \mathbb{E}_{\widetilde{\pi}, \mathbb{A}}\left[\sum_{t=1}^{\tau} \mathbb{1}(A_t = i)\, \mathrm{KL}(\widetilde{\pi}_i(t), \widetilde{\nu}_i(t))\right] \\
&\overset{(1)}{=} \sum_{i \in \{a,b\}} \mathbb{E}_{\widetilde{\pi}, \mathbb{A}}\left[\sum_{t=1}^{\tau} \mathbb{1}(A_t = i)\, \mathrm{KL}(\widetilde{\pi}_i(t), \widetilde{\nu}_i(t))\right] \\
&\overset{(2)}{\leq} n\Big(\mathrm{KL}(\mathcal{N}(\mu_a, 1), \mathcal{N}(\mu_b, 1)) + \mathrm{KL}(\mathcal{N}(\mu_b, 1), \mathcal{N}(\mu_a, 1))\Big) \\
&= n(\mu_a - \mu_b)^2 \overset{(3)}{\leq} n(\mu^\star - \mu_b)^2
\end{aligned}$$

where (1) is because $\widetilde{\nu}$ and $\widetilde{\pi}$ only differ in the distributions of arms a and b, (2) is because this difference only holds for the distributions of the first n rewards and (3) is because $\mu_b \leq \mu_a$ since $b \notin G_{\beta_\varepsilon}(\mu)$. Therefore, the inequality above is simplified to

$$\mathbb{P}_{\nu, \mathbb{A}}(N_a(\tau) > n) - (\mu^\star - \mu_b)\sqrt{n/2} \leq 3\mathbb{P}_{\nu, \mathbb{A}}(N_b(\tau) > n).$$

Note that the inequality above holds trivially when $a = b$. Now, for a fixed b, by summing the inequality over all arms $a \in G_{\beta_\varepsilon}(\mu)$ we get

$$\sum_{a \in G_{\beta_\varepsilon}(\mu)} \mathbb{P}_\nu(N_a(\tau) > n) - |G_{\beta_\varepsilon}(\mu)|(\mu^\star - \mu_b)\sqrt{n/2} \leq 3|G_{\beta_\varepsilon}(\mu)|\mathbb{P}_\nu(N_b(\tau) > n).$$

Hence the statement of the lemma. ■

Remark 1.6 In the proof above, we built the non-stationary instance $\widetilde{\nu}$ (resp. $\widetilde{\pi}$) so that it "simulates" the multi-armed bandit ν (resp. $\pi(\nu)$), i.e., it generates rewards from the same distributions for the first n pulls of arms b (resp. arm a). Hence no algorithm can distinguish between ν and $\widetilde{\nu}$ (resp. $\pi(\nu)$ and $\widetilde{\pi}$) unless it has a non-zero probability of pulling arm b (resp. arm a) more than n times. Unlike the KL-contraction proof where we did a single change-of-measure $\nu \longrightarrow \lambda$, the simulator technique relies on perfoming 3 changes-of-measure: $\nu \xrightarrow{(1)} \widetilde{\nu} \xrightarrow{(2)} \widetilde{\pi} \xrightarrow{(3)} \pi(\nu)$ (see the inequalities leading to (1.58)). The underlying intuition is that the sampling behaviour of $\mathbb{A}$, represented by the probabilities of the event $(N_a(\tau) > n)$, will not differ much between ν and $\pi(\nu)$ if:

1. $\widetilde{\nu}$ is almost indistinguishable from ν
2. n is small enough that $\mathrm{KL}(\mathbb{P}_{\widetilde{\pi}}, \mathbb{P}_{\widetilde{\nu}})$ is negligible
3. $\widetilde{\pi}$ is almost indistinguishable from $\pi(\nu)$.

The right choice of n will be dictated by the next Lemma. ■

The second step in proving Theorem 1.2 is to show that arms in $G_{\beta_\varepsilon}(\mu)$ must be pulled $\Omega(1/\beta_\varepsilon^2)$ times because underestimating their means by β_ε may cause the algorithm to declare arm $m+1$ as ε-optimal.

Lemma 1.4 For all integers $n \geq 1$,

$$1 - 2\delta - |G_{\beta_\varepsilon}(\mu)|\beta_\varepsilon\sqrt{n}/2 \leq \sum_{a \in G_{\beta_\varepsilon}(\mu)} \mathbb{P}_{\nu,\mathbb{A}}\left(N_a(\tau) > n\right).$$

Proof. Let $\eta > 0$. We define the instances λ and $\widetilde{\nu}$ such that

λ :

Arm	All rewards
For $a \in G_{\beta_\varepsilon}(\mu)$	$\sim \mathcal{N}(\mu^\star - \beta_\varepsilon - \eta, 1)$
For $k \in [K] \setminus G_{\beta_\varepsilon}(\mu)$	$\sim \mathcal{N}(\mu_k, 1)$

and

$\widetilde{\nu}$:

Arm	First n rewards	Next rewards
For $a \in G_{\beta_\varepsilon}(\mu)$	$\sim \mathcal{N}(\mu_a, 1)$	$\sim \mathcal{N}(\mu^\star - \beta_\varepsilon - \eta, 1)$
For $k \in [K] \setminus G_{\beta_\varepsilon}(\mu)$	$\sim \mathcal{N}(\mu_k, 1)$	$\sim \mathcal{N}(\mu_k, 1)$

By considering the event $E = (\forall a \in G_{\beta_\varepsilon}(\mu),\ N_a(\tau) \leq n)$, one can show in a similar fashion to the proof of Lemma 1.3 that

$$\mathrm{TV}(\mathbb{P}_{\widetilde{\nu}}, \mathbb{P}_\nu) \leq \mathbb{P}_\nu(\exists a \in G_{\beta_\varepsilon}(\mu),\ N_a(\tau) > n) \leq \sum_{a \in G_{\beta_\varepsilon}(\mu)} \mathbb{P}_\nu\left(N_a(\tau) > n\right). \tag{1.59}$$

Recall that $m+1 \in \arg\min_{k \notin G_\varepsilon(\mu)} \mu^\star - \varepsilon - \mu_k$ and observe that $m+1$ becomes an ε-optimal arm under λ (see (1.54)). Thus we have $\mathbb{P}_\lambda(m+1 \notin \widehat{G}_\varepsilon) \leq \delta$ while $\mathbb{P}_\nu(m+1 \notin \widehat{G}_\varepsilon) \geq 1 - \delta$.

Therefore

$$\begin{aligned}
1-2\delta &\leq \mathbb{P}_\nu(m+1 \notin \widehat{G}_\varepsilon) - \mathbb{P}_\lambda(m+1 \notin \widehat{G}_\varepsilon) \\
&\leq \mathrm{TV}(\mathbb{P}_\nu, \mathbb{P}_\lambda) \\
&\leq \mathrm{TV}(\mathbb{P}_\nu, \mathbb{P}_{\widetilde{\nu}}) + \mathrm{TV}(\mathbb{P}_{\widetilde{\nu}}, \mathbb{P}_\lambda) \\
&\overset{(a)}{\leq} \sum_{a \in G_{\beta_\varepsilon}(\mu)} \mathbb{P}_\nu\left(N_a(\tau) > n\right) + \sqrt{\frac{\mathrm{KL}(\mathbb{P}_{\widetilde{\nu}}, \mathbb{P}_\lambda)}{2}} \\
&= \sum_{a \in G_{\beta_\varepsilon}(\mu)} \mathbb{P}_\nu\left(N_a(\tau) > n\right) + \sqrt{\frac{n\sum_{a \in G_{\beta_\varepsilon}(\mu)}(\mu_a - \mu^\star + \beta_\varepsilon + \eta)^2/2}{2}} \\
&\overset{(b)}{=} \sum_{a \in G_{\beta_\varepsilon}(\mu)} \mathbb{P}_\nu\left(N_a(\tau) > n\right) + \sqrt{\frac{n|G_{\beta_\varepsilon}(\mu)|(\beta_\varepsilon + \eta)^2}{4}} \\
&\overset{(c)}{=} \sum_{a \in G_{\beta_\varepsilon}(\mu)} \mathbb{P}_\nu\left(N_a(\tau) > n\right) + |G_{\beta_\varepsilon}(\mu)|(\beta_\varepsilon + \eta)\sqrt{n}/2
\end{aligned}$$

where (a) comes from (1.59) and Pinsker's inequality, (b) is because all arms in $G_{\beta_\varepsilon}(\mu)$ satisfy $\mu^\star - \beta_\varepsilon \leq \mu_a \leq \mu^\star$ and (c) comes from the fact that $\sqrt{|G_{\beta_\varepsilon}(\mu)|} \leq |G_{\beta_\varepsilon}(\mu)|$. Note that the inequality above holds for all $\eta > 0$. We get the final result by taking the limit $\eta \to 0$. ■

In the final step of the proof, we combine the results of Lemmas 1.3 and 1.4 to get for all $b \in [K] \setminus G_{\beta_\varepsilon}(\mu)$

$$\frac{1-2\delta}{3|G_{\beta_\varepsilon}(\mu)|} - (\mu^\star - \mu_b + \beta_\varepsilon)\sqrt{n}/6 \leq \mathbb{P}_\nu(N_b(\tau) > n).$$

Thus by choosing $n = \left\lceil \frac{(1-2\delta)^2}{|G_{\beta_\varepsilon}(\mu)|^2(\mu^\star - \mu_b + \beta_\varepsilon)^2} \right\rceil$ we get

$$\frac{1-2\delta}{6|G_{\beta_\varepsilon}(\mu)|} \leq \mathbb{P}_\nu(N_b(\tau) > n) \leq \mathbb{P}_\nu\left(N_b(\tau) \geq \frac{(1-2\delta)^2}{|G_{\beta_\varepsilon}(\mu)|^2(\mu_1 - \mu_b + \beta_\varepsilon)^2}\right),$$

which implies by Markov's inequality that for all $b \in [K] \setminus G_{\beta_\varepsilon}(\mu)$,

$$\frac{(1-2\delta)^3}{6|G_{\beta_\varepsilon}(\mu)|^3(\mu_1 - \mu_b + \beta_\varepsilon)^2} \leq \mathbb{E}_\nu[N_b(\tau)].$$

The final result is obtained by summing the inequality over arms in $[K] \setminus G_{\beta_\varepsilon}(\mu)$, adding (1.55) and noting that for $\delta \leq 1/10, (1-2\delta)^3 \geq 1/2$.

Open question 1.2 At a high level, the simulator technique relies on the fact that all instances in $\{\pi(\nu)\}_{\pi \in \mathbf{S}_K}$ are somewhat equivalent, since only the indexing of arms changes from one permutation to another. Therefore, a lower bound averaged over all instances in that set still reflects the hardness of the bandit that our algorithm is facing. However, this property no longer holds when we consider other bandit settings with structure, e.g. Lipschitz bandits (Magureanu et al., 2014) or Linear bandits (Soare et al., 2014). For instance in Lipschitz bandits, given arms $(x_a)_{a \in [K]} \in [0,1]^K$ the mean rewards of arms must satisfy $\forall (i,j),\ |\mu_i - \mu_j| \leq L|x_i - x_j|$ for some constant $L > 0$. Hence, permutating the arms may break the Lipschitz property. This raises the following question: How to generalize the simulator technique to other bandit settings where some structure is embedded into the arms distributions? In particular, what is a possible class of "equivalent" bandits that one can use to prove a refined problem-dependent lower

bound in such settings?

1.6.3 Covering an MDP: minimum flows in graphs, submodular optimization and zero-sum games

So far we have seen in Section 1.6.1 how to derive an instance-dependent lower bound for BPI in discounted MDPs using the KL contraction method. We also briefly sketched an exploration strategy that leads to an asymptotically optimal algorithm by solving the max-min program of this bound and following the resulting allocation vector $\omega^\star(\mathcal{M})$. In Section 1.6.2, we explained in a MAB setting why the KL contraction lower bounds can be loose in the moderate δ-regime. In addition, we showed through the example of All-ε-BAI how to derive tighter bounds for pure exploration using the simulator technique. This motivates us to look for algorithmic guarantees beyond the asymptotic $\delta \to 0$ regime, by seeking to design algorithms with sample complexity upper bounds that hold for all $\delta \in (0, 1)$.

To that end, we developed an efficient method to *cover* an MDP, i.e., a sampling rule that collects observations from any desired subset of state-action pairs using a minimal number of episodes. We refer to this task as *coverage* of an MDP and we shall see in Chapters 3 and 4 how efficient coverage is a powerful tool for designing pure exploration algorithms. Notably, Contribution 1.1 would not have been possible without the study and use of a near-optimal coverage algorithm.

In this section, we will present some results established in this thesis for coverage in the case of deterministic transitions. As it turns out, there are some interesting connections between the coverage of an MDP, solving flow problems on a graph and submodular optimization. We will also briefly sketch some results for the general case of stochastic MDPs, which will be further improved in Chapter 3. The contents are extracted from appendices B and D of the conference paper:

Andrea Tirinzoni, Aymen Al Marjani, and Emilie Kaufmann. **Near instance-optimal PAC reinforcement learning for deterministic MDPs**. In *Advances in Neural Information Processing Systems* (NeurIPS), 2022.

1.6.3.1 Deterministic MDPs as directed acyclic graphs

We consider the setting of episodic MDPs (Section 1.3.2). We are interested in the case where transition kernels are deterministic, i.e., when for all $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$, there exists a unique state s' such that $p_h(s'|s, a) = 1$. Under this property, we can equivalently represent the transitions by a sequence of deterministic functions $\{f_h : \mathcal{S} \times \mathcal{A} \to \mathcal{S}\}_{h\in[H]}$ such that $f_h(s, a)$ is the unique s' defined earlier. A deterministic MDP then becomes the tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, H, \{f_h\}_{h\in[H]}, \{q_h\}_{h\in[H]}, s_1)$.

Assumption 1.4 We assume that the transitions $\{f_h\}_{h\in[H]}$ are knwon to the learner.

Assumption 1.4 is without loss of generality. Indeed if the transitions are unknown, Proposition 2 in (Ortner, 2010) shows how we can recover them using no more than SAH episodes. A procedure that achieves this is as follows: At the beginning of any episode t, given the "known part" of the transitions, we find the closest state with an unexplored action. We reach this state and play the action in question. Since there are altogether SAH triplets (h, s, a) to explore, the total number of episodes needed is at most SAH.

Our second observation is that, if we ignore the reward distributions, $\mathcal{M}$ can be represented as a *layered directed acyclic graph* (DAG) $\mathcal{G}(\mathcal{M}) := (\mathcal{N}, \mathcal{E}, s_1, s_{H+1})$ with nodes $\mathcal{N} := \{(s, h) : h \in [H], s \in \mathcal{S}\}$, arcs $\mathcal{E} := \{(s, a, h) : h \in [H], s \in \mathcal{S}, a \in \mathcal{A}\}$, a unique *source* node $(s_1, 1)$, and a fictitious *sink* node $(s_{H+1}, H + 1)$ which is the endpoint of every arc

$(s, a, H) \in \mathcal{E}$. In particular, for node $(s, h) \in \mathcal{N}$, there is one arc for each $a \in \mathcal{A}$ which connects the node to $(f_h(s, a), h + 1)$. The graph is *layered*, in the sense that the set of nodes can be partitioned into H subsets $(\{(s, h) : s \in \mathcal{S}\})_{h\in[H]}$, one for each stage, and transitions are possible only between adjacent stages. Let $\mathcal{I}_h(s) := \{(s', a') \in \mathcal{S} \times \mathcal{A} \mid s' \in \mathcal{S}_{h-1}, a' \in \mathcal{A}_{h-1}(s), f_{h-1}(s', a') = s\}$ be the set of incoming arcs into (s, h).

1.6.3.2 The minimum flow problem and its properties

We define a *flow* as any non-negative function $\eta : \mathcal{E} \to [0, \infty)$ that satisfies the navigation constraints

$$\sum_{(s',a')\in\mathcal{I}_h(s)} \eta_{h-1}(s', a') = \sum_{a\in\mathcal{A}} \eta_h(s, a) \quad \forall h > 1, s \in \mathcal{S}. \tag{1.60}$$

We let Ω be the set of all flows. The value of a flow η is given by $\varphi(\eta) := \sum_{a\in\mathcal{A}} \eta_1(s_1, a)$. Let $c : \mathcal{E} \to [0, \infty)$ be a non-negative *target function*. We say that a flow η is *feasible* if

$$\eta_h(s, a) \geq c_h(s, a) \quad \forall (s, a, h) \in \mathcal{E}.$$

That is, $c_h(s, a)$ acts as a lower bound on the flow we require through arc (s, a, h). The minimum flow for the target function c is the solution to the linear program,

$$\varphi^\star(c) := \min_{\eta\in\Omega} \sum_{a\in\mathcal{A}} \eta_1(s_1, a) \quad \text{s.t.} \quad \eta_h(s, a) \geq c_h(s, a) \quad \forall (s, a, h) \in \mathcal{E}. \tag{1.61}$$

Intuitively, the goal is to minimize the amount of flow leaving the initial state while satisfying the navigation and demand constraints. From the MDP perspective, if $\eta_h(s, a)$ is the number of times our algorithm $\mathbb{A}$ visited a triplet (h, s, a) then $\sum_{a\in\mathcal{A}} \eta_1(s_1, a)$ is the total number of episodes played by $\mathbb{A}$ (since each episode starts by playing an action at the initial state s_1). Therefore, computing a minimum flow corresponds to minimizing the number of episodes that are required to visit each triplet at least the amount of times prescribed by the target function c.

We now state some simple properties of flows which will be useful later on. The first two lemmas below can be immediately derived from the LP formulation.

Lemma 1.5 — Monotonicity. Let $c^1, c^2 : \mathcal{E} \to [0, \infty)$ be such that $c^1_h(s, a) \leq c^2_h(s, a)$ for all $(s, a, h) \in \mathcal{E}$. Then

$$\varphi^\star(c^1) \leq \varphi^\star(c^2).$$

Lemma 1.6 Let c^1, c^2 be two non-negative lower bound functions and $\alpha > 0$. Then,

$$\varphi^\star(\alpha c^1 + c^2) \leq \alpha\varphi^\star(c^1) + \varphi^\star(c^2).$$

Lemma 1.7 — Flow bounds. For any lower bound function c,

$$\max_{h\in[H]} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} c_h(s, a) \leq \varphi^\star(c) \leq \sum_{h\in[H]} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} c_h(s, a).$$

Proof. Both inequalities are easy to see from the navigation constraints and the definition of the value of a minimum flow. First, note that the navigation constraints imply that for

flow vector η and any $h \in [H]$,

$$\begin{aligned}\varphi(\eta) &= \sum_{a\in\mathcal{A}} \eta_1(s_1, a) \\ &\overset{(i)}{=} \sum_{s\in\mathcal{S},a\in\mathcal{A}} \eta_h(s, a) \\ &\overset{(ii)}{\geq} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} c_h(s, a),\end{aligned}$$

where (i) is thanks to the navigation constraints (1.60) and (ii) is because the flow is feasible. Taking the maximum over h, we get the lower bound on $\varphi^\star(c)$. The upper bound is trivial since we can construct a feasible flow in the following fashion. For each (h, s, a), we define a flow vector $\eta^{hsa} \in \Omega$ by starting with $\eta^{hsa}_h(s, a) = c_h(s, a)$ and $\eta^{hsa}_\ell(\widetilde{s}, \widetilde{a}) = 0$ elsewhere. Then we propagate the flow into the adjacent layers. In other words, for the successor state $s' = f_h(s, a)$ we choose one action a' and set $\eta^{hsa}_{h+1}(s', a') = c_h(s, a)$. Similarly, we choose one predecessor state-action pair $(s^-, a^-) \in \mathcal{I}_h(s)$ and set $\eta^{hsa}_{h-1}(s^-, a^-) = c_h(s, a)$. By doing this recursively, we build a flow vector η^{hsa} that satisfies $\eta^{hsa}_h(s, a) \geq c_h(s, a)$ and whose value is exactly $c_h(s, a)$. Then a feasible flow is given by the sum vector

$$\eta = \sum_{h,s,a} \eta^{hsa}.$$

By Lemma 1.6, $\varphi(\eta) = \sum_{h,s,a} \varphi(\eta^{hsa}) = \sum_{h,s,a} c_h(s, a)$. Therefore the value of the minimum flow is at most this quantity. ■

1.6.3.3 Minimum policy covers and minimum flows

A crucial problem that arises when trying to solve ε-BPI in a deterministic MDP is the problem of computing a *minimum policy cover*. Imagine that we have run our ε-BPI algorithm for $t \geq 1$ episodes and collected $n^t_h(s, a)$ observations from each triplet (h, s, a). Using these, we built high- probability confidence intervals on the optimal action-values

$$Q^\star_h(s, a) \in [\underline{Q}_h(s, a), \overline{Q}_h(s, a)] \quad \forall (h, s, a).$$

Based on the confidence intervals we can already establish that for every $(h, s) \in [H] \times \mathcal{S}$, actions a such that $\overline{Q}_h(s, a) < \max_{b\in\mathcal{A}} \underline{Q}_h(s, b)$ are sub-optimal. Such triplets (h, s, a) no longer need to be explored since we know that no optimal policy plays a at (h, s). We say that they are *eliminated*. Therefore, we only want to collect observations from a subset of triplets $(h, s, a) \in \{[H] \times \mathcal{S} \times \mathcal{A} : \ \overline{Q}_h(s, a) \geq \max_{b\in\mathcal{A}} \underline{Q}_h(s, b)\}$. This motivates us to study the *minimum policy cover* problem.

Formally, given a subset $\mathcal{E}' \subseteq \mathcal{E}$ of the arcs (i.e., of the state-action-stage triplets), the goal is to find a set of policies $\Pi_{\text{cover}} \subseteq \Pi$ of *minimum size* such that

$$\forall (s, a, h) \in \mathcal{E}', \exists \pi \in \Pi_{\text{cover}} : (s^\pi_h, a^\pi_h) = (s, a).$$

That is, Π_{cover} is the smallest set of policies that, played together, visit all arcs in $\mathcal{E}'$. This problem can be easily reduced to a minimum flow problem with target function

$$c_h(s, a) := \mathbb{1}\left((s, a, h) \in \mathcal{E}'\right),$$

which intuitively demands at least one visit to all $(s, a, h) \in \mathcal{E}'$, and zero visits from the other triplets. Moreover, since c is integer-valued, an integer minimum flow exists which can be computed by existing algorithms (e.g., Brandizi et al., 2012). Suppose that $\eta^\star$ is one such integer minimum flow. A policy cover can be easily extracted from it by the procedure shown in Algorithm 4, which is similar to the method proposed by (Brandizi et al., 2012) to obtain a minimum path cover in a layered DAG.

Algorithm 4 Static Maximum Coverage

Input: deterministic MDP (without reward) $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \{f_h\}_{h\in[H]}, s_1, H)$
Solve LP (1.61) with targets $c_h(s,a) = \mathbb{1}\left((s,a,h) \in \mathcal{E}'\right)$ to get $\eta^\star$
Set $\eta \leftarrow \eta^\star$
Initialize $\Pi_{\text{cover}} \leftarrow \emptyset$
while $\varphi(\eta) > 0$ **do**
  Initialize a policy π with arbitrary actions
  for $h = 1, \dots, H$ **do**
    $\pi_h(s_h) \leftarrow \arg\max_{a\in\mathcal{A}_h(s_h)} \eta_h(s,a)$
    $\eta_h(s_h, \pi_h(s_h)) \leftarrow \eta_h(s_h, \pi_h(s_h)) - 1$
    $s_{h+1} \leftarrow f_h(s_h, \pi_h(s_h))$
  end for
  $\Pi_{\text{cover}} \leftarrow \Pi_{\text{cover}} \cup \{\pi\}$
end while

Lemma 1.8 — size of policy cover. Let Π_{cover} be the size of the policy cover returned by Algorithm 4. Then

$$|\Pi_{\text{cover}}| = \varphi^\star(\mathbb{1}_{\mathcal{E}'}).$$

Proof. Note that at every iteration of Algorithm 4, the value of the flow η is decreased by one while the cardinal of Π_{cover} is increased by the same amount. Since Algorithm 4 only stops when the value of the update flow is zero, this means that $|\Pi_{\text{cover}}| = \varphi(\eta^\star) = \varphi^\star(\mathbb{1}_{\mathcal{E}'})$. ■

1.6.3.4 Dynamic Maximum Coverage and submodular maximization

While Static Maximum Coverage solves the minimum policy cover problem with optimal sample complexity, it is not the most intuitive strategy one would think of to explore an MDP. We would like to analyze a simpler strategy, named Dynamic Maximum Coverage and hopefully prove some sample complexity guarantees for it too. At every iteration, Dynamic Maximum Coverage solves a Dynamic Program for some reward $\widetilde{r}$ and plays the resulting policy. The exploration reward $\widetilde{r}$ is initialized as an indicator reward over all triplets in $\mathcal{E}'$: $\widetilde{r}_h(s,a) := \mathbb{1}\left((s,a,h) \in \mathcal{E}'\right)$, then updated each time by setting zero reward for the triplets that were visited. The pseudo-code of Dynamic Maximum Coverage is reported in Algorithm 5.

Algorithm 5 Dynamic Maximum Coverage

Input: deterministic MDP (without reward) $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \{f_h\}_{h\in[H]}, s_1, H)$
Initialize number of visits $n_h(s,a) \leftarrow O$ for all (h,s,a)
while $\min_{(h,s,a)\in\mathcal{E}'} n_h(s,a) < 1$ **do**
  Compute $\pi^t \leftarrow \arg\max_{\pi\in\Pi_{\text{D}}} \sum_{h=1}^{H} \mathbb{1}\left((h,s,a) \in \mathcal{E}',\ n_h(s_h^\pi, a_h^\pi) < 1\right)$
  for $h = 1, \dots, H$ **do**
    Play action $\pi_h(s_h)$
    $n_h(s_h, \pi_h(s)) \leftarrow n_h(s_h, \pi_h(s)) + 1$
    $s_{h+1} \leftarrow f_h(s_h, \pi_h(s_h))$
  end for
end while

Reduction to submodular maximization Let us define the set function $C : 2^{\Pi_D} \to [0, \infty)$ as

$$C(\Pi') := \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathbb{1}\left((h, s, a) \in \mathcal{E}',\ \exists \pi \in \Pi' : (s_h^\pi, a_h^\pi) = (s, a)\right) \quad \forall \Pi' \subseteq \Pi_D.$$

Moreover, let $\bar{\Pi}_i$ be the set containing the first i policies played by Dynamic Maximum Coverage. We note that the policy selection strategy of Dynamic Maximum Coverage (Line 4 of Algorithm 5) is essentially a greedy algorithm approximating the maximization of C. In fact, maximizing C corresponds to finding a set of policies that visit all triplets in $\mathcal{E}'$. Instead of directly maximizing the set function C, Dynamic Maximum Coverage greedily builds the set $\bar{\Pi}_i$ by adding, at each round where it is used, the policy visiting most of these unvisited triplets. Let us prove some of the important properties of C.

First, we relate the maximization of C to the computation of a minimum flow with target function $c_h(s, a) \leftarrow \mathbb{1}\left((h, s, a) \in \mathcal{E}'\right)$, i.e., the same one used by Static Maximum Coverage.

Proposition 1.3 — Maximization vs minimum flow. For each $v \geq \varphi^\star(\mathbb{1}_{\mathcal{E}'})$,

$$\max_{\Pi' \subseteq \Pi_D : |\Pi'| \leq v} C(\Pi') = \max_{\Pi' \subseteq \Pi_D} C(\Pi') = |\mathcal{E}'|.$$

Proof. Clearly, $C(\Pi') \leq |\mathcal{E}'|$ for all $\Pi' \subseteq \Pi_D$, which is attained when all state-action-stage triplets in $\mathcal{E}'$ are visited at least once. When the cardinality of Π' can be at least $\varphi^\star(\mathbb{1}_{\mathcal{E}'})$, we can choose Π' to include a set of $\varphi^\star(\mathbb{1}_{\mathcal{E}'})$ policies realizing a minimum 1-flow (i.e., a minimum policy cover as the one computed by Static Maximum Coverage). These, by definition, cover all under-visited triplets and thus attain the maximal value $|\mathcal{E}'|$. ■

Observe that if $\Pi' \subseteq \Pi''$ then Π'' must visit at least all the triplets visited by Π'. Therefore, the following proposition holds.

Proposition 1.4 — Monotonicity. For each $\Pi' \subseteq \Pi'' \subseteq \Pi_D$, $C(\Pi') \leq C(\Pi'')$.

Proposition 1.5 — Sub-modularity. Function C is sub-modular, i.e., for every $\Pi' \subseteq \Pi'' \subseteq \Pi_D$ and $\bar{\pi} \in \Pi_D \setminus \Pi''$,

$$C(\Pi' \cup \{\bar{\pi}\}) - C(\Pi') \geq C(\Pi'' \cup \{\bar{\pi}\}) - C(\Pi'').$$

Proof. Note that

$$\begin{aligned}
&C(\Pi' \cup \{\bar{\pi}\}) - C(\Pi') \\
&\quad := \sum_{(h,s,a) \in \mathcal{E}'} \mathbb{1}\left((s_h^{\bar{\pi}}, a_h^{\bar{\pi}}) = (s, a), \neg \exists \pi \in \Pi' : (s_h^\pi, a_h^\pi) = (s, a)\right) \\
&\quad = \sum_{h=1}^{H} \mathbb{1}\left(\neg \exists \pi \in \Pi' : (s_h^\pi, a_h^\pi) = (s_h^{\bar{\pi}}, a_h^{\bar{\pi}})\right) \\
&\quad \geq \sum_{h=1}^{H} \mathbb{1}\left(\neg \exists \pi \in \Pi'' : (s_h^\pi, a_h^\pi) = (s_h^{\bar{\pi}}, a_h^{\bar{\pi}})\right) \\
&\quad = C(\Pi'' \cup \{\bar{\pi}\}) - C(\Pi''),
\end{aligned}$$

where the inequality holds since $\Pi' \subseteq \Pi''$. ■

Proposition 1.6 — Greedy maximization. Let $\bar{\Pi}_i$ be the set containing the first $i \geq 0$ policies computed by Dynamic Maximum Coverage. Then, for any positive integer v,

$$C(\bar{\Pi}_i) \geq (1 - e^{-(i+1)/v}) \max_{\Pi' \subseteq \Pi_D : |\Pi'| \leq v} C(\Pi').$$

Proof. This is a simple consequence of Theorem 1.5 of (Krause & Golovin, 2014) on greedy maximization of submodular functions. We just need to show that Dynamic Maximum Coverage is greedily maximizing the function C. To that end, observe that at iteration $i+1$ of Dynamic Maximum Coverage, we can rewrite the objective in Line 4 if Algorithm 5 as

$$\begin{aligned} f(\pi) &:= \sum_{h=1}^{H} \mathbb{1}\left((h, s, a) \in \mathcal{E}',\ n_h(s_h^\pi, a_h^\pi) < 1\right) \\ &= \sum_{(h,s,a) \in \mathcal{E}'} \mathbb{1}\left(\neg \exists \bar{\pi} \in \bar{\Pi}_i :\ (s_h^\pi, a_h^\pi) = (s_h^{\bar{\pi}}, a_h^{\bar{\pi}})\right) \\ &= \sum_{h \in [H], s \in \mathcal{S}, a \in \mathcal{A}} \mathbb{1}\left((h, s, a) \in \mathcal{E}',\ (s_h^{\bar{\pi}}, a_h^{\bar{\pi}}) = (s, a), \neg \exists \bar{\pi} \in \bar{\Pi}_i :\ (s_h^{\bar{\pi}}, a_h^{\bar{\pi}}) = (s, a)\right) \\ &= C(\bar{\Pi}_i \cup \{\pi\}) - C(\bar{\Pi}_i). \end{aligned}$$

■

Contribution 1.5 We now state the main theorem of this section.

Theorem 1.3 — (Theorem 10, Tirinzoni et al., 2022). The number of episodes played by Dynamic Maximum Coverage is upper-bounded

$$d \leq \varphi^\star(\mathbb{1}_{\mathcal{E}'})(\log(H) + 1).$$

Thus we see that the sample complexity of Dynamic Maximum Coverage is nearly optimal, as we lose at most a logarithmic factor in the horizon compared to solving the minimum flow LP.

Proof. Let $\underline{i} := \sup_{i \in \mathbb{N}} \{i : C(\bar{\Pi}_i) \leq |\mathcal{E}'| - \varphi^\star(\mathbb{1}_{\mathcal{E}'})\}$ be the last iteration at which at least $\varphi^\star(\mathbb{1}_{\mathcal{E}'})$ triplets still need to be visited by the algorithm. Then, by Proposition 1.6 combined with Proposition 1.3,

$$\begin{aligned} |\mathcal{E}'| - \varphi^\star(\mathbb{1}_{\mathcal{E}'}) &\geq C(\bar{\Pi}_{\underline{i}}) \\ &\geq (1 - e^{-(\underline{i}+1)/\varphi^\star(\mathbb{1}_{\mathcal{E}'})}) \max_{\Pi' \subseteq \Pi : |\Pi'| \leq \varphi^\star(\mathbb{1}_{\mathcal{E}'})} C(\Pi') \\ &= (1 - e^{-(\underline{i}+1)/\varphi^\star(\mathbb{1}_{\mathcal{E}'})}) |\mathcal{E}'|. \end{aligned}$$

Thus,

$$(\underline{i} + 1) \leq \varphi^\star(\mathbb{1}_{\mathcal{E}'}) \log(|\mathcal{E}'|/\varphi^\star(\mathbb{1}_{\mathcal{E}'})) \leq \varphi^\star(\mathbb{1}_{\mathcal{E}'}) \log(H),$$

where the second inequality holds since $\varphi^\star(\mathbb{1}_{\mathcal{E}'}) \geq \max_{h \in [H]} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathbb{1}\left((h, s, a) \in \mathcal{E}'\right)$ by Lemma 1.7 and $|\mathcal{E}'| \leq H \max_{h \in [H]} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathbb{1}\left((h, s, a) \in \mathcal{E}'\right)$. This implies that $\underline{i} \leq \varphi^\star(\mathbb{1}_{\mathcal{E}'}) \log(H) - 1$. Finally, note that $d \leq \underline{i} + \varphi^\star(\mathbb{1}_{\mathcal{E}'})$ since at iteration $\underline{i} + 1$ less than $\varphi^\star(\mathbb{1}_{\mathcal{E}'})$ triplets are missing and the algorithm visits at least a new one every episode. ■

1.6.3.5 Two-player zero-sum games for efficient coverage in the stochastic case

In this section, we go back to the general case of coverage, i.e. when the target function c is not necessarily the indicator over some subset of triplets and the transition kernel of $\mathcal{M}$ is no longer assumed to be deterministic.

Definition 1.7 — Active coverage algorithms. Given a failure probability $\delta \in (0,1)$, we want an algorithm that explores the MDP and with probability at least $1-\delta$ collects a $c_h(s,a)$ observations from every triplet (h,s,a). We say that such an algorithm is (δ, c)-correct for coverage.

Motivation We shall see that a (δ, c)-correct coverage algorithm, named CovGame, is the backbone of an RFE algorithm and an ε-BPI algorithm respectively presented in Chapters 3 and 4. In fact, we will show that one can use CovGame in a plug-and-play fashion to solve the RFE problem, simply by setting an appropriate target function $c_h(s,a) \propto \sup_{\pi \in \Pi_D} p_h^\pi(s,a)$. On the other hand, for ε-BPI we combine CovGame with a new technique that we call *"Implicit policy eliminations"* to get an algorithm that enjoys instance-dependent sample complexity, see Section 1.6.4. Here we present a simplified version of CovGame, as it conveys the main ideas behind solving coverage while still being quite simple to analyze.

Intuition Let $\mathcal{X} := \{(h,s,a) : c_h(s,a) > 0\}$ be the set of triplets to be covered and define $p_h^{\pi^{exp}}(s,a) := \mathbb{P}_{\mathcal{M},\pi^{exp}}(s_h = s, a_h = s)$. We will prove in Theorem 3.1 that any (δ, c)-correct coverage algorithm needs roughly more than

$$\varphi^\star(c) := \inf_{\pi^{exp} \in \Pi_S} \max_{(s,a,h) \in \mathcal{X}} \frac{c_h(s,a)}{p_h^{\pi^{exp}}(s,a)}\,, \tag{1.62}$$

episodes (in expectation) to complete the coverage task. Now, observe that

$$\begin{aligned}
\frac{1}{\varphi^\star(c)} &= \sup_{\pi^{exp} \in \Pi_S} \min_{(s,a,h) \in \mathcal{X}} \frac{p_h^{\pi^{exp}}(s,a)}{c_h(s,a)} \\
&= \sup_{\pi^{exp} \in \Pi_S} \inf_{\lambda \in \Sigma_{\mathcal{X}}} \sum_{h,s,a} \frac{p_h^{\pi^{exp}}(s,a)\lambda_h(s,a)}{c_h(s,a)} \\
&= \sup_{\pi^{exp} \in \Pi_S} \inf_{\lambda \in \Sigma_{\mathcal{X}}} \mathbb{E}_{\mathcal{M},\pi^{exp}}\left[\sum_{h,s,a} \frac{\mathbb{1}(s_h = s, a_h = a)\lambda_h(s,a)}{c_h(s,a)}\right],
\end{aligned}$$

where $\Sigma_{\mathcal{X}} := \{\omega \in \mathbb{R}_+^{|\mathcal{X}|} : \sum_i \omega_i = 1\}$ is the simplex with support over $\mathcal{X}$. We see that the inverse of the lower bound above is the value of a two-player zero-sum game between a first player that plays a policy π^{exp} to explore the MDP and a second player that plays a weight vector λ in the simplex $\Sigma_{\mathcal{X}}$. Moreover, the objective of the max-min program above is the value function of π^{exp} for a particular reward $\widetilde{r}_h(s,a) := \lambda_h(s,a)/c_h(s,a)$. The previous observations suggest to use a gaming approach that shares some similarities with Dynamic Maximum Coverage. Specifically, we use the same idea of designing a suitable exploration reward which we will try to maximize by running an algorithm for regret minimization as a subroutine. However, instead of handcrafting the reward ourselves, we let another competing algorithm design it for us. This gives rise to the meta-algorithm described in Algorithm 6, which employs a regret minimizer $\mathcal{A}^\Pi$ and an online learner $\mathcal{A}^\lambda$ as a subroutine. The idea is that $\mathcal{A}^\lambda$ is penalized whenever is outputs large weights for some triplets (h,s,a) that were easily visited by $\mathcal{A}^\Pi$. By doing so, we make $\mathcal{A}^\lambda$ challenge $\mathcal{A}^\Pi$ more by putting higher rewards in triplets that are hard to reach, thereby making the exploration process efficient.

Remark 1.7 We shall prove in Lemma 3.1 that the quantity $\varphi^\star(c)$ defined in (1.62) boils down to the minimum flow defined in (1.61) whenever $\mathcal{M}$ has deterministic transitions. ■

Algorithm 6 Simplified CovGame

Input: target function c, regret minimization algorithm $\mathcal{A}^\Pi$, online learner $\mathcal{A}^\lambda$, risk δ.
Initialize dataset of episodes $\mathcal{D}_0 \leftarrow \emptyset$
Set target set $\mathcal{X} \leftarrow \{(s,a,h) \in [H] \times \mathcal{S} \times \mathcal{A} : c_h(s,a) > 0\}$
Normalize targets $\widetilde{c}_h(s,a) \leftarrow c_h(s,a)/c_{min} \quad (c_{min} := \min_{(h,s,a)\in\mathcal{X}} c_h(s,a))$
Initialize challenger weights $\lambda_h^1(s,a) \leftarrow \mathbb{1}((h,s,a) \in \mathcal{X})/|\mathcal{X}|$ for all h,s,a
for $t = 1, 2, \ldots$ **do**
  Define reward $R_h^t(s,a) = \mathbb{1}((h,s,a) \in \mathcal{X})\lambda_h^t(s,a)/\widetilde{c}_h(s,a)$ for all h,s,a
  Feed $\mathcal{A}^\Pi$ with R^t, confidence $\delta/2$ and get exploration policy π^t
  Play π^t and observe trajectory $\mathcal{H}_t := \{(s_h^t, a_h^t, s_{h+1}^t)\}_{1\le h\le H-1}$
  Update dataset $\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} \cup \mathcal{H}_t$.
  Feed $\mathcal{A}^\lambda$ with loss

$$\ell^t(\lambda) = \sum_{(h,s,a)\in\mathcal{X}} \lambda_h(s,a) \frac{\mathbb{1}(s_h^t = s, a_h^t = a)}{\widetilde{c}_h(s,a)}$$

  and get new weight vector λ^{t+1}
  If $\forall (h,s,a),\ n_h(s,a;\mathcal{D}_t) \ge c_h(s,a)$**:** Stop and return $\mathcal{D}_t$
end for

Assumption 1.5 There exists a sublinear fonction $T \mapsto \mathcal{R}^\lambda(T)$ that bounds the regret of $\mathcal{A}^\lambda$ anytime, i.e.

$$\forall T \in \mathbb{N}^*,\ \sum_{t=1}^T \ell^t(\lambda^t) - \min_{\lambda\in\Sigma_\mathcal{X}} \sum_{t=1}^T \ell^t(\lambda) \le \mathcal{R}^\lambda(T) \text{ a.s. .} \tag{1.63}$$

Furthermore, there exists a sublinear fonction $T \mapsto \mathcal{R}^\Pi(T,\delta)$ that upper bounds the dynamic regret of $\mathcal{A}^\Pi$ with high-probability, i.e., for any sequence of reward functions $(R^t)_{t\ge1} \in (\Sigma_\mathcal{X})^\mathbb{N}$,

$$\mathbb{P}_{\mathcal{M},\mathcal{A}^\Pi}\left(\forall T \in \mathbb{N}^*,\ \sum_{t=1}^T \sup_\pi V_1^\pi(s_1;R^t) - \sum_{t=1}^T V_1^{\pi^t}(s_1;R^t) \le \mathcal{R}^\Pi(T,\delta)\right) \ge 1-\delta\ . \tag{1.64}$$

Now we state the main result of this section, which is adapted from Theorem 3.2.

Theorem 1.4 Under the previous assumption, with probability at least $1-\delta$, for all $T \ge 1$,

$$\min_{(h,s,a)\in\mathcal{X}} \frac{n_h^T(s,a)}{c_h(s,a)} \ge \frac{T}{\varphi^\star(c)} - \frac{1}{c_{min}}\left[\mathcal{R}^\lambda(T) + \mathcal{R}^\Pi(T,\delta/2) + \sqrt{T\log\left(\frac{4T^2}{\delta}\right)}\right]$$

The theorem above shows that the number of observations collected by Simplified Covgame grows at a nearly optimal rate. Indeed, $T/\varphi^\star(c)$ is the rate at which the expectation of the ratio $n_h^T(s,a)/c_h(s,a)$ would increase after T episodes if we had an oracle that provides the optimal π^{exp} solution to the lower bound (1.62) and played such policy. However, since we

do not have access to such an oracle, the observations increase at the optimal rate minus a $o(T)$ term, which represents the cost of learning how to explore the MDP.

Corollary 1.1 Using $\mathcal{A}^{\Pi}$ = UCBVI(Azar et al., 2017) [a] and $\mathcal{A}^{\lambda}$ = HEDGE (Freund & Schapire, 1997), we have $\mathcal{R}^{\Pi}(T,\delta) \leq 32\log(T+1)\sqrt{SAH^2T(\log(2SAH/\delta)+S)}$ and $\mathcal{R}^{\lambda}(T) \leq \sqrt{2T\log(SAH)} + \sqrt{\log(SAH)/8}$.

This yields that with probability at least $1-\delta$, Simplified CovGame instantiated with the algorithms above has sample complexity

$$\tau \leq 2\varphi^{\star}(c) + \widetilde{\mathcal{O}}\left(\left(\frac{\varphi^{\star}(c)}{c_{\min}}\right)^2 SH^2A(\log(1/\delta)+S)\right),$$

where $\widetilde{\mathcal{O}}$ hides logarithmic factors in $S, A, H, 1/c_{\min}$ and $\varphi^{\star}(c)$.

[a] A modified version of UCBVI to handle changing rewards, see appendix C of (Al-Marjani et al., 2023).

Proof sketch of Theorem 1.4

We denote by $n^T = [n_h^T(s,a)]_{h,s,a}$ the vector of the number of visits to all the triplets. For two vectors $x = [x_i]_i$ and $y = [y_i]_i$, $x/y := [x_i/y_i]_i$ is the entry-wise division of x by y. The first is structured in three steps. First, we relate the counts to the loss of the $\mathcal{A}^{\lambda}$:

$$\begin{aligned}
c_{\min}\min_{(h,s,a)\in\mathcal{X}}\frac{n_h^T(s,a)}{c_h(s,a)} &= \inf_{\lambda\in\Sigma_{\mathcal{X}}}\lambda\cdot(n^T/\widetilde{c}) && \text{(definitions of } \widetilde{c} \text{ and } \Sigma_{\mathcal{X}}\text{)}\\
&= \inf_{\lambda\in\Sigma_{\mathcal{X}}}\sum_{(h,s,a)\in\mathcal{X}}\lambda_h(s,a)\sum_{t=1}^{T}\frac{\mathbb{1}\left(s_h^t=s,a_h^t=a\right)}{\widetilde{c}_h(s,a)}\\
&= \inf_{\lambda\in\Sigma_{\mathcal{X}}}\sum_{t=1}^{T}\ell^t(\lambda) && \text{(definition of } \ell_t(\lambda)\text{)}\\
&\geq \sum_{t=1}^{T}\ell^t(\lambda^t) - \mathcal{R}^{\lambda}(T). && \text{(regret bound of } \mathcal{A}^{\lambda}\text{)}
\end{aligned}$$

Second, we go from the loss of $\mathcal{A}^{\lambda}$ to the optimal value function of $\mathcal{A}^{\Pi}$:

$$\begin{aligned}
\sum_{t=1}^{T}\ell^t(\lambda^t) &= \sum_{t=1}^{T}\sum_{h,s,a}\frac{\mathbb{1}((h,s,a)\in\mathcal{X})\lambda_h^t(s,a)}{\widetilde{c}_h(s,a)}\left(\mathbb{1}\left(s_h^t=s,a_h^t=a\right)\pm p_h^{\pi^t}(s,a)\right)\\
&\qquad\qquad\text{(definition of } \ell_t(\lambda^t)\text{)}\\
&= \sum_{t=1}^{T}\sum_{h,s,a}p_h^{\pi^t}(s,a)R_h^t(s,a) + \sum_{t=1}^{T}\sum_{h,s,a}R_h^t(s,a)\left(\mathbb{1}\left(s_h^t=s,a_h^t=a\right)-p_h^{\pi^t}(s,a)\right)\\
&= \sum_{t=1}^{T}V_1^{\pi^t}\left(s_1;R^t\right) + M_T && \text{(definition of } V_1^{\pi}(s_1;R) \text{ + martingale)}\\
&\geq \sup_{\pi}\sum_{t=1}^{T}V_1^{\pi}\left(s_1;R^t\right) - \mathcal{R}^{\Pi}(T,\delta/2) - \sqrt{T\log\left(\frac{4T^2}{\delta}\right)}.\\
&\qquad\qquad\text{(Regret of } \mathcal{A}^{\Pi} \text{ + Azuma-Hoeffding's inequality)}
\end{aligned}$$

Finally, we move from the optimal value function of $\mathcal{A}^{\Pi}$ to the lower bound (1.62).

$$\begin{aligned}
\sup_{\pi} \sum_{t=1}^{T} V_1^{\pi}\left(s_1; R^t\right) &= \sup_{\pi} \sum_{t=1}^{T} \sum_{h,s,a} p_h^{\pi}(s,a) \frac{\mathbb{1}((h,s,a) \in \mathcal{X}) \lambda_h^t(s,a)}{\widetilde{c}_h(s,a)} \\
&= T \sup_{\pi} \sum_{h,s,a} \left(p_h^{\pi}(s,a) \frac{\mathbb{1}((h,s,a) \in \mathcal{X})}{\widetilde{c}_h(s,a)} \right) \left(\frac{\sum_{t=1}^{T} \lambda_h^t(s,a)}{T} \right) \\
&\geq T \sup_{\pi} \min_{(h,s,a) \in \mathcal{X}} \frac{p_h^{\pi}(s,a)}{\widetilde{c}_h(s,a)} = c_{min} \frac{T}{\varphi^{\star}(c)}.
\end{aligned}$$

Wrapping up everything, we get

$$\begin{aligned}
& c_{\min} \min_{(h,s,a) \in \mathcal{X}} \frac{n_h^T(s,a)}{c_h(s,a)} \geq c_{min} \frac{T}{\varphi^{\star}(c)} - \mathcal{R}^{\lambda}(T) - \mathcal{R}^{\Pi}(T, \delta/2) - \sqrt{T \log\left(\frac{4T^2}{\delta}\right)} \\
\implies & \min_{(h,s,a) \in \mathcal{X}} \frac{n_h^T(s,a)}{c_h(s,a)} \geq \frac{T}{\varphi^{\star}(c)} - \frac{1}{c_{min}} \left[\mathcal{R}^{\lambda}(T) + \mathcal{R}^{\Pi}(T, \delta/2) + \sqrt{T \log\left(\frac{4T^2}{\delta}\right)} \right].
\end{aligned}$$

■

Remark 1.8 When the target function c is uniform: $c_h(s,a) = N\mathbb{1}((h,s,a) \in \mathcal{X})$, the sample complexity showcased in Corollary 1.1 is nearly optimal. By "near-optimal", we mean that when $N \to \infty$ the dominating term is $2\varphi^{\star}(c)$. Hence, in this regime, we are able to match the lower bound up to a factor of 2. However, if the target function is unbalanced, meaning that the ratio c_{max}/c_{min} is large, the second term in the bound above is no longer negligible and we can not claim to be near-optimal. We will explain in Chapter 3 how to improve Algorithm 6 in order to solve this issue. ■

1.6.4 Implicit policy eliminations for computationally-efficient approximate BPI

In Chapter 4, we will derive a problem-dependent lower bound for ε-BPI. We will also see that PEDEL, an ε-BPI algorithm proposed by (Wagenmaker & Jamieson, 2022) for the general case of linear MDPs, nearly matches our lower bound when we instantiate it for the tabular MDP setting. However, PEDEL has exponential time and memory complexities as it needs to enumerate the set of deterministic policies Π_{D} in order to eliminate the suboptimal ones. The structure of PEDEL is briefly sketched in Algorithm 7. **Notation:** We let $p_h^{\pi}(s,a) := \mathbb{P}_{\mathcal{M},\pi}(s_h = s, a_h = a)$ and $\widehat{p}_h^{\pi,k}(s,a) := \mathbb{P}_{\widehat{\mathcal{M}}_k,\pi}(s_h = s, a_h = a)$ where $\widehat{\mathcal{M}}_k$ is the empirical MDP constructed after k iterations of PEDEL. $\Omega(\mathcal{M}) := \{[p_h^{\pi}(s,a)]_{h \in [H], s \in \mathcal{S}, a \in \mathcal{A}} : \pi \in \Pi_{\mathrm{S}}\}$ denotes the set of all valid state-action distributions.

Algorithm 7 General structure of PEDEL

1: **Input:** precision ε, risk δ.
2: Initialize set of candidate policies $\Pi_0 \leftarrow \Pi_{\mathrm{D}}$
3: **for** $k = 0, 1, \ldots$ **do**
4: Run exploration procedure to collect $n_h^k(s,a)$ observations from each $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$ such that

$$\forall h \in [H],\ \max_{\pi \in \Pi_k} \sum_{s,a} \frac{\widehat{p}_h^{\pi,k}(s,a)^2}{n_h^k(s,a)} \leq \varepsilon_k. \tag{1.65}$$

5: Update the set of candidate policies

$$\Pi_{k+1} \leftarrow \Pi_k \setminus \left\{ \pi \in \Pi_k :\ \widehat{V}_1^{\pi} < \max_{\pi' \in \Pi_k} \widehat{V}_1^{\pi'} - 2^{1-k} \right\}$$

6: **If** $|\Pi_{k+1}| = 1$ **or** $2^{-k} \leq \varepsilon$**:**
7: Stop and return any $\widehat{\pi} \in \Pi_{k+1}$
8: **end if**
9: **end for**

At every iteration k, PEDEL keeps a set of candidate policies Π_k which is initialized as $\Pi_0 := \Pi_{\mathrm{D}}$. The exploration procedure aims to collect observations that will reduce the size of a confidence interval over the values of policies in Π^k below a certain threshold ε_k[11]. Then at the end of iteration k, the algorithm updates the set of candidate policies by removing those that are provably suboptimal. PEDEL stops when there remains only a single policy in the candidate set or it has reached a precision that is below ε. Hence, we see that needs $\Omega\big((SH)^A\big)$ operations and memory space in its exploration procedure to check whether the condition (1.65) holds and to eliminate suboptimal policies (line 5 of Algorithm 7). So how can we eliminate policies while keeping a polynomial time-memory complexity?

This is where our technique of *implicit policy eliminations* comes into the picture. We exploit two basic properties of MDPs. The first is that the value of any Markovian policy π is linear in its state-action distribution $[p_h^{\pi}(s,a)]_{h \in [H], s \in \mathcal{S}, a \in \mathcal{A}}$:

$$V_1^{\pi} = \sum_{h,s,a} p_h^{\pi}(s,a) r_h(s,a). \tag{1.66}$$

The second property is the fact that the set of state-action distributions $\Omega(\mathcal{M})$ is a *polytope* defined by linear constraints. Precisely, we know that (e.g., Puterman, 1994) that

$$\Omega(\mathcal{M}) = \left\{ \rho \in \mathbb{R}_+^{SAH} :\ \sum_{a \in \mathcal{A}} \rho_1(s,a) = 1,\ \sum_{a \in \mathcal{A}} \rho_h(s,a) = \sum_{(s',a')} \rho_{h-1}(s',a') p_{h-1}(s|s',a')\ \forall (h,s) \right\}.$$

Therefore, instead of performing operations over sets of policies we define sets of candidate state-action distributions $(\Omega_k)_{k \geq 1}$. The general idea is that to eliminate suboptimal policies, we take the set $\Omega(\widehat{\mathcal{M}}_k)$ and add to it a linear constraint of the shape

$$\sum_{h,s,a} \rho_h(s,a) r_h(s,a) > \sup_{\eta \in \Omega_k} \sum_{h,s,a} \eta_h(s,a) r_h(s,a) - 2^{1-k}. \tag{1.67}$$

[11] See (Wagenmaker & Jamieson, 2022) for the precise tuning of the sequence $(\varepsilon_k)_k$

This defines the set Ω_k. The advantages over line of Algorithm 7 are that: (1) computing the supremum in (1.67) can be done by solving a Linear Program (LP) for which there are several algorithms that run in polynomial time; (2) storing the constraint above only requires a memory space that is linear in SAH. As for the exploration procedure, observe that if we run CovGame with target function $c_h(s,a) = \sup_{\rho\in\Omega_k} \rho_h(s,a)/\varepsilon_k$, then with probability at least $1-\delta$ we would collect $(n_h^k(s,a))_{h,s,a}$ observations such that

$$\forall h \in [H], \quad \sup_{\rho\in\Omega_k} \sum_{s,a} \frac{\rho_h(s,a)^2}{n_h^k(s,a)} \leq \varepsilon_k \sup_{\rho\in\Omega_k} \sum_{s,a} \rho_h(s,a) = \varepsilon_k.$$

Thus, we have achieved a sufficient condition for (1.65) by reducing the size of the confidence interval over the values of policies that satisfy $[\widehat{p}_h^{\pi,k}(s,a)] \in \Omega_k$. Observe that such an operation can still be done in polynomial time since computing the targets amounts to solving another LP. In conclusion, we have implicitly eliminated the suboptimal policies such that $[\widehat{p}_h^{\pi,k}(s,a)]$ does not satisfy (1.67) using polynomial time and memory!

We give more details on the resulting ε-BPI algorithm in Chapter 4.

2. Asymptotic Navigation for Problem-Dependent Best Policy Identification

In this chapter, we present Navigate-and-Stop, an algorithm for exact Best Policy Identification (Section 1.4.1) that uses mixing properties of Markov Chains to converge to any allocation vector $\omega^\star(\mathcal{M})$. The contents of this chapter are based on the conference paper:

Aymen Al-Marjani, Aurélien Garivier, and Alexandre Proutiere. **Navigating to the Best Policy in Markov Decision Processes**. In *Advances in Neural Information Processing Systems* (NeurIPS), 34, 2021.

Contents

2.1 On the Optimization Objective and the Optimal Allocation

In Chapter 1, we assumed that the solution to the optimization problem in (1.32) is unique. In reality, we do not know whether this is truly the case. Indeed, let us define the quantity

$$T(\mathcal{M},\omega) := \left(\inf_{\mathcal{M}'\in \mathrm{Alt}(\mathcal{M})} \sum_{s,a} \omega_{sa} \mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a)\right)^{-1}. \tag{2.1}$$

Recall that we proved in Theorem 1.1 a lower bound on the sample complexity of BPI, which is written as

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_{\mathbb{A},\mathcal{M}}[\tau]}{\log(1/\delta)} \geq T^\star(\mathcal{M}) = \inf_{\omega\in\Omega(\mathcal{M})} T(\mathcal{M},\omega), \tag{2.2}$$

where the equality can be checked easily from the definitions of $T(\mathcal{M},\omega)$ and $T^\star(\mathcal{M})$. Then we showed in (Al-Marjani & Proutiere, 2021) that even for a "toy" MDP of 2 states and 2 actions, the minimization problem in (2.1) is not convex. As such, it is difficult to obtain theoretical guarantees on the uniqueness of its solution, let alone come up with a tractable method to compute it. In the same paper, we also provided the following tractable upper bound on $T(\mathcal{M},\omega)$. Some notations are due before stating the result.

Notation We recall that $\Sigma := \{\omega \in \mathbb{R}_+^{SA} : \sum_{i=1}^{SA} \omega_i = 1\}$ refers to the simplex of dimension $SA-1$, while

$$\Omega(\mathcal{M}) := \{\omega\in\Sigma : \forall s\in\mathcal{S}, \sum_{a\in\mathcal{A}} \omega_{sa} = \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} p(s|s',a')\omega_{s'a'}\}$$

denotes the set of allocation vectors that satisfy the navigation constraints. $\mathfrak{M}_{\star,1}$ denotes the class of discounted MDPs with a unique optimal policy. $\Delta(s,a) := V^\star(s) - Q^\star(s,a)$ is the suboptimality gap of state-action pair (s,a). We use the shorthand $\pi^\star$ to denote the unique optimal policy of $\mathcal{M}$. $\Delta_{\min}(\mathcal{M}) := \min_{a\neq\pi^\star(s)} \Delta(s,a)$ denotes the minimum positive suboptimality gap. For a real-valued function $f : \mathcal{X}\to\mathbb{R}$, $\mathrm{sp}(f) := \sup_{x,x'\in\mathcal{X}} |f(x') - f(x)|$ denotes the span of f.

Lemma 2.1 — Theorem 1, (Al-Marjani & Proutiere, 2021). For all vectors $\omega\in\Sigma$ and MDPs $\mathcal{M}\in\mathfrak{M}_{\star,1}$, it holds that $T(\mathcal{M},\omega)\leq U(\mathcal{M},\omega)$, where[a]

$$U(\mathcal{M},\omega) := \max_{(s,a):a\neq\pi^\star(s)} \frac{H_{sa}}{\omega_{sa}} + \frac{H^\star}{S\min_s \omega_{s,\pi^\star(s)}}, \tag{2.3}$$

$$\text{and} \quad \begin{cases} H_{sa} := \dfrac{2}{\Delta(s,a)^2} + \max\left(\dfrac{16\mathrm{Var}_{s'\sim p(s,a)}[V^\star_{\mathcal{M}}(s')]}{\Delta(s,a)^2}, \dfrac{6\,\mathrm{sp}(V^\star_{\mathcal{M}})^{4/3}}{\Delta(s,a)^{4/3}}\right), \\ H^\star := \dfrac{2S}{[\Delta_{\min}(\mathcal{M})(1-\gamma)]^2} + \mathcal{O}\left(\dfrac{S}{\Delta_{\min}(\mathcal{M})^2(1-\gamma)^3}\right). \end{cases} \tag{2.4}$$

[a] The exact definition of $H^\star$ is given in Section 2.6.

Using $U(\mathcal{M},\omega)$, we obtain the following upper bound on the characteristic time in (2.2):

$$T^\star(\mathcal{M}) \leq U^\star(\mathcal{M}) := \inf_{\omega\in\Omega(\mathcal{M})} U(\mathcal{M},\omega). \tag{2.5}$$

The advantages of this upper bound $U^\star(\mathcal{M})$ are that:

1. $U^\star(\mathcal{M})$ is still a problem-specific quantity as it depends on the gaps and variances of the value function in $\mathcal{M}$.
2. The function $\omega \mapsto U(\mathcal{M}, \omega)$ is strictly convex and the feasible set in (2.5) is also convex. Therefore there is a unique allocation vector that solves (2.5).
3. Since the optimization problem in (2.5) has a convex objective and convex constraints, we can easily compute its solution using either the Franke-Wolfe method or the projected subgradient-descent algorithm.

Definition 2.1 For the rest of this chapter, we shall define the optimal allocation vector as the one that solves (2.5),

$$\omega^\star(\mathcal{M}) := \underset{\omega \in \Omega(\mathcal{M})}{\arg\min}\, U(\mathcal{M}, \omega). \tag{2.6}$$

Remark 2.1 One can easily check that $\omega^\star_{sa}(\mathcal{M}) > 0$ for all state-action pairs (s, a). Indeed, from the definition of $U(\mathcal{M}, \omega)$, the objective function of an allocation vector ω that has a null component is infinite. Therefore, such an allocation cannot be optimal. ■

Remark 2.2 While our algorithm's design implements this particular choice of an allocation vector, the results that we will present can be applied in a straightforward fashion if (i) the solution to (2.2) is unique; (ii) one assumes access to an optimization oracle that solves that problem. ■

2.2 C-Navigation: A Sampling Rule for Asymptotic Optimality

We introduce a few notations to simplify the presentation. Any stationary Markov policy π induces a finite Markov chain on $\mathcal{S} \times \mathcal{A}$ whose transition matrix is defined by $P_\pi\big((s,a),\ (s',a')\big) := p_{\mathcal{M}}(s'|s,a)\pi(a'|s')$. It also induces a Markov chain on the state space $\mathcal{S}$ whose transition matrix is given by $\widetilde{P}_\pi(s,s') := \sum_{a\in\mathcal{A}} \pi(a|s) p_{\mathcal{M}}(s'|s,a)$. With some abuse of notation, we will use P_π to refer to both Markov chains. P_π^n denotes the n-th power of P_π. A standard result in Markov chain theory states that P_π^n is the transition matrix corresponding to n-th step Markov chain. We denote by π_u the uniform policy, i.e., $\pi_u(a|s) = 1/A$ for all pairs (s,a). For a pair of policies π_1 and π_2, the *mixture policy* with parameter ε is defined through $\overline{\pi}(a|s) := \varepsilon\pi_1(a|s) + (1-\varepsilon)\pi_2(a|s)$ for all (s,a). In that case, we will simply write $\overline{\pi} := \varepsilon\pi_1 + (1-\varepsilon)\pi_2$. Finally, we define the vector of visitation-frequencies at time t, $\mathbf{N}(t)/t := \big(N_{sa}(t)/t\big)_{(s,a)\in\mathcal{S}\times\mathcal{A}}$.

Before we proceed, we need to make the following assumptions.

Assumption 2.1 We assume that $\mathcal{M}$ is *communicating*, i.e., we can reach any state s' starting from any other state s. This means that for all $(s,s') \in \mathcal{S}$, there exists a deterministic Markovian policy $\pi \in \Pi_{\mathrm{D}}$ and an integer $t \geq 1$ such that

$$P^t(s,s') = \mathbb{P}_{\mathcal{M},\pi}(s_t = s' | s_1 = s) > 0, \tag{2.7}$$

where $\mathbb{P}_{\mathcal{M},\pi}$ is the probability distribution of trajectories induced by playing π in $\mathcal{M}$.

We restrict our attention to the case where $\mathcal{M}$ is communicating, for otherwise, there would be a non-zero probability that the algorithm enters a set of states from which there is no possible comeback. In this case, it becomes impossible to identify the optimal policy.

Assumption 2.2 P_{π_u} is aperiodic.

This assumption is mild as it is enough to have only one state $\tilde{s}$ and one action $\tilde{a}$ such that $P_{\mathcal{M}}(\tilde{s}|\tilde{s},\tilde{a}) > 0$ for it to be satisfied. Furthermore, Assumptions 2.1 and 2.2 combined imply that P_{π_u} is ergodic (because it is irreducible and aperiodic). This is still less restrictive than the *"$\mathcal{M}$ is ergodic"* assumption which is ubiquitous in RL literature (Burnetas & Katehakis, 1997; Tarbouriech & Lazaric, 2019; Pesquerel & Maillard, 2022). Indeed, assuming that the MDP is ergodic means that the Markov chains of *all* policies are ergodic.

2.2.1 Building-up the intuition

In contrast with the settings of MABs and MDPs with a generative model where one could converge to any allocation vector in Σ through C-tracking, see (Garivier & Kaufmann, 2016; Al-Marjani & Proutiere, 2021), here we face the *challenge of navigation*. Namely, the agent can only choose a *sequence of actions* $(a_t)_{t\geq 1}$ and follow the resulting trajectory whose law is determined by the transition kernel: $s_{t+1} \sim p_{\mathcal{M}}(.|s_t, a_t)$. Therefore, one might wonder whether the convergence to the optimal allocation can be achieved by following a simple policy. A natural candidate is the *oracle policy* defined by

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \quad \pi_{\omega^\star}(\mathcal{M})(a|s) := \frac{\omega^\star_{sa}(\mathcal{M})}{\sum_{b\in\mathcal{A}} \omega^\star_{sb}(\mathcal{M})}. \tag{2.8}$$

We will use $\pi_{\omega^\star}(\mathcal{M})$ to denote the oracle policy of $\mathcal{M}$ and $\pi_{\omega^\star}$ whenever the MDP under consideration is clear from the context. The oracle policy satisfies the following property.

Proposition 2.1 $\omega^\star(\mathcal{M})$ is the unique stationary distribution of the Markov chain whose transition matrix is $P_{\pi_{\omega^\star}}$.

Proof. It is immediate to check that $P_{\pi_{\omega^\star}}\omega^\star(\mathcal{M}) = \omega^\star(\mathcal{M})$ using the fact that $\omega^\star(\mathcal{M}) \in \Omega(\mathcal{M})$. Hence $\omega^\star(\mathcal{M})$ is a stationary distribution of $P_{\pi_{\omega^\star}}$. The uniqueness is guaranteed by the irreducibility of $P_{\pi_{\omega^\star}}$, see Proposition 1.7 in (Levin et al., 2006) for instance. Indeed, a direct consequence of Assumption 2.1 is that for every couple of state-action pairs $\big((s,a),(s',a')\big) \in (\mathcal{S}\times\mathcal{A})^2$, there exists $\pi \in \Pi_{\mathrm{D}}$ and $t \geq 1$ such that $P^t_\pi\big((s,a),\ (s',a')\big) = \mathbb{P}_{\mathcal{M},\pi}(s_t = s', a_t = a | s_1 = s, a_t = a) > 0$. By taking $\eta = \min_{s\in\mathcal{S},a\in\mathcal{A}} \pi_{\omega^\star}(a|s)$, we have that $\eta > 0$ (see Remark 2.1) and for all $\big((s,a),(s',a')\big) \in (\mathcal{S}\times\mathcal{A})^2$,

$$P_{\pi_{\omega^\star}}\big((s,a),\ (s',a')\big) = p_{\mathcal{M}}(s'|s,a)\pi_{\omega^\star}(a'|s') \geq \eta p_{\mathcal{M}}(s'|s,a)\pi(a'|s') = \eta P_\pi\big((s,a),\ (s',a')\big).$$

Therefore it also holds that $P^t_{\pi_{\omega^\star}}\big((s,a),\ (s',a')\big) > 0$, which mean that $P_{\pi_{\omega^\star}}$ is irreducible. ■

$\pi_{\omega^\star}(\mathcal{M})$ is the "target" policy that we would like to play since, by the Ergodic theorem (Theorem 4.16 in (Levin et al., 2006)), executing it guarantees convergence of the visitation-frequencies $\mathbf{N}(t)/t$ to the stationary distribution $\omega^\star(\mathcal{M})$. However, because the rewards and transitions of $\mathcal{M}$ are unknown to the algorithm, so is $\pi_{\omega^\star}(\mathcal{M})$. We circumvent this issue by using the oracle policy for the empirical MDP $\widehat{\mathcal{M}}_t$ whose reward function and transition kernel are the Maximum Likelihood Estimate (MLE) of $r_{\mathcal{M}}$ and $p_{\mathcal{M}}$. Provided that the MLEs are consistent $\widehat{\mathcal{M}}_t \underset{t\to\infty}{\longrightarrow} \mathcal{M}$, we can hope that using $\pi_{\omega^\star}(\widehat{\mathcal{M}}_t)$ for exploration will lead to the same asymptotic results than if we had used $\pi_{\omega^\star}(\mathcal{M})$ instead. To achieve the previous requirement, we *force exploration* by playing a mixture with the uniform policy. This ensures that all actions in all states are played sufficiently enough so that $N_{sa}(t) \underset{t\to\infty}{\longrightarrow} \infty$ for all (s,a).

Definition 2.2 — C-Navigation. Given a decreasing sequence of mixture parameters $(\varepsilon_t)_{t\geq 1}$ and a sequence of empirical estimates $\left(\widehat{\mathcal{M}}_t\right)_{t\geq 1}$, the C-Navigation sampling rule plays an action $a_t \sim \pi_t(.|s_t)$ where

$$\pi_t := \varepsilon_t \pi_u + (1-\varepsilon_t)\frac{\sum_{j=0}^{t-1} \pi_{\omega^\star}(\widehat{\mathcal{M}}_j)}{t}, \quad \forall t \geq 1. \tag{2.9}$$

Observe that we explore using a Cesàro-mean of oracle policies instead of the current estimate of the oracle policy. This ensures the stability of the non-homogeneous Markov chain, a property that will be crucial for our convergence guarantees.

2.2.2 Tuning the mixture parameters

We begin by defining an important parameter that describes how well-connected are the states through the transition kernel of $\mathcal{M}$.

Definition 2.3 We define the communication parameter m as the maximum number of transitions that are needed to travel between any pair of states in $\mathcal{M}$ with positive probability:

$$m := \max_{(s,s')\in\mathcal{S}^2} \min\{n \geq 1 : \ \exists \pi : \mathcal{S} \to \mathcal{A}, P_\pi^n(s,s') > 0\}.$$

Remark 2.3 Note that if it takes m steps to move between any pair of states (s,s') with non-zero probability, then we can move between any pair of state-actions $((s,a),(s',a'))$ in at most $m+1$ steps. Indeed, by playing action a at s, we move to some intermediate state $\widetilde{s}$. From there, we have at most m steps to reach s' and play a'. ■

If m is small, e.g. $m = 1$, then all states are reachable from any other state within a one-step transition. As a result, it takes only a small effort to explore all states and actions. On the other hand, m can be as large as $S-1$ in the worst case[1]. In such a scenario, the navigation challenge becomes harder since the agent may need to go through several "lucky" transitions to cover all the states in a short time. Given these observations, it is only natural that m quantifies how much forced-exploration the algorithm must perform. Our next result is a lemma showing a sufficient condition on the sequence $(\varepsilon_t)_{t\geq 1}$ to guarantee forced exploration with high probability.

2.2.2.1 Sufficient conditions

Lemma 2.2 — High probability forced exploration. Denote by $\tau_k(s,a)$ the k-th time that the algorithm visits the state-action pair (s,a). Suppose that the exploration rate of C-Navigation satisfies $\varepsilon_t \geq t^{-\frac{1}{2(m+1)}}$ for all $t \geq 1$. Then there exists a parameter $\eta > 0$ that only depends on $\mathcal{M}$ such that

$$\forall \alpha \in (0,1), \quad \mathbb{P}\left(\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \ \forall k \geq 1, \ \tau_k(s,a) \leq \lambda_\alpha k^4\right) \geq 1-\alpha,$$

where $\lambda_\alpha := \frac{(m+1)^2}{\eta^2}\log^2(1+\frac{SA}{\alpha})$.

By inverting the inequality on the hitting times above, we immediately get the following Corollary.

[1] $S-1$ corresponds to the length of the shortest path between any pair of nodes in a graph whose nodes are the states of $\mathcal{M}$ and where all edges have weight one.

Corollary 2.1 Denote by $N_{sa}(t)$ the number of times the agent visits state-action (s,a) up to and including time step t. Then under the same condition of the lemma above we have

$$\forall \alpha \in (0,1), \quad \mathbb{P}\left(\forall (s,a) \in \mathcal{S} \times \mathcal{A},\ \forall t \geq 1,\ N_{sa}(t) \geq \left(\frac{t}{\lambda_\alpha}\right)^{1/4} - 1\right) \geq 1 - \alpha.$$

Remark 2.4 When the communication parameter m is unknown to the learner, one can always replace it with its worst-case value $m_{\max} = S - 1$. However, when prior knowledge is available, using a faster-decreasing sequence $\varepsilon_t = t^{-\frac{1}{2(m+1)}}$ instead of $t^{-\frac{1}{2S}}$ can be useful to accelerate convergence, especially when the states of $\mathcal{M}$ are "densely connected", i.e., $m \ll S - 1$. ∎

Proof. For the sake of simplicity, we let $P_t := P_{\pi_t}$ be the transition matrix induced by the policy that C-Navigation plays at time step t. We also denote a state-action pair by z instead of (s,a). Let f be some increasing function such that $f(\mathbb{N}) \subset \mathbb{N}$ and $f(0) = 0$ and define the event $\mathcal{E} := \big(\forall z \in \mathcal{S} \times \mathcal{A},\ \forall k \geq 1,\ \tau_k(z) \leq f(k)\big)$. We will prove the following more general result:

$$\mathbb{P}(\mathcal{E}^c) \leq SA \sum_{k=1}^{\infty} \prod_{j=0}^{\lfloor \frac{f(k)-f(k-1)-1}{m+1} \rfloor - 1} \left[1 - \eta \prod_{l=1}^{m+1} \varepsilon_{f(k-1)+(m+1)j+l}\right], \tag{2.10}$$

where η is a constant depending on $\mathcal{M}$. Then we will tune $f(k)$ and ε_t so that the right-hand side is less than α. First, observe that

$$\mathcal{E}^c = \bigcup_{z \in \mathcal{S}\times\mathcal{A}} \bigcup_{k=1}^{\infty} \left(\tau_k(z) > f(k) \text{ and } \forall j \leq k-1,\ \tau_j(z) \leq f(j)\right).$$

Using the decomposition above, we upper bound the probability of $\mathcal{E}^c$ by the sum of probabilities for $k \geq 1$ that the k-th excursion from and back to z takes too long:

$$\begin{aligned}
\mathbb{P}(\mathcal{E}^c) &\leq \sum_{z \in \mathcal{S}\times\mathcal{A}} \left[\mathbb{P}(\tau_1(z) > f(1)) + \sum_{k=2}^{\infty} \mathbb{P}\left(\tau_k(z) > f(k) \text{ and } \forall j \leq k-1,\ \tau_j(z) \leq f(j)\right)\right] \\
&\leq \sum_{z \in \mathcal{S}\times\mathcal{A}} \left[\mathbb{P}(\tau_1(z) > f(1)) + \sum_{k=2}^{\infty} \mathbb{P}\left(\tau_k(z) > f(k) \text{ and } \tau_{k-1}(z) \leq f(k-1)\right)\right] \\
&\leq \sum_{z \in \mathcal{S}\times\mathcal{A}} \left[\mathbb{P}(\tau_1(z) > f(1)) + \sum_{k=2}^{\infty} \mathbb{P}\left(\tau_k(z) - \tau_{k-1}(z) > f(k) - f(k-1),\ \tau_{k-1}(z) \leq f(k-1)\right)\right] \\
&\leq \sum_{z \in \mathcal{S}\times\mathcal{A}} \Bigg[\mathbb{P}(\tau_1(z) > f(1)) \\
&\qquad + \sum_{k=2}^{\infty} \sum_{n=1}^{f(k-1)} \mathbb{P}\big(\tau_k(z) - \tau_{k-1}(z) > f(k) - f(k-1) \big| \tau_{k-1}(z) = n\big)\mathbb{P}(\tau_0(z) = n)\Bigg] \\
&= \sum_{z \in \mathcal{S}\times\mathcal{A}} \left[a_1(z) + \sum_{k=2}^{\infty} \sum_{n=1}^{f(k-1)} a_{k,n}(z)\mathbb{P}(\tau_{k-1}(s) = n)\right],
\end{aligned} \tag{2.11}$$

where

$$a_1(z) := \mathbb{P}(\tau_1(z) > f(1)) \ ,$$
$$\forall k \geq 2 \ \forall n \in [|1, f(k-1)|], \ a_{k,n}(z) := \mathbb{P}\big(\tau_k(z) - \tau_{k-1}(z) > f(k) - f(k-1) \big| \tau_{k-1}(z) = n\big) \ .$$

We will now prove an upper bound on $a_{k,n}(z)$ for a fixed $z \in \mathcal{S} \times \mathcal{A}$ and $k \geq 1$.

1) Upper bounding the probability that an excursion takes too long: Let us rewrite P_t as

$$P_t = \left(\begin{array}{c|c} Q_t(z) & [P_t(z', z)]_{z' \neq z} \\ \hline [P_t(z, z')]_{z' \neq z}^T & P_t(z, z) \end{array}\right) \ ,$$

so that state-action z corresponds to the last row and last column and $Q_t(z) := [P_t(z', z")]_{z', z" \in \mathcal{S} \times \mathcal{A} \setminus \{z\}}$. Further let $p_t(z', \neg z) := [P_t(z', z")]_{z" \neq z}$ denote the vector of probabilities of transitions at time t from z' to states $z"$ different from z. Using a simple recurrence on N, one can prove that for all $k, N, n \geq 1$ we have:

$$\mathbb{P}\Big(\tau_k(z) - \tau_{k-1}(z) > N \Big| \tau_{k-1}(z) = n\Big) = p_n(z, \neg z)^\top \Big(\prod_{j=n+1}^{n+N-1} Q_j(z)\Big) \mathbb{1} \ . \tag{2.12}$$

Observe that the matrices $(Q_j)_j$ are sub-stochastic, i.e, they each have at least one line whose sum is strictly smaller than 1 (the line corresponds to the state-action pair from which one can move to z within one transition using the uniform policy). Using Lemma 2.7, there exists $\eta > 0$ (that only depends on $\mathcal{M}$) such that for all $n \geq 1$ and all sequences $(\pi_t)_{t \geq 1}$ that satisfy $\pi_t \geq \varepsilon_t \pi_u$ we have

$$\Big\| \prod_{l=n+1}^{n+m+1} Q_l(z) \Big\|_\infty \leq 1 - \eta \prod_{l=n+1}^{n+m+1} \varepsilon_l \ . \tag{2.13}$$

Therefore using (2.12) for $N = f(k) - f(k-1)$ and breaking the matrix product into smaller product terms of $(m+1)$ matrices, we get for $k \geq 2$

$$\begin{aligned}
a_{k,n}(z) &= \mathbb{P}\Big(\tau_k(z) - \tau_{k-1}(z) > f(k) - f(k-1) \big| \ \tau_{k-1}(z) = n\Big) \\
&\overset{(a)}{=} \mathbb{E}\Big[\mathbb{P}\Big(\tau_k(s) - \tau_{k-1}(s) > f(k) - f(k-1) \Big| \tau_{k-1}(z) = n, (\pi_t)_{t \geq 1}\Big)\Big] \\
&= \mathbb{E}\Big[p_n(z, \neg z)^\intercal \Big(\prod_{j=n+1}^{n+f(k)-f(k-1)-1} Q_j(z)\Big) \mathbb{1}\Big] \\
&\overset{(b)}{\leq} \Big\| \prod_{l=n+1}^{n+f(k)-f(k-1)-1} Q_l(z) \Big\|_\infty \\
&\leq \Big\| \prod_{l=(m+1)\lfloor \frac{f(k)-f(k-1)-1}{m+1} \rfloor + 1}^{f(k)-f(k-1)-1} Q_{n+l}(z) \Big\|_\infty \times \prod_{j=0}^{\lfloor \frac{f(k)-f(k-1)-1}{m+1} \rfloor - 1} \Big\| \prod_{l=1}^{m+1} Q_{n+(m+1)j+l}(z) \Big\|_\infty \\
&\overset{(c)}{\leq} \prod_{j=0}^{\lfloor \frac{f(k)-f(k-1)-1}{m+1} \rfloor - 1} \Big[1 - \eta \prod_{l=1}^{m+1} \varepsilon_{n+(m+1)j+l}\Big] \\
&\overset{(d)}{\leq} \prod_{j=0}^{\lfloor \frac{f(k)-f(k-1)-1}{m+1} \rfloor - 1} \Big[1 - \eta \prod_{l=1}^{m+1} \varepsilon_{f(k-1)+(m+1)j+l}\Big] := b_k \ ,
\end{aligned} \tag{2.14}$$

where (a) uses the law of total expectation, (b) uses that $\|p_n(z,\neg z)\|_1 \leq 1$, (c) uses the fact that the matrices Q are substochastic while (d) is due to the facts that $n \leq f(k-1)$ and $t \mapsto \varepsilon_t$ is decreasing. Similarly, one can prove that

$$a_1(z) \leq \prod_{j=0}^{\lfloor \frac{f(1)-1}{m+1} \rfloor - 1} \left[1 - \eta \prod_{l=1}^{m+1} \varepsilon_{(m+1)j+l}\right] = \prod_{j=0}^{\lfloor \frac{f(1)-f(0)-1}{m+1} \rfloor - 1} \left[1 - \eta \prod_{l=1}^{m+1} \varepsilon_{f(0)+(m+1)j+l}\right] := b_1 \,, \tag{2.15}$$

where we used the fact that $f(0) = 0$. Now we only have to tune $f(k)$ and ε_t so that $\sum_{k=1}^{\infty} b_k < \frac{\alpha}{SA}$ and conclude using (2.11), (2.14) and (2.15).

2) Tuning f and the exploration rate: Since the sequence $(\varepsilon_t)_{t\geq 1}$ is decreasing we have:

$$\begin{aligned} b_k &= \prod_{j=0}^{\lfloor \frac{f(k)-f(k-1)-1}{m+1} \rfloor - 1} \left[1 - \eta \prod_{l=1}^{m+1} \varepsilon_{f(k-1)+(m+1)j+l}\right] \\ &\leq \prod_{j=0}^{\lfloor \frac{f(k)-f(k-1)-1}{m+1} \rfloor - 1} \left[1 - \eta \big(\varepsilon_{f(k-1)+(m+1)j+S}\big)^{m+1}\right] \\ &\leq \left[1 - \eta \big(\varepsilon_{f(k)}\big)^{m+1}\right]^{\lfloor \frac{f(k)-f(k-1)-1}{m+1} \rfloor} . \end{aligned}$$

For $f(k) = \lambda.k^4$ where $\lambda \in \mathbb{N}^\star$ and $\varepsilon_t = t^{-\frac{1}{2(m+1)}}$ we have: $\lfloor \frac{f(k)-f(k-1)-1}{m+1} \rfloor \geq \frac{\lambda k^3}{(m+1)}$ and $\big(\varepsilon_{f(k)}\big)^{m+1} = \frac{1}{\sqrt{\lambda}k^2}$, implying:

$$b_k \leq \left[1 - \frac{\eta}{\sqrt{\lambda}k^2}\right]^{\frac{\lambda k^3}{(m+1)}} \leq \exp\left(\frac{-\lambda k^3 \eta}{(m+1)\sqrt{\lambda}k^2}\right) = \exp\left(-\frac{\lambda^{1/2} k\eta}{m+1}\right) .$$

Summing the last inequality, along with (2.11), (2.14) and (2.15) we get:

$$\mathbb{P}(\mathcal{E}^c) \leq SA \sum_{k=1}^{\infty} b_k \leq SA \sum_{k=1}^{\infty} \exp\left(-\frac{\lambda^{1/2} k\eta}{m+1}\right) = \frac{SA \exp\left(-\frac{\lambda^{1/2}\eta}{m+1}\right)}{1 - \exp\left(-\frac{\lambda^{1/2}\eta}{m+1}\right)} := g(\lambda) \,.$$

For $\lambda_\alpha := \frac{(m+1)^2}{\eta^2} \log^2(1 + \frac{SA}{\alpha})$, we have $g(\lambda_\alpha) = \alpha$, which gives the desired result. ■

We complement the previous result with another lemma which shows that we can use a slightly smaller rate of exploration if we only want to establish almost-sure forced exploration.

Lemma 2.3 C-Navigation with any decreasing sequence $(\varepsilon_t)_{t\geq 1}$ such that $\forall t \geq 1,\ \varepsilon_t \geq t^{-\frac{1}{m+1}}$ satisfies

$$\mathbb{P}_{\mathcal{M},\mathbb{A}}\big(\forall (s,a) \in \mathcal{S} \times \mathcal{A},\ \lim_{t\to\infty} N_{sa}(t) = \infty\big) = 1 \,.$$

Proof. Consider the event $\mathcal{E} := \big(\exists z \in \mathcal{S} \times \mathcal{A},\ \exists M > 0,\ \forall t \geq 1,\ N_z(t) < M\big)$. Observe that $\mathcal{E} = \bigcup_{z\in\mathcal{S}\times\mathcal{A}} \mathcal{E}_z$, where for $z \in \mathcal{S} \times \mathcal{A}$, $\mathcal{E}_z := \big(\exists M > 0,\ \forall t \geq 1,\ N_z(t) < M\big)$. We will prove that $\mathbb{P}(\mathcal{E}_{z'}) = 0$ for all z', which implies the desired result. From Remark 2.3, we have

$$\forall (z,z') \in (\mathcal{S} \times \mathcal{A})^2,\ \exists r \in [|1, m+1|],\ \exists \pi \in \Pi_{\mathrm{D}},\ P_\pi^r(z,z') > 0, \tag{2.16}$$

where P_π^r is the r-th power of the transition matrix induced by policy π. Therefore,

$$\eta := \min_{z,z'} \max_{\substack{1\le r\le m+1 \\ \pi\in\Pi_{\mathrm{D}}}} P_\pi^r(z,z')$$

is positive. Fix $z \in \mathcal{S}\times\mathcal{A}$ and let π, r be a policy and an integer satisfying the property (2.16) above for the pair (z,z'). Observe that

$$P_t \ge \varepsilon_t P_{\pi_u} \ge \frac{\varepsilon_t}{A} P_\pi,$$

where the matrix inequality is entry-wise. Now define the stopping times $(\tau_k(z))_{k\ge1}$ where the agent reaches state-action z for the k-th time[2]. Also, denote by X_t the state-action pair at the t-th step of the Markov Chain. Then

$$\begin{aligned}
\mathbb{P}\big(\mathcal{E}_{z'}|\ (\pi_t)_{t\ge1},\ (\tau_k(z))_{k\ge1}\big) &\le \mathbb{P}\big(\exists N\ge1, \forall k\ge N, X_{\tau_k(z)+r}\neq z' \mid (\pi_t)_{t\ge1},\ (\tau_k(z))_{k\ge1}\big) \\
&\overset{(a)}{\le} \sum_{N=1}^{\infty}\prod_{k=N}^{\infty} \mathbb{P}\big(X_{\tau_k(z)+r}\neq z' \mid \tau_k(z),\ (\pi_t)_{t\in[|\tau_k(z)+1,\tau_k(z)+r|]}\big) \\
&= \sum_{N=1}^{\infty}\prod_{k=N}^{\infty}\left[1-\left(\prod_{t=\tau_k(z)+1}^{\tau_k(z)+r} P_{\pi_t}\right)(z,z')\right] \\
&\le \sum_{N=1}^{\infty}\prod_{k=N}^{\infty}\left[1-\left(\prod_{t=\tau_k(z)+1}^{\tau_k(z)+r} \frac{\varepsilon_t}{A}P_\pi\right)(z,z')\right] \\
&\le \sum_{N=1}^{\infty}\prod_{k=N}^{\infty}\left[1-\frac{\eta}{A^r}\prod_{t=\tau_k(z)+1}^{\tau_k(z)+r}\varepsilon_t\right] \\
&\overset{(b)}{\le} \sum_{N=1}^{\infty}\prod_{k=N}^{\infty}\left[1-\frac{\eta}{A^{m+1}}\prod_{t=\tau_k(z)+1}^{\tau_k(z)+m+1}\varepsilon_t\right] n
\end{aligned}$$

where (a) comes from a union bound and the strong Markov property[3] and (b) comes from the fact that $r \le m+1$ and $\varepsilon_t \le 1$. Now observe that the inequality above holds for all realizations of the sequences $(\pi_t)_{t\ge1}$. Therefore, integrating that inequality over all possible sequences of policies yields:

$$\forall z\in\mathcal{S}\times\mathcal{A},\ \mathbb{P}\big(\mathcal{E}_{z'}|\ (\tau_k(z))_{k\ge1}\big) \le \sum_{N=1}^{\infty}\prod_{k=N}^{\infty}\left[1-\frac{\eta}{A^{m+1}}\prod_{t=\tau_k(z)+1}^{\tau_k(z)+m+1}\varepsilon_t\right].$$

We can already see that if state-action z is visited "frequently enough" ($\tau_k(z)\sim c.k$ for some constant c) then the right-hand side above will be zero. Since we know that a least one state-action z is visited frequently enough, we consider the product over all state-action pairs z of the probabilities above:

$$\begin{aligned}
\prod_{z\in\mathcal{S}\times\mathcal{A}}\mathbb{P}\Big(\mathcal{E}_{z'}|\ (\tau_k(z))_{k\ge1}\Big) &\le \sum_{(N_1,\dots,N_{SA})\in(\mathbb{N}^\star)^{SA}}\ \prod_{z\in\mathcal{S}\times\mathcal{A}}\prod_{k=N_z}^{\infty}\left[1-\frac{\eta}{A^{m+1}}\prod_{t=\tau_k(z)+1}^{\tau_k(z)+m+1}\varepsilon_t\right] \qquad (2.17)\\
&:= \sum_{(N_1,\dots,N_{SA})} a_{(N_1,\dots,N_{SA})}\ .
\end{aligned}$$

[2] We restrict our attention to departure state-action pairs z that are visited infinitely often. Such pairs always exist, therefore $\tau_k(z)$ is well defined.

[3] This property is sometimes referred to as: "Markov Chains start afresh after stopping times."

We will now show that $a_{(N_1,\dots,N_{SA})} = 0$ for all tuples $(N_1,\dots,N_{SA})$:

$$\begin{aligned} a_{(N_1,\dots,N_{SA})} &\leq \prod_{z\in\mathcal{S}\times\mathcal{A}} \prod_{k=\max_z N_z}^{\infty} \left[1 - \frac{\eta}{A^{m+1}} \prod_{t=\tau_k(z)+1}^{\tau_k(z)+m+1} \varepsilon_t\right] \\ &= \prod_{k=\max_z N_z}^{\infty} \prod_{z\in\mathcal{S}\times\mathcal{A}} \left[1 - \frac{\eta}{A^{m+1}} \prod_{t=\tau_k(z)+1}^{\tau_k(z)+m+1} \varepsilon_t\right] . \end{aligned}$$

Now by the pigeon-hole principle, for all $k \geq 1$ there exists $z_k \in \mathcal{S}\times\mathcal{A}$ such that $\tau_k(z_k) \leq SAk$, i.e., at least one state-action has been visited k times before time step SAk. For that particular choice of z_k and since $(\varepsilon_t)_{t\geq 1}$ is decreasing, we get

$$\begin{aligned} a_{(N_1,\dots,N_{SA})} &\leq \prod_{k=\max_z N_z}^{\infty} \left[1 - \frac{\eta}{A^{m+1}} \prod_{t=\tau_k(z_k)+1}^{\tau_k(z_k)+m+1} \varepsilon_t\right] \\ &\leq \prod_{k=\max_z N_z}^{\infty} \left[1 - \frac{\eta}{A^{m+1}} \prod_{t=SA.k+1}^{SA.k+m+1} \varepsilon_t\right]. \end{aligned}$$

For the choice of $\varepsilon_t = t^{-\frac{1}{m+1}}$ the right-hand side above is zero. To sum up, for all realizations of $(\tau_k(z))_{z\in\mathcal{S}\times\mathcal{A},k\geq 1}$:

$$\prod_{z\in\mathcal{S}\times\mathcal{A}} \mathbb{P}\Big(\mathcal{E}_{z'} |\ (\tau_k(z))_{k\geq 1}\Big) = 0 .$$

Therefore, for all $z', \mathbb{P}(\mathcal{E}_{z'}) = 0$ and consequently $\mathbb{P}(\mathcal{E}) = 0$. ■

2.2.2.2 A necessary condition

When m is unknown, replacing it by its worst-case value gives the forced exploration rates of $t^{-\frac{1}{S}}$ in Lemma 2.3 (resp. $t^{-\frac{1}{2S}}$ in Lemma 2.2). These rates vanish quite slowly when the number of states is large. Therefore we ask the question:

Are these rates really necessary to guarantee sufficient exploration in communicating MDPs?

We give a partially positive answer to this question, by showing that a rate of at least $t^{-\frac{1}{S-1}}$ is necessary in the worst case. Specifically, we show that if the sequence of policies $(\pi_t)_{t\geq 1}$ is such that $\min_{s,a} \pi_t(a|s) = t^{-\alpha}$ decays polynomially, then we need $\alpha < 1/(S-1)$ in order to visit all states infinitely often.

To that end, consider a variant of the classical RiverSwim MDP (Strehl & Littman, 2008) with state (resp. action) space $\mathcal{S} = [|1, S|]$, (resp. $\mathcal{A} = \{\text{LEFT, RIGHT}\}$). After playing RIGHT the agent makes a transition of one step to the right while playing LEFT moves the agent all the way back to state 1. Now suppose that the agent starts at $s = 1$ and allocates a sequence of probabilities $(\varepsilon_t)_{t\geq 1}$[4] to explore the states to the right:

$$\forall s \in \mathcal{S}, \forall t \geq 1,\ \pi_t(\text{RIGHT}|s) = \varepsilon_t = t^{-\alpha}.$$

This induces the non-homogeneous Markov Chain depicted in Figure 2.1.

[4]For simplicity, we assume that this probability is the same for all states.

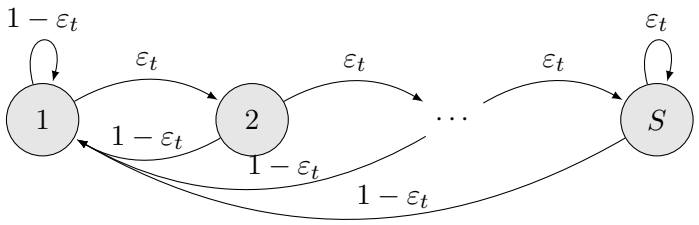


Figure 2.1: Non-homogeneous Markov Chain. An exploration rate of at least $t^{-\frac{1}{S-1}}$ is needed.

Lemma 2.4 If the agent uses any $\alpha > \frac{1}{S-1}$, with non-zero probability she will visit state S only a finite number of times.

Proof. Indeed if the agent visits state S at time k, then the last $S-1$ transitions before k must have been to the right, i.e., $\mathbb{P}(s_k = S) \leq \prod_{j=k-S+1}^{k-1} \varepsilon_j \leq (\varepsilon_{k-S+1})^{S-1}$. Therefore $\mathbb{E}[N_S(t)] \leq \sum_{k=S}^{t} (k-S+1)^{-\alpha(S-1)}$. In particular this implies that for $\alpha > \frac{1}{S-1}$, $\limsup_{t\to\infty} \mathbb{E}[N_S(t)] = M < \infty$. Therefore, using the reverse Fatou lemma and Markov's inequality we get

$$\begin{aligned}
\mathbb{P}(\forall t \geq 1, N_S(t) \leq 2M) &= \mathbb{E}\big[\limsup_{t\to\infty} \prod_{k=1}^{t} \mathbb{1}\left(N_S(k) \leq 2M\right)\big] \\
&\geq \limsup_{t\to\infty} \mathbb{E}\big[\prod_{k=1}^{t} \mathbb{1}\left(N_S(k) \leq 2M\right)\big] \\
&= \limsup_{t\to\infty} \mathbb{E}\big[\mathbb{1}\left(N_S(t) \leq 2M\right)\big] \\
&= \limsup_{t\to\infty} \mathbb{P}\big(N_S(t) \leq 2M\big) \geq \frac{1}{2}.
\end{aligned}$$

■

This proves that, in a worst-case instance like the one above, the probabilities of playing any action must decay at a rate larger than $t^{-1/(S-1)}$. Otherwise, the algorithm only visits some state a finite number of times. This would be problematic since we want to establish the convergence (1.42).

Case of Ergodic MDPs: An MDP is ergodic if the agent can reach any state from any other state using any policy. In other words, for any Markovian policy $\pi \in \Pi_S$, P_π is ergodic. For such MDPs, we can select $\varepsilon_t = 1/t^\alpha$ where $\alpha < 1$ without compromising the conclusion of Lemma 2.3. The proof is deferred to Appendix 2.8.

2.2.3 Convergence of visitation frequencies to the optimal allocation

To establish the convergence of $\mathbf{N}(t)/t$ to $\omega^\star(\mathcal{M})$, we make use of an Ergodic Theorem for non-homogeneous Markov Chains derived by (Fort et al., 2011) which we state below. Its proof can be found in Appendix D of (Al-Marjani et al., 2021).

Notation: For a probability measure μ and a function f, $\mu(f) = \mathbb{E}_{X\sim\mu}[f(X)]$ denotes the mean of f w.r.t. μ. Finally, for two policies π and π' we define $D(\pi,\pi') := \|P_\pi - P_{\pi'}\|_\infty = \max_{z\in\mathcal{S}\times\mathcal{A}} \|P_\pi(z,.) - P_{\pi'}(z,.)\|_1$. We define the $(\mathcal{S}\times\mathcal{A})\times\Pi_S$-valued process $\{(z_t,\pi_t), t \geq 1\}$ where $z_t := (s_t, a_t)$ is the t-th state-action pair on the trajectory of the algorithm. Observe that (z_t,π_t) is $\mathcal{F}_t$-adapted and that for any bounded measurable function f, $\mathbb{E}[f(z_{t+1})|\mathcal{F}_t] = \sum_{z'\in\mathcal{S}\times\mathcal{A}} P_{\pi_t}(z_t,z') f(z')$. We recall the simplified notation $P_t := P_{\pi_t}$.

Proposition 2.2 (Corollary 2.9, (Fort et al., 2011)) Assume that:

(C1) $\forall t \geq 1,\ P_t$ is ergodic. We denote by ω_t its stationary distribution.

(C2) There exists an ergodic kernel P such that $\|P_t - P\|_\infty \underset{t\to\infty}{\longrightarrow} 0$ almost surely.

(C3) There exists two constants C_t and ρ_t such that for all $n \geq 1,\ \|P_t^n - W_t\|_\infty \leq C_t \rho_t^n$, where W_t is a rank-one matrix whose rows are equal to ω_t^T.

(C4) Denote by $L_t := C_t(1-\rho_t)^{-1}$. Then $\limsup_{t\to\infty} L_t < \infty$ almost surely.

(C5) $D(\pi_{t+1}, \pi_t) \underset{t\to\infty}{\longrightarrow} 0$ almost surely.

Finally, denote by $\omega^\star$ the stationary distribution of P. Then for any bounded non-negative function $f : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^+$, it holds that, almost surely,

$$\frac{\sum_{k=1}^{t} f(z_k)}{t} \underset{t\to\infty}{\longrightarrow} \omega^\star(f).$$

Theorem 2.1 Using C-Navigation as a sampling rule, it holds that $\lim_{t\to\infty} \mathbf{N}(t)/t = \omega^\star(\mathcal{M})$ almost surely.

Proof. We will show that C-Navigation satisfies the conditions of Proposition 2.2 for $P = P_{\pi_{\omega^\star(\mathcal{M})}}$ and $\omega^\star = \omega^\star(\mathcal{M})$. The statement of the Theorem follows immediately by applying the Proposition for the functions $f(\tilde{z}) = \mathbb{1}\{\tilde{z} = z\}$, where z is any fixed state-action pair.

(C1): This is a direct consequence of the fact that P_{π_u} is ergodic (due to Assumptions 2.1 and 2.2) which implies by construction that P_t is also ergodic. **(C2):** By Lemma 2.3 we have

$N_{sa}(t) \xrightarrow{a.s} \infty$ for all (s,a). Hence $\big(\widehat{r}_t(s,a), \widehat{p}_t(.|s,a)\big) \xrightarrow{a.s} \big(r_{\mathcal{M}}(s,a), p_{\mathcal{M}}(.|s,a)\big)$. Berge's Maximum theorem (e.g. Theorem 17.31 in Aliprantis & Border, 2006) guarantees that $\omega^\star(\widehat{\mathcal{M}}_t) \xrightarrow{a.s} \omega^\star(\mathcal{M})$ and by continuity of the mapping $\omega \mapsto \pi_\omega$ (see (2.8)), $\pi_{\omega^\star}(\widehat{\mathcal{M}}_t) \xrightarrow{a.s} \pi_{\omega^\star}(\mathcal{M})$. This implies that

$$P_t = \varepsilon_t P_{\pi_u} + (1-\varepsilon_t)\frac{\sum_{k=0}^{t-1} P_{\pi_{\omega^\star}(\widehat{\mathcal{M}}_k)}}{t} \xrightarrow{a.s} P_{\pi_{\omega^\star}(\mathcal{M})}. \tag{2.18}$$

(C5): Since $\big(P_{t+1} - P_t\big)\big((s,a),(s',a')\big) = \big[\pi_{t+1}(a'|s') - \pi_t(a'|s')\big]p_{\mathcal{M}}(s'|s,a)$, it holds that $\|P_{t+1} - P_t\|_\infty \leq \|\pi_{t+1} - \pi_t\|_\infty$, where π_{t+1} and π_t are viewed as vectors of $\mathbb{R}^{SA}$. Next we introduce the notation $\overline{\pi}^t_{\omega^\star} := \frac{\sum_{k=0}^{t-1} \pi_{\omega^\star}(\widehat{\mathcal{M}}_k)}{t}$ for the Cesàro-mean of oracle policies and write

$$\begin{aligned}
\pi_{t+1} - \pi_t &= (\varepsilon_t - \varepsilon_{t+1})(\overline{\pi}^{t+1}_{\omega^\star} - \pi_u) + (1-\varepsilon_t)(\overline{\pi}^{t+1}_{\omega^\star} - \overline{\pi}^t_{\omega^\star}) \\
&= (\varepsilon_t - \varepsilon_{t+1})(\overline{\pi}^{t+1}_{\omega^\star} - \pi_u) + (1-\varepsilon_t)\Big(\frac{t\times\overline{\pi}^t_{\omega^\star} + \pi_{\omega^\star}(\widehat{\mathcal{M}}_t)}{t+1} - \overline{\pi}^t_{\omega^\star}\Big) \\
&= (\varepsilon_t - \varepsilon_{t+1})(\overline{\pi}^{t+1}_{\omega^\star} - \pi_u) + (1-\varepsilon_t)\frac{\pi_{\omega^\star}(\widehat{\mathcal{M}}_t) - \overline{\pi}^t_{\omega^\star}}{t+1}
\end{aligned}$$

Therefore

$$\begin{aligned} D(\pi_{t+1}, \pi_t) &= \|P_{t+1} - P_t\|_\infty \\ &\leq \|\pi_{t+1} - \pi_t\|_\infty \\ &\leq (\varepsilon_t - \varepsilon_{t+1}) + \frac{1}{t+1} \underset{t\to\infty}{\longrightarrow} 0. \end{aligned}$$

(C3): By Lemma 2.9, P_t satisfies (C3) for $C_t = 2\theta(\varepsilon_t, \overline{\pi}^t_{\omega^\star}, \omega_t)^{-1}$ and $\rho_t = \theta(\varepsilon_t, \overline{\pi}^t_{\omega^\star}, \omega_t)^{1/r}$ where $(\varepsilon, \pi, \omega) \mapsto \theta(\varepsilon, \pi, \omega)$ was defined in Appendix 2.5.
(C4): By definition, we have

$$\begin{aligned} \sigma(\varepsilon_t, \overline{\pi}^t_{\omega^\star}, \omega_t) &= \left(\varepsilon_t^r + \left[(1-\varepsilon_t) A \min_{s,a} \overline{\pi}^t_{\omega^\star}(a|s)\right]^r\right) \sigma_u\left(\min_z \frac{\omega_u(z)}{\omega_t(z)}\right) \\ &\xrightarrow{a.s} \left(A \min_{s,a} \pi_{\omega^\star}(a|s)\right) \sigma_u \min_z \frac{\omega_u(z)}{\omega^\star(z)} := \sigma^\star, \end{aligned} \tag{2.19}$$

where the convergence was established in the proof of (C1). Note that $\sigma^\star > 0$ since $\omega_u > 0$ (ergodicity of P_{π_u}), $\omega^\star < 1$ and $\pi_{\omega^\star} > 0$ entry-wise. Similarly, it is trivial that $\sigma^\star < 1$ since $A \min_{s,a} \pi_{\omega^\star}(a|s) < 1, \min_z \frac{\omega_u(z)}{\omega^\star(z)} < 1$ and $\sigma_u \leq 1$. Therefore $\theta(\varepsilon_t, \overline{\pi}^t_{\omega^\star}, \omega_t) = 1 - \sigma(\varepsilon_t, \overline{\pi}^t_{\omega^\star}, \omega_t) \xrightarrow{a.s} 1 - \sigma^\star := \theta^\star \in (0,1)$ and

$$\begin{aligned} \limsup_{t\to\infty} L_t &= \limsup_{t\to\infty} C_t (1-\rho_t)^{-1} \\ &= \limsup_{t\to\infty} \frac{2}{\theta(\varepsilon_t, \overline{\pi}^t_{\omega^\star}, \omega_t)\left[1 - \theta(\varepsilon_t, \overline{\pi}^t_{\omega^\star}, \omega_t)^{1/r}\right]} \\ &= \frac{2}{\theta^\star\left[1 - (\theta^\star)^{\frac{1}{r}}\right]} < \infty. \end{aligned} \tag{2.20}$$

■

2.3 Navigate-and-Stop

Navigate-and-Stop (NaS) is a model-based algorithm inspired by the lower bound. The lower bound suggests that to identify the best policy in a sample-efficient manner, an algorithm must collect samples from state-action pair (s,a) proportionally to $\omega^\star_{sa}(\mathcal{M})$. For that, we use C-Navigation which satisfies (1.42). C-Navigation is then combined with a Generalized Likelihood Ratio Test (GLRT)[5]. If we denote by $\widehat{\pi}^\star_t$ the optimal policy in the empirical MDP $\widehat{\mathcal{M}}_t$, the GLRT stops as soon as we are confident that $\widehat{\pi}^\star_t = \pi^\star$ with probability at least $1-\delta$. The pseudo-code for NaS is given in Algorithm 8.

2.3.1 Pseudo-code

NaS starts by drawing a random MDP with a unique optimal policy (for example, it can select Bernoulli rewards with means drawn from the uniform distribution on $[0,1]$ and transitions from a Dirichlet distribution $\mathcal{D}(1,\dots,1)$) that will serve as an initial estimate $\widehat{\mathcal{M}}_0$ of $\mathcal{M}$. The algorithm maintains, after t time steps, an empirical estimate $\widehat{\mathcal{M}}_t$ of the true MDP. Based on this estimate, NaS computes the empirical oracle policy $\pi_{\omega^\star}(\widehat{\mathcal{M}}_t)$ defined in (2.8), and selects the action to play depending on the current state $a_t \sim \pi_t(.|s_t)$, where π_t is given by either of our sampling rules. After each observation, $\widehat{\mathcal{M}}_t$ is updated. Finally, the algorithm checks if the stopping condition in (2.23) is satisfied, in which case

[5] Rather a proxy of the GLRT, see Section 2.3.2 for details

Algorithm 8 Navigate-and-Stop (NaS)

Input: risk $\delta \in (0,1)$, ERGODIC boolean variable, communication parameter m or an upper bound.
if ERGODIC = True **:**
  Set $(\varepsilon_t)_{t\geq 1} = (1/\sqrt{t})_{t\geq 1}$
else:
  Set $(\varepsilon_t)_{t\geq 1} = (t^{-\frac{1}{m+1}})_{t\geq 1}$
Set $t \leftarrow 0$ and $N_{sa}(t) \leftarrow 0$, for all (s,a)
Initialize empirical estimate $\widehat{\mathcal{M}}_0$ by drawing an arbitrary MDP from $\mathfrak{M}_{\star,1}$
for $t = 1, 2, \ldots$ **do**
  Compute $\omega^\star(\widehat{\mathcal{M}}_{t-1})$ by solving (2.6) and the corresponding policy $\pi_{\omega_\star}(\widehat{\mathcal{M}}_{t-1})$ by normalization (2.8)
  Set $\pi_t \leftarrow \varepsilon_t \pi_u + (1-\varepsilon_t)\frac{\sum_{j=0}^{t-1} \pi_{\omega^\star}(\widehat{\mathcal{M}}_j)}{t}$
  Play $a_t \sim \pi_t(.|s_t)$ and observe reward R_t and next state s_{t+1}.
  Update empirical estimates $\big(\widehat{r}_t(s,a), \widehat{p}_t(s,a)\big)_{s,a}$ and counts $(N_{sa}(t))_{sa}$
  if $t \cdot U\big(\widehat{\mathcal{M}}_t, \mathbf{N}(t)/t\big)^{-1} \geq \beta(t,\delta)$ **:**
    Stop and return $\widehat{\pi}_t^\star := \pi^\star(\widehat{\mathcal{M}}_t)$
  end if
end for

it stops and returns the empirical optimal policy $\widehat{\pi}_\tau^\star$. The exploration rate used by NaS depends on a boolean variable that indicates whether we have prior knowledge that $\mathcal{M}$ is ergodic or not.

2.3.2 Stopping rule

Assumption 2.3 We assume that the reward distributions $q_{\mathcal{M}'}(s,a)$ for MDPs in $\mathfrak{M}_{\star,1}$ come from a single-parameter exponential family (SPEF) and can therefore be parametrized by their respective means $r_{\mathcal{M}'}(s,a)$.

Under the previous assumption, we can easily build MLE estimates of the reward distributions by computing the empirical mean. For any (s,a) and $t \geq 1$ such that $N_{sa}(t) > 0$, we let $\widehat{q}_{s,a}(t)$ denote the distribution belonging to the SPEF of our model, whose mean is the empirical average $\widehat{r}_t(s,a) = \frac{\sum_{k=1}^{t} R_k \mathbb{1}(s_t=s, a_t=a)}{N_{sa}(t)}$.

2.3.2.1 Some Intuition on the GLRT

To implement a Generalized Likelihood Ratio Test (GLRT), we define $\ell_{\mathcal{M}'}(t)$, the likelihood of the observations under some MDP $\mathcal{M}' \in \mathfrak{M}_{\star,1}$ by

$$\ell_{\mathcal{M}'}(t) := \prod_{k=1}^{t-1} p_{\mathcal{M}'}(s_{k+1}|s_k,a_k) q_{\mathcal{M}'}(R_k|s_k,a_k),$$

where at step k the algorithm is in state s_k, plays action a_k and observes the reward R_k and s_{k+1} (the next state). Performing a GLRT at step t consists in (1) computing the optimal policy $\widehat{\pi}_t^\star$ for the estimated MDP $\widehat{\mathcal{M}}_t$; (2) comparing the likelihood of observations under the most likely model where $\widehat{\pi}_t^\star$ is optimal to the likelihood under the most likely

model where $\widehat{\pi}_t^\star$ is sub-optimal. To that end, we define the ratio

$$\mathrm{GLR}(t;\widehat{\pi}_t^\star) := \log \frac{\sup_{\mathcal{M}'\in\mathfrak{M}_{\star,1}:\pi^\star(\mathcal{M}')=\widehat{\pi}_t^\star} \ell_{\mathcal{M}'}(t)}{\sup_{\mathcal{M}'\in\mathfrak{M}_{\star,1}:\pi^\star(\mathcal{M}')\neq\widehat{\pi}_t^\star} \ell_{\mathcal{M}'}(t)}.$$

Intuitively, if $\mathrm{GLR}(t;\widehat{\pi}_t^\star)$ is large, then the evidence in favor of $(\widehat{\pi}_t^\star = \pi^\star)$ is stronger than the evidence for $(\widehat{\pi}_t^\star \neq \pi^\star)$. Therefore, we reject the hypothesis $(\widehat{\pi}_t^\star \neq \pi^\star)$ as soon as this ratio of likelihoods becomes greater than some threshold $\beta(t,\delta)$, properly tuned to ensure that the algorithm is δ-PAC.

2.3.2.2 Tuning the threshold of the stopping time

Notation: To simplify the presentation, we write $\widehat{q}_{s,a}(t) := \widehat{q}_t(.|s,a)$ and $\widehat{p}_{s,a}(t) := \widehat{p}_t(.|s,a)$ for the empirical reward and transition distributions at step t. Similarly, we denote $q_{\mathcal{M}'}(s,a) := q_{\mathcal{M}'}(.|s,a)$ and $p_{\mathcal{M}'}(s,a) := p_{\mathcal{M}'}(.|s,a)$.

The next Lemma gives a simplified expression of the GLR that will be useful in the design of our stopping rule. Its proof is deferred to Appendix 2.11.

Lemma 2.5 It holds that

$$\mathrm{GLR}(t;\widehat{\pi}_t^\star) = t\, T\big(\widehat{\mathcal{M}}_t, \mathbf{N}(t)/t\big)^{-1}, \tag{2.21}$$

$$= \inf_{\mathcal{M}'\in\mathrm{Alt}(\widehat{\mathcal{M}}_t)} \sum_{s,a} N_{sa}(t)\Big[\mathrm{KL}(\widehat{q}_{s,a}(t), q_{\mathcal{M}'}(s,a)) + \mathrm{KL}(\widehat{p}_{s,a}(t), p_{\mathcal{M}'}(s,a))\Big], \tag{2.22}$$

where $(\mathcal{M},\omega)\mapsto T(\mathcal{M},\omega)$ was defined in (2.1).

(2.22) suggests that we need concentration inequalities on the Kullback-Leibler divergence of transitions and reward distributions to set a proper threshold $\beta(t,\delta)$. This is the purpose of the next Lemma, whose proof can be found in Appendix E of (Al-Marjani et al., 2021).

Lemma 2.6 Define the thresholds for the transitions and rewards respectively,

$$\beta_p(t,\delta) := \log(1/\delta) + (S-1)\sum_{(s,a)} \log\big(e\big[1+N_{sa}(t)/(S-1)\big]\big),$$

$$\beta_r(t,\delta) := SA\,\varphi\big(\log(1/\delta)/SA\big) + 3\sum_{s,a}\log\big(1+\log(N_{sa}(t))\big),$$

where $x\mapsto\varphi(x)$ is defined in the Appendix E of (Al-Marjani et al., 2021) and satisfies $\varphi(x)\underset{\infty}{\sim} x$. Then for the threshold $\beta(t,\delta) := \beta_r(t,\delta/2) + \beta_p(t,\delta/2)$ we have that

$$\mathbb{P}_{\mathcal{M},\mathbb{A}}\left(\exists t\geq 1, \sum_{s,a} N_{sa}(t)\big[\mathrm{KL}\big(\widehat{q}_{s,a}(t), q_{\mathcal{M}}(s,a)\big) + \mathrm{KL}\big(\widehat{p}_{s,a}(t), p_{\mathcal{M}}(s,a)\big)\big] \geq \beta(t,\delta)\right) \leq \delta.$$

Remark 2.5 Observe that $\beta(t,\delta)\underset{\delta\to 0}{\sim} 2\log(1/\delta)$. This will be crucial when analyzing the sample complexity of NaS. ∎

Computing the likelihood ratio $\mathrm{GLR}(t;\widehat{\pi}_t^\star)$ can be difficult, since that is equivalent to solving (2.1), see Section2.1. We circumvent this issue by using a lower bound on the GLR, which leads to the following Theorem.

Theorem 2.2 Combining C-Navigation with the stopping rule

$$\tau_\delta := \inf\left\{ t \geq 1 :\ t\, U\big(\widehat{\mathcal{M}}_t, \mathbf{N}(t)/t\big)^{-1} \geq \beta(t,\delta) \right\} \tag{2.23}$$

yields a δ-PAC algorithm for BPI, i.e., $\mathbb{P}_{\mathcal{M},\mathbb{A}}(\tau_\delta < \infty,\ \widehat{\pi}^\star_{\tau_\delta} = \pi^\star) \geq 1-\delta$.

Proof. Observe that $\beta(t,\delta) = \mathcal{O}_{t\to\infty}(\log(t))$. On the other hand, by Theorem 2.1 we have that $t\, U\big(\widehat{\mathcal{M}}_t, \mathbf{N}(t)/t\big)^{-1} \underset{t\to\infty}{\sim} U(\mathcal{M}, \omega^\star(\mathcal{M}))\cdot t$ almost surely. Therefore τ_δ is finite almost surely. Now assume that the algorithm stops at time step t while $\widehat{\pi}^\star_t \neq \pi^\star$. This means that $\mathcal{M} \in \mathrm{Alt}(\widehat{\mathcal{M}}_t)$. Hence,

$$\begin{aligned}
\mathbb{P}(\widehat{\pi}^*_\tau \neq \pi^*, \tau_\delta < \infty) &= \mathbb{P}\Big(\exists t \geq 1 :\ t\, U\big(\widehat{\mathcal{M}}_t, \mathbf{N}(t)/t\big)^{-1} \geq \beta(t,\delta),\ \widehat{\pi}^*_t \neq \pi^*\Big) \\
&\overset{(a)}{\leq} \mathbb{P}\Big(\exists t \geq 1 :\ t\, T\big(\widehat{\mathcal{M}}_t, \mathbf{N}(t)/t\big)^{-1} \geq \beta(t,\delta),\ \mathcal{M} \in \mathrm{Alt}(\widehat{\mathcal{M}}_t)\Big) \\
&\overset{(b)}{=} \mathbb{P}\Big(\exists t \geq 1 :\ \inf_{\mathcal{M}'\in\mathrm{Alt}(\widehat{\mathcal{M}}_t)} \sum_{s,a} N_{sa}(t)\big[\mathrm{KL}\big(\widehat{q}_{s,a}(t), q_{\mathcal{M}'}(s,a)\big) + \mathrm{KL}\big(\widehat{p}_{s,a}(t), p_{\mathcal{M}'}(s,a)\big)\big] \\
&\qquad\qquad \geq \beta(t,\delta),\ \mathcal{M} \in \mathrm{Alt}(\widehat{\mathcal{M}}_t)\Big) \\
&\leq \mathbb{P}\Big(\exists t \geq 1 :\ \sum_{s,a} N_{sa}(t)\big[\mathrm{KL}\big(\widehat{q}_{s,a}(t), q_{\mathcal{M}}(s,a)\big) + \mathrm{KL}\big(\widehat{p}_{s,a}(t), p_{\mathcal{M}}(s,a)\big)\big] \\
&\qquad\qquad \geq \beta(t,\delta)\Big) \\
&\overset{(c)}{\leq} \delta
\end{aligned}$$

where (a), (b) and (c) use Lemmas 2.1, 2.5 and 2.6 respectively. ∎

2.4 Sample Complexity of Navigate-and-Stop

2.4.1 Main Theorem

Now we present the main guarantees on the sample complexity of NaS. We will only prove the first statement of the next Theorem. For the second statement, we just give a proof sketch as the full proof is somewhat involved.

Theorem 2.3 (i) NaS stops almost surely and its stopping time satisfies

$$\mathbb{P}_{\mathcal{M},\mathbb{A}}\left(\limsup_{\delta\to 0} \frac{\tau_\delta}{\log(1/\delta)} \leq 2U^\star(\mathcal{M}) \right) = 1,$$

where $U^\star(\mathcal{M})$ was defined in (2.5);
(ii) The stopping time of NaS has a finite expectation for all $\delta \in (0,1)$ and

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\mathcal{M},\mathbb{A}}[\tau_\delta]}{\log(1/\delta)} \leq 2U^\star(\mathcal{M}).$$

2.4.2 Proof of the almost-sure asymptotic complexity

Consider the event $\mathcal{E} = \big(\forall (s,a) \in \mathcal{S}\times\mathcal{A},\ \lim_{t\to\infty} \frac{N_{sa}(t)}{t} = \omega^\star_{sa}(\mathcal{M})$ and $\widehat{\mathcal{M}}_t \underset{t\to\infty}{\to} \mathcal{M}\big)$. By Lemma 2.3 and Theorem 2.1, we have $\mathbb{P}_{\mathcal{M},\mathbb{A}}(\mathcal{E}) = 1$. We will prove that under $\mathcal{E}$, $\limsup_{\delta\to 0} \frac{\tau_\delta}{\log(1/\delta)} \le 2U_o(\mathcal{M})$. Fix $\eta > 0$. Under $\mathcal{E}$ and using the continuity of $(\mathcal{M},\omega) \mapsto U(\mathcal{M},\omega)$. There exits t_η such that for all $t \ge t_\eta$

$$U\big(\widehat{\mathcal{M}}_t, \mathbf{N}(t)/t\big)^{-1} \ge (1-\eta) U\big(\mathcal{M},\omega^\star\big)^{-1}, \tag{2.24}$$

$$\beta(t,\delta) \le \log(1/\delta) + SA\,\varphi\big(\log(1/\delta)/SA\big) + \eta U\big(\mathcal{M},\omega^\star\big)^{-1} t, \tag{2.25}$$

where the last inequality comes from the fact that the threshold satisfies $\beta(t,\delta) = \log(1/\delta) + SA\,\varphi\big(\log(1/\delta)/SA\big) + \mathcal{O}_{t\to\infty}\big(\log(t)\big)$. Combining the inequalities above with the definition of τ_δ, we get

$$\begin{aligned}\tau_\delta &\le \inf\left\{ t \ge t_\eta, (1-2\eta) t U\big(\mathcal{M},\omega^\star\big)^{-1} \ge \log(1/\delta) + SA\,\varphi\big(\log(1/\delta)/SA\big)\right\} \\ &= \max\left(t_\eta,\ \frac{\Big[\log(1/\delta) + SA\,\varphi\big(\log(1/\delta)/SA\big)\Big] U\big(\mathcal{M},\omega^\star\big)}{1-2\eta}\right).\end{aligned}$$

Since $\varphi(x) \underset{\infty}{\sim} x$, then the last inequality implies that $\limsup_{\delta\to 0} \frac{\tau_\delta}{\log(1/\delta)} \le \frac{2U(\mathcal{M},\omega^\star)}{1-2\eta}$. Taking the limit when η goes to zero finishes the proof.

2.4.3 Proof sketch for the expected sample complexity

The starting point of our proof is a concentration event of the empirical estimates $\widehat{\mathcal{M}}_t$ around $\mathcal{M}$. For $\xi > 0$ and $T \ge 1$, we define

$$\mathcal{C}^1_T(\xi) := \bigcap_{t=T^{1/4}}^{T} \left(\left\|\widehat{\mathcal{M}}_t - \mathcal{M}\right\| \le \xi, \left\|\pi_{\omega^\star}(\widehat{\mathcal{M}}_t) - \pi_{\omega^\star}(\mathcal{M})\right\|_\infty \le \xi \right),$$

where $\|\mathcal{M}' - \mathcal{M}\| := \max_{s,a} |r_{\mathcal{M}'}(s,a) - r_{\mathcal{M}}(s,a)| + \|p_{\mathcal{M}'}(.|s,a) - p_{\mathcal{M}}(.|s,a)\|_1$ is a semi-distance on MDPs in $\mathfrak{M}_{\star,1}$. Thanks to Lemma 2.2, we show in Lemma 18 of (Al-Marjani et al., 2021) that $\mathcal{C}^1_T(\xi)$ holds with high probability in the sense that

$$\forall T \ge 1, \quad \mathbb{P}_{\mathcal{M},\mathbb{A}}\left(\mathcal{C}^1_T(\xi)\right) \ge 1 - \mathcal{O}_{T\to\infty}\big(1/T^2\big). \tag{2.26}$$

In a second step, we adapt the proof of (Fort et al., 2011) to derive a finite-time version of Theorem 2.1 which results into the following proposition.

Proposition 2.3 — Proposition 19, (Al-Marjani et al., 2021). Under C-Navigation, for all $\xi > 0$, there exists a time T_ξ such that for all $T \ge T_\xi$, all $t \ge T^{3/4}$ and all functions $f : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^+$, we have

$$\mathbb{P}_{\mathcal{M},\mathbb{A}}\left(\left| \frac{\sum_{k=1}^t f(s_k,a_k)}{t} - \mathbb{E}_{(s,a)\sim\omega^\star(\mathcal{M})}[f(s,a)] \right| \ge K_\xi \|f\|_\infty \xi \,\middle|\, \mathcal{C}^1_T(\xi) \right) \le 2\exp\big(-t\xi^2\big).$$

where $\xi \mapsto K_\xi$ is a mapping with values in $(1,\infty)$ such that $\limsup_{\xi\to 0} K_\xi < \infty$.

Let us define

$$\mathcal{C}^2_T(\xi) := \bigcap_{t=T^{3/4}}^{T} \big(\|\mathbf{N}(t)/t - \omega^\star(\mathcal{M})\|_\infty \le K_\xi\, \xi \big).$$

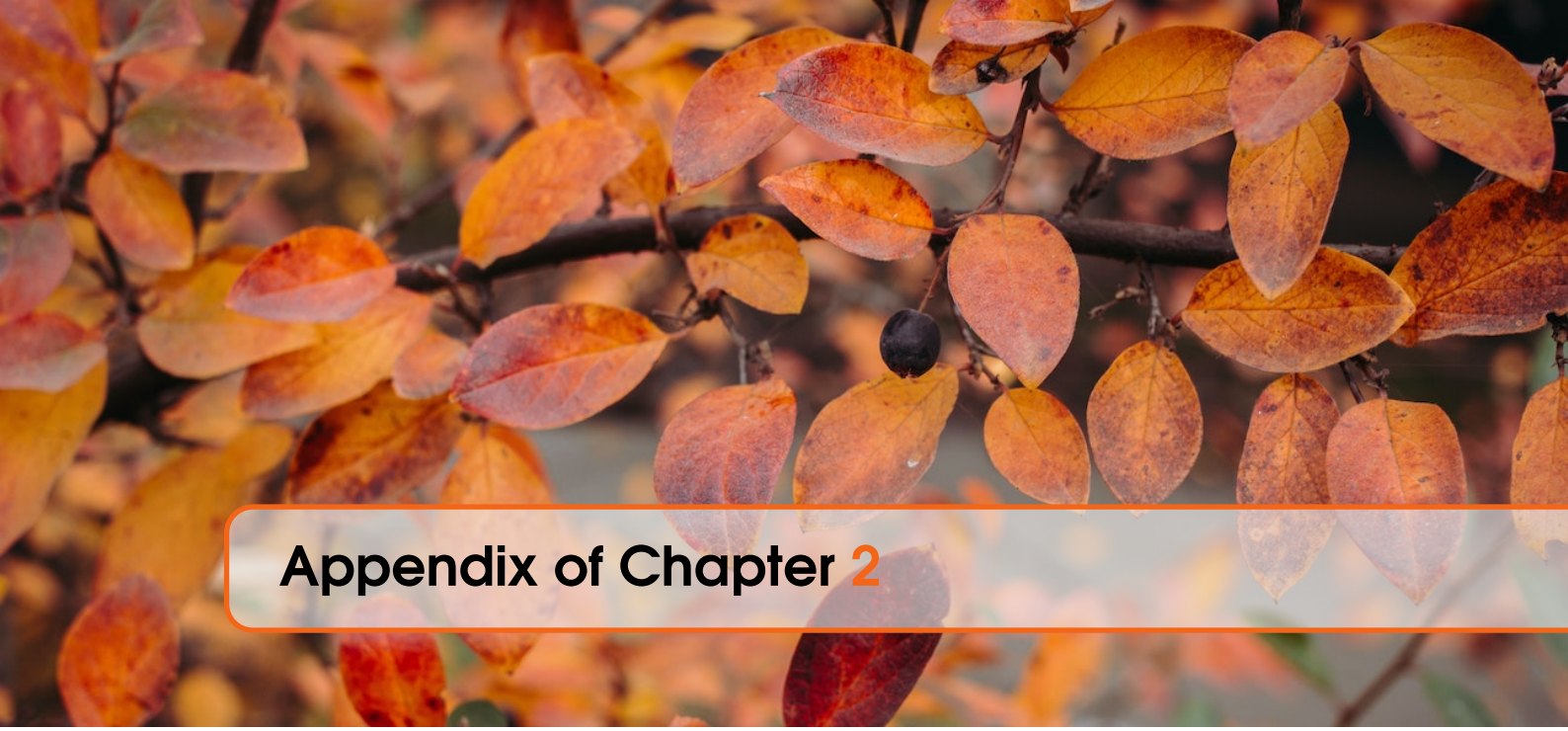
Proposition 2.3 and Eq. (2.26) combined imply that for T large enough, the event $\mathcal{C}_T^1(\xi) \cap \mathcal{C}_T^2(\xi)$ holds with high probability. From this, we conclude that the expected stopping time is finite on the complementary event, $\mathbb{E}_{\mathcal{M},\mathbb{A}}\big[\tau_\delta \mathbb{1}\big(\overline{\mathcal{C}_T^1(\xi)} \cup \overline{\mathcal{C}_T^2(\xi)}\big)\big] < \infty$. On the other hand, given the asymptotic shape of the threshold $\beta(t,\delta) \underset{\delta\to 0}{\sim} 2\log(1/\delta)$, we may informally write

$$\mathbb{E}\big[\tau_\delta \mathbb{1}\left(\mathcal{C}_T^1(\xi) \cap \mathcal{C}_T^2(\xi)\right)\big] \underset{\delta\to 0}{\preceq} 2\log(1/\delta) \sup_{(\mathcal{M}',\omega')\in B_\xi} U^\star\big(\mathcal{M}',\omega'\big),$$

where $B_\xi = \{(\mathcal{M}',\omega') : \|\mathcal{M}' - \mathcal{M}\| \le \xi, \|\omega' - \omega^\star(\mathcal{M})\|_\infty \le K_\xi\, \xi\}$. Dividing by $\log(1/\delta)$ and taking the limits when δ and ξ go to zero respectively concludes the proof.

2.5 Discussion

We have designed a sampling rule that overcomes the navigation challenge and achieves the optimality recipe (1.42) for any mapping of allocation vectors $\mathcal{M} \mapsto \omega^\star(\mathcal{M})$. Such sampling rule can lead to many different algorithms that enjoy instance-dependent guarantees, simply by changing the definition of the objective in (2.3) to another problem-dependent quantity. One limitation of our results is that they only cover the asymptotic regime $\delta \to 0$. In the future, it would be interesting to derive instance-dependent bounds that hold for any $\delta \in (0,1)$. We note that such finite-time bounds have been obtained only recently for the simpler setting of finite-armed bandits, see (Barrier et al., 2022) for instance.



Appendix of Chapter 2

2.6 Definition of $H^\star$

Let $\mathrm{Var}^\star_{\max}[V^\star_{\mathcal{M}}] = \max_{s\in\mathcal{S}} \mathrm{Var}_{s'\sim p_{\mathcal{M}}(.|s,\pi^\star(s))}[V^\star_{\mathcal{M}}(s')]$ denote the maximum variance of the value function on the trajectory of the optimal policy. Further let $\mathrm{sp}(f) := \sup_{x,x'\in\mathcal{X}} |f(x') - f(x)|$ denote the span of a function real-valued f. Then (Al-Marjani & Proutiere, 2021) define:

$$
\begin{aligned}
H^\star &:= S(T_3(\mathcal{M}) + T_4(\mathcal{M})) \\
T_3(\mathcal{M}) &:= \frac{2}{\Delta^2_{\min}(1-\gamma)^2}, \\
T_4(\mathcal{M}) &:= \min\left(\frac{27}{\Delta^2_{\min}(1-\gamma)^3},\ \max\left(\frac{16\mathrm{Var}^\star_{\max}[V^\star_{\mathcal{M}}]}{\Delta^2_{\min}(1-\gamma)^2}, \frac{6\,\mathrm{sp}(V^\star_{\mathcal{M}})^{4/3}}{\Delta^{4/3}_{\min}(1-\gamma)^{4/3}}\right)\right).
\end{aligned}
$$

Note that $H^\star = \mathcal{O}\big(\frac{S}{\Delta^2_{\min}(1-\gamma)^3}\big)$.

2.7 Upper Bound on the Norm of Products of Substochastic Matrices

Before we proceed with the lemma, we lay out some definitions. $\eta_1 := \min\big\{P_{\pi_u}(z,z') \,\big|(z,z') \in (\mathcal{S}\times\mathcal{A})^2, P_{\pi_u}(z,z') > 0\big\}$ denotes the minimum positive probability of transition in $\mathcal{M}$. Similarly define $\eta_2 := \min\big\{P^n_{\pi_u}(z,z') \,\big|(z,z') \in \mathcal{S}\times\mathcal{A}^2, n \in [|1, m+1|], P^n_{\pi_u}(z,z') > 0\big\}$ the minimal probability of reaching some state-action pair z' from any other state-action z after $n \le m+1$[6] transitions in the Markov chain induced by the uniform random policy. Finally, $\eta := \eta_1\eta_2$.

Lemma 2.7 Fix some state-action z and let P_t be the transition matrix under some policy π_t satisfying $\pi_t(a|s) \ge \varepsilon_t \pi_u(a|s)$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Define the substochastic matrix

[6] Refer to Remark 2.3 for more detail.

Q_t obtained by removing from P_t the row and the column corresponding to z:

$$P_t = \left(\begin{array}{c|c} Q_t & [P_t(z',z)]_{z'\neq z} \\ \hline [P_t(z,z')]^T_{z'\neq z} & P_t(z,z) \end{array}\right).$$

Then we have:

$$\forall n \geq 1, \left\| \prod_{l=n+1}^{n+m+1} Q_l \right\|_\infty \leq 1 - \eta \prod_{l=n+1}^{n+m+1} \varepsilon_l.$$

Proof. Define $r_k(n_1,n_2) = \sum_{j=1}^{SA-1} \left(\prod_{l=n_1+1}^{n_2} Q_l \right)_{kj}$ the sum of the k-th row in the product of matrices Q_l for $l \in [|n_1+1, n_2|]$. We will prove that or all $i \in [|1, SA-1|]$: $r_i(n, n+m+1) \leq 1 - \eta \prod_{l=n+1}^{n+m+1} \varepsilon_l$. The result follows immediately by noting that $\left\| \prod_{l=n+1}^{n+m+1} Q_l \right\|_\infty = \max_{i\in[|1,SA-1|]} r_i(n, n+m+1)$.

Consider z' such that $P_{\pi_u}(z',z) \geq \eta_1$ (such z' always exists since $\mathcal{M}$ is communicating) and let $k^\star$ be the index of the row corresponding to z' in Q_t. Then for all $n_1 \geq 1$:

$$\begin{aligned} r_{k^\star}(n_1, l = n_1+1) &= \sum_{j=1}^{SA-1} (Q_{n_1+1})_{k^\star j} \\ &= 1 - P_{n_1+1}(z',z) \\ &\leq 1 - \eta_1 \varepsilon_{n_1+1}. \end{aligned} \tag{2.27}$$

Now for $n_1, n_2 \geq 1$ we have:

$$\begin{aligned} r_{k^\star}(n_1, n_1+n_2) &= \sum_{j_1=1}^{SA-1} \left(\prod_{l=n_1+1}^{n_1+n_2} Q_l \right)_{k^\star j_1} \\ &= \sum_{j_1=1}^{SA-1} \sum_{j_2=1}^{SA-1} \left(\prod_{l=n_1+1}^{n_1+n_2-1} Q_l \right)_{k^\star j_2} (Q_{n_1+n_2})_{j_2 j_1} \\ &= \sum_{j_2=1}^{SA-1} \left(\prod_{l=n_1+1}^{n_1+n_2-1} Q_l \right)_{k^\star j_2} \left[\sum_{j_1=1}^{SA-1} (Q_{n_1+n_2})_{j_2 j_1} \right] \\ &= \sum_{j_2=1}^{SA-1} \left(\prod_{l=n_1+1}^{n_1+n_2-1} Q_l \right)_{k^\star j_2} r_{j_2}(n_1+n_2-1, n_1+n_2) \\ &\leq r_{k^\star}(n_1, n_1+n_2-1) \\ &\vdots \\ &\leq r_{k^\star}(n_1, n_1+1) \\ &\leq 1 - \eta_1 \varepsilon_{n_1+1}, \end{aligned} \tag{2.28}$$

where in the fifth line we use the fact that for all j_2, a, b: $r_{j_2}(a,b) \leq 1$ since the matrices Q_l are substochastic. The last line comes from (2.27). Now for all other indexes $i \in [|1, SA-1|]$

we have:

$$\begin{aligned}
\forall n_1 \in [|1, m|],\ r_i(n, n+m+1) &= \sum_{j_1=1}^{SA-1} \left(\prod_{l=n+1}^{n+n_1} Q_l \times \prod_{l=n+n_1+1}^{n+m+1} Q_l \right)_{ij_1} \\
&= \sum_{j_1=1}^{SA-1} \sum_{j_2=1}^{SA-1} \left(\prod_{l=n+1}^{n+n_1} Q_l \right)_{ij_2} \left(\prod_{l=n+n_1+1}^{n+m+1} Q_l \right)_{j_2 j_1} \\
&= \sum_{j_2=1}^{SA-1} \left(\prod_{l=n+1}^{n+n_1} Q_l \right)_{ij_2} \sum_{j_1=1}^{SA-1} \left(\prod_{l=n+n_1+1}^{n+m+1} Q_l \right)_{j_2 j_1} \\
&= \sum_{j_2=1}^{SA-1} \left(\prod_{l=n+1}^{n+n_1} Q_l \right)_{ij_2} r_{j_2}(n+n_1, n+m+1) \\
&\leq (1 - \eta_1 \varepsilon_{n+n_1+1}) \left(\prod_{l=n+1}^{n+n_1} Q_l \right)_{ik^\star} + \sum_{j_2 \neq k^\star} \left(\prod_{l=n+1}^{n+n_1} Q_l \right)_{ij_2} \\
&\leq (1 - \eta_1 \varepsilon_{n+n_1+1}) \left(\prod_{l=n+1}^{n+n_1} Q_l \right)_{ik^\star} + 1 - \left(\prod_{l=n+1}^{n+n_1} Q_l \right)_{ik^\star} \\
&= 1 - \eta_1 \varepsilon_{n+n_1+1} \left(\prod_{l=n+1}^{n+n_1} Q_l \right)_{ik^\star}, \qquad (2.29)
\end{aligned}$$

where we used (2.28) and the fact that the matrix $\prod_{l=n+1}^{n+n_1} Q_l$ is substochastic. Now since $\mathcal{M}$ is communicating then we can reach state-action z' from any other state-action $z_i \in [|1, SA-1|]$, after some $n_i \leq m+1$ steps in the Markov chain corresponding to the random uniform policy. In other words, if i is the index corresponding to z_i then there exists $n_i \leq m+1$, such that $(P_{\pi_u}^{n_i})_{ik^\star} \geq \eta_2 > 0$. Therefore

$$\begin{aligned}
\left(\prod_{l=n+1}^{n+n_i} Q_l \right)_{ik^\star} &\geq \left(\prod_{l=n+1}^{n+n_i} \varepsilon_l P_{\pi_u} \right)_{ik^\star} \\
&= \left(\prod_{l=n+1}^{n+n_i} \varepsilon_l \right) (P_{\pi_u}^{n_i})_{ik^\star} \\
&\geq \eta_2 \prod_{l=n+1}^{n+n_i} \varepsilon_l. \qquad (2.30)
\end{aligned}$$

Thus, combining (2.29) for $n_1 = n_i$ and (2.30) we get:

$$\begin{aligned}
\forall i \in [|1, SA-1|],\ r_i(n, n+m+1) &\leq 1 - \eta_1 \eta_2 \prod_{l=n+1}^{n+n_i} \varepsilon_l \\
&\leq 1 - \eta_1 \eta_2 \prod_{l=n+1}^{n+m+1} \varepsilon_l \\
&= 1 - \eta \prod_{l=n+1}^{n+m+1} \varepsilon_l.
\end{aligned}$$

■

2.8 Minimal Exploration Rate for Ergodic MDPs

This is a consequence of Proposition 2 (Burnetas & Katehakis, 1997), stating that there exist $c_1, c_2, C > 0$ such that for all s and t large enough, $\mathbb{P}_{\mathcal{M},\mathbb{A}}[N_s(t) > c_1 t] \geq 1 - Ce^{-c_2 t}$. A union bound yields: $\mathbb{P}_{\mathcal{M},\mathbb{A}}[\forall s, N_s(t) > c_1 t] \geq 1 - CSe^{-c_2 t}$. To extend this result to the numbers of visits at the various state-action pairs, we can derive a lower bound on $N_{sa}(t)$ given that $N_s(t) > c_1 t$ by observing that a worst scenario (by monotonicity of ε_s) occurs when s is visited only in the $c_1 t$ rounds before t. We get $\mathbb{E}[N_{sa}(t)|N_s(t) > c_1 t] \geq c_3 t^{(1-\alpha)}$. Remarking that $N_{sa}(t+1) - N_s(t)\varepsilon_t$ is a sub-martingale with bounded increments, standard concentration arguments then imply that $\mathbb{P}_{\mathcal{M},\mathbb{A}}[\forall s, a, N_{sa}(t) > \frac{c_3}{2} t^{(1-\alpha)}] \geq \varphi(t)$, where $\varphi(t) \to 1$. Next, define the random variable $Z_t = \prod_{s,a} \mathbb{1}\{N_{sa}(t) > \frac{c_3}{2} t^{(1-\alpha)}\}$. Applying the reverse Fatou lemma, we get $1 = \limsup_t \mathbb{E}[Z_t] \leq \mathbb{E}[\limsup_t Z_t]$. From there, we directly deduce (by monotonicity of $t \mapsto N_{sa}(t)$) that a.s. $\lim_{t\to\infty} N_{sa}(t) = \infty$.

2.9 Geometric Convergence of Iterates of an Ergodic Chain

The following lemma is adapted from the proof of the Convergence theorem (Theorem 4.9, (Levin et al., 2006)).

Lemma 2.8 Let P be a stochastic matrix with stationary distribution vector ω. Suppose that there exist $\sigma > 0$ and an integer r such that $P^r(s, s') \geq \sigma\omega(s')$ for all (s, s'). Let W be a rank-one matrix whose rows are equal to $\omega^\intercal$. Then:

$$\forall n \geq 1, \ \|P^n - W\|_\infty \leq 2\theta^{\frac{n}{r}-1}$$

where $\theta := 1 - \sigma$.

Proof. We write $P^r = (1-\theta)W + \theta Q$ where Q is a stochastic matrix. Note that $WP^k = W$ for all $k \geq 0$ since $\omega^\intercal = \omega^\intercal P$. Furthermore $MW = W$ for all stochastic matrices M since all rows of W are equal. Using these properties, we will show by induction that $P^{rk} = (1-\theta^k)W + \theta^k Q^k$. For $k = 1$ the result is trivial. Now suppose that $P^{rk} = (1-\theta^k)W + \theta^k Q^k$. Then

$$\begin{aligned}
P^{r(k+1)} &= P^{rk}P^r \\
&= [(1-\theta^k)W + \theta^k Q^k]P^r \\
&= (1-\theta^k)WP^r + (1-\theta)\theta^k Q^k W + \theta^{k+1}Q^{k+1} \\
&= (1-\theta^k)W + (1-\theta)\theta^k W + \theta^{k+1}Q^{k+1} \\
&= (1-\theta^{k+1})W + \theta^{k+1}Q^{k+1}.
\end{aligned}$$

Therefore the result holds for all $k \geq 1$. Therefore $P^{rk+j} - W = \theta^k(Q^k P^j - W)$ which implies that

$$\begin{aligned}
\forall n = rk + j \geq 1, \ \|P^n - W\|_\infty &\leq \theta^k \left\|Q^k P^j - W\right\|_\infty \\
&\leq 2\theta^k = 2\theta^{\lfloor \frac{n}{r} \rfloor} \leq 2\theta^{\frac{n}{r}-1}.
\end{aligned}$$

■

2.10 Geometric Ergodicity of C-Navigation

Since P_{π_u} is ergodic, there exists $r > 0$ such that $P^r_{\pi_u}(z, z') > 0$ for all z, z' (Proposition 1.7, (Levin et al., 2006)). For a stationary distribution vector ω and a state-action pair z,

we denote by $\omega(z)$ the component of ω corresponding to z. Moreover, we define

$$r := \min\{\ell \geq 1 : \ \forall (z,z') \in (\mathcal{S}\times\mathcal{A})^2, \ P^{\ell}_{\pi_u}(z,z') > 0\}, \tag{2.31}$$

$$\sigma_u := \min_{(z,z')\in(\mathcal{S}\times\mathcal{A})^2} \frac{P^r_{\pi_u}(z,z')}{\omega_u(z')}, \tag{2.32}$$

where ω_u is the stationary distribution of P_{π_u}.

Lemma 2.9 Let $\pi^o_t := \pi_{\omega^\star}(\widehat{\mathcal{M}}_t)$ (resp. $\overline{\pi}^t_{\omega^\star} := \sum_{j=1}^t \pi_{\omega^\star}(\widehat{\mathcal{M}}_j)/t$) denote the oracle policy of $\widehat{\mathcal{M}}_t$ (resp. the Cesaro-mean of oracle policies up to time t). Further define the functions

$$\sigma(\varepsilon,\pi,\omega) := \sigma_u\left(\varepsilon^r + \left[(1-\varepsilon)A\min_{s,a}\pi(a|s)\right]^r\right)\min_{z\in\mathcal{S}\times\mathcal{A}}\frac{\omega_u(z)}{\omega(z)},$$
$$\theta(\varepsilon,\pi,\omega) := 1-\sigma(\varepsilon,\pi,\omega),$$
$$\mathcal{L}(\varepsilon,\pi,\omega) := \frac{2}{\theta(\varepsilon,\pi,\omega)\left[1-\theta(\varepsilon,\pi,\omega)^{1/r}\right]}.$$

Then for C-Navigation it holds that

$$\forall n \geq 1, \ \|P^n_t - W_t\|_\infty \leq C_t\rho^n_t,$$

where $C_t := 2\theta(\varepsilon_t,\overline{\pi}^t_{\omega^\star},\omega_t)^{-1}$ and $\rho_t := \theta(\varepsilon_t,\overline{\pi}^t_{\omega^\star},\omega_t)^{1/r}$. In particular $C_t(1-\rho_t)^{-1} = \mathcal{L}(\varepsilon_t,\overline{\pi}^t_{\omega^\star},\omega_t)$.

Proof. Recall that: $P_t = \varepsilon_t P_{\pi_u} + (1-\varepsilon_t)P_{\overline{\pi}^t_{\omega^\star}}$. Therefore for all $(z,z') \in (\mathcal{S}\times\mathcal{A})^2$,

$$\begin{aligned}
P^r_t(z,z') &\geq [\varepsilon^r_t P^r_{\pi_u} + (1-\varepsilon_t)^r P^r_{\overline{\pi}^t_{\omega^\star}}](z,z') \\
&\overset{(a)}{\geq} \left(\varepsilon^r_t + \left[(1-\varepsilon_t)A\min_{s,a}\overline{\pi}^t_{\omega^\star}(a|s)\right]^r\right)P^r_{\pi_u}(z,z') \\
&\overset{(b)}{\geq} \left(\varepsilon^r_t + \left[(1-\varepsilon_t)A\min_{s,a}\overline{\pi}^t_{\omega^\star}(a|s)\right]^r\right)\sigma_u\omega_u(z') \\
&\geq \underbrace{\left(\varepsilon^r_t + \left[(1-\varepsilon_t)A\min_{s,a}\overline{\pi}^t_{\omega^\star}(a|s)\right]^r\right)\sigma_u\left(\min_z\frac{\omega_u(z)}{\omega_t(z)}\right)}_{:=\sigma_t}\omega_t(z') \\
&= \sigma(\varepsilon_t,\overline{\pi}^t_{\omega^\star},\omega_t)\omega_t(z').
\end{aligned}$$

where (a) comes from the fact that $P_{\overline{\pi}^t_{\omega^\star}} \geq A\min_{s,a}\overline{\pi}^t_{\omega^\star}(a|s)P_{\pi_u}$ entry-wise and (b) is due to (2.32). Using Lemma 2.8 we conclude that or all $n \geq 1$

$$\|P^n_t - W_t\|_\infty \leq 2\theta(\varepsilon_t,\overline{\pi}^t_{\omega^\star},\omega_t)^{\frac{n}{r}-1},$$

where $\theta(\varepsilon_t,\overline{\pi}^t_{\omega^\star},\omega_t) = 1-\sigma(\varepsilon_t,\overline{\pi}^t_{\omega^\star},\omega_t)$. Therefore P_t satisfies $\|P^n_t - W_t\|_\infty \leq C_t\rho^n_t$ for $C_t = 2\theta(\varepsilon_t,\overline{\pi}^t_{\omega^\star},\omega_t)^{-1}$ and $\rho_t = \theta(\varepsilon_t,\overline{\pi}^t_{\omega^\star},\omega_t)^{1/r}$. ∎

2.11 Simplifed Expression of the Generalized Likelihood Ratio

Proof. Observe that by definition $\widehat{\mathcal{M}}_t$ is the MDP that maximizes the likelihood of observations. Hence, we have

$$\begin{aligned}
\mathrm{GLR}(t;\widehat{\pi}_t^\star) &:= \log \frac{\sup\limits_{\mathcal{M}'\in\mathfrak{M}_{\star,1} \text{ s.t: } \pi^\star(\mathcal{M}')=\widehat{\pi}_t^\star} \ell_{\mathcal{M}'}(t)}{\sup\limits_{\mathcal{M}'\in\mathfrak{M}_{\star,1} \text{ s.t: } \pi^\star(\mathcal{M}')\neq\widehat{\pi}_t^\star} \ell_{\mathcal{M}'}(t)} \\
&= \log \frac{\ell_{\widehat{\mathcal{M}}_t}(t)}{\sup\limits_{\mathcal{M}'\in\mathrm{Alt}(\widehat{\mathcal{M}}_t)} \ell_{\mathcal{M}'}(t)} \\
&= \inf_{\mathcal{M}'\in\mathrm{Alt}(\widehat{\mathcal{M}}_t)} \log \frac{\ell_{\widehat{\mathcal{M}}_t}(t)}{\ell_{\mathcal{M}'}(t)}. \qquad (2.33)
\end{aligned}$$

Now we simplify the expression of the likelihhod ratio,

$$\begin{aligned}
\log \frac{\ell_{\widehat{\mathcal{M}}_t}(t)}{\ell_{\mathcal{M}'}(t)} &= \sum_{k=1}^{t-1} \left[\log \frac{q_{\widehat{\mathcal{M}}_t}(R_k|s_k,a_k)}{q_{\mathcal{M}'}(R_k|s_k,a_k)} + \log \frac{p_{\widehat{\mathcal{M}}_t}(s_{k+1}|s_k,a_k)}{p_{\mathcal{M}'}(s_{k+1}|s_k,a_k)} \right] \\
&= \sum_{k=1}^{t-1} \log \frac{q_{\widehat{\mathcal{M}}_t}(R_k|s_k,a_k)}{q_{\mathcal{M}'}(R_k|s_k,a_k)} + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{s'\in\mathcal{S}} N_{sas'}(t) \log \frac{p_{\widehat{\mathcal{M}}_t}(s'|s,a)}{p_{\mathcal{M}'}(s'|s,a)} \\
&\overset{(a)}{=} \sum_{k=1}^{t-1} \log \frac{q_{\widehat{\mathcal{M}}_t}(R_k|s_k,a_k)}{q_{\mathcal{M}'}(R_k|s_k,a_k)} + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} N_{sa}(t) \left[\sum_{s'\in\mathcal{S}} p_{\widehat{\mathcal{M}}_t}(s'|s,a) \log \frac{p_{\widehat{\mathcal{M}}_t}(s'|s,a)}{p_{\mathcal{M}'}(s'|s,a)} \right] \\
&= \sum_{k=1}^{t-1} \log \frac{q_{\widehat{\mathcal{M}}_t}(R_k|s_k,a_k)}{q_{\mathcal{M}'}(R_k|s_k,a_k)} + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} N_{sa}(t) \mathrm{KL}\big(\widehat{p}_{s,a}(t), p_{\mathcal{M}'}(s,a)\big) \\
&\overset{(b)}{=} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} N_{sa}(t) \big[\mathrm{KL}\big(\widehat{q}_{s,a}(t), q_{\mathcal{M}'}(s,a)\big) + \mathrm{KL}\big(\widehat{p}_{s,a}(t), p_{\mathcal{M}'}(s,a)\big) \big] \qquad (2.34)
\end{aligned}$$

where $N_{sas'}(t) := \sum_{k=1}^{t} \mathbb{1}\left(s_k = s, a_k = a, s_{k+1} = s'\right)$ is the number of times we observed the transition $(s,a) \to s'$ up to time step t, (a) uses that $p_{\widehat{\mathcal{M}}_t}(s'|s,a) = N_{sas'}(t)/N_{sa}(t)$ and (b) uses Lemma A.2 from (Degenne et al., 2020). Therfore, combining (2.33) and (2.34) we get that

$$\begin{aligned}
\mathrm{GLR}(t;\widehat{\pi}_t^\star) &= \inf_{\mathcal{M}'\in\mathrm{Alt}(\widehat{\mathcal{M}}_t)} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} N_{sa}(t) \big[\mathrm{KL}\big(\widehat{q}_{s,a}(t), q_{\mathcal{M}'}(s,a)\big) + \mathrm{KL}\big(\widehat{p}_{s,a}(t), p_{\mathcal{M}'}(s,a)\big) \big] \\
&= t\, T\big(\widehat{\mathcal{M}}_t, \mathbf{N}(t)/t\big)^{-1}
\end{aligned}$$

where the last inequality uses the definition of $(\mathcal{M},\omega) \mapsto T(\mathcal{M},\omega)$. ■

3. Active Coverage and Reward-Free Exploration in Episodic MDPs

In this chapter, we present and study the problem of active coverage. In particular, we design an algorithm, CovGame, that efficiently solves this problem. Then, we will show how an almost plug-and-play version of CovGame solves RFE with an *instance-dependent* complexity. The contents of this chapter are based on the conference paper:

Aymen Al Marjani, Andrea Tirinzoni, and Emilie Kaufmann. **Active Coverage for PAC Reinforcement Learning**. In *Proceedings of the 36th Conference On Learning Theory* (COLT), 2023.

Contents

3.1 Background on Coverage and RFE

The quality of the available data, whether it is actively gathered through *online* interactions with the environment or provided as a fixed *offline* dataset, plays a fundamental role in characterizing the performance of any reinforcement learning (RL, Sutton & Barto, 2018) agent. An important concept to quantify such quality is *coverage*, a property measuring the extent to which data spreads across the state-action space. The notion of coverage, through the so-called *concentrability coefficients*, is ubiquitous in the vast literature on offline RL (e.g., Munos, 2003; Munos & Szepesvári, 2008; Farahmand et al., 2009; Farahmand et al., 2010; Chen & Jiang, 2019; Xie & Jiang, 2020; Xie & Jiang, 2021; Jin et al., 2021; Foster et al., 2022). Intuitively, the better data covers the state space, the better performance one can expect from an offline RL method. Recently, (Xie et al., 2022) showed that a similar phenomenon also occurs in online RL: the sole existence of a good covering data distribution implies sample-efficient online RL with non-linear function approximation, even if such a distribution is unknown and inaccessible by the agent.

While these works treat coverage as a property of some *given* data or environment, a large body of literature focuses on *actively* collecting good covering data. This falls under the umbrella of *reward-free exploration* (RFE, Jin et al., 2020), a setting where the agent interacts with an unknown environment without any reward feedback. The objective is to collect sufficient data to enable the computation of a near-optimal policy for any reward function provided at downstream, e.g., by planning on top of an estimated model of the environment. Many provably-efficient algorithms exist for this problem that mostly differ in their exploration strategy. Some try to gather a minimum number of samples from each reachable state (Jin et al., 2020; Zhang et al., 2021c), while others adaptively optimize a reward function proportional to their uncertainty over the environment (Kaufmann et al., 2021; Ménard et al., 2021) or more simply a zero reward (Chen et al., 2022). All these approaches provably guarantee that the collected data is sufficient to learn any reward function provided at test time. Another popular technique is to seek data distributions that maximize the entropy over the state-space (Hazan et al., 2019; Cheung, 2019; Zahavy et al., 2021; Mutti et al., 2022). Finally, there is a long recent line of empirical works focusing on RFE, where the problem is often called *unsupervised RL* (e.g., Laskin et al., 2021; Eysenbach et al., 2019; Burda et al., 2019; Yarats et al., 2021).

The RFE literature mostly focuses on collecting data with the *specific* properties needed for the task under consideration (e.g., achieving zero-shot RL at test time). Motivated by the crucial role of coverage in RL, in this chapter we treat the problem at a higher level of generality. We formulate and study the problem of *active coverage* in episodic MDPs, where the goal is to interact online with the environment so as to collect data that satisfies some given coverage constraints. Following (Tarbouriech et al., 2021) who considered a similar problem in reset-free MDPs, we formalize such constraints as a set of sampling requirements that the learner must fulfill during learning. This gives our framework a high flexibility, as one can require different notions of coverage simply by changing the sampling requirements. Moreover, the applications are numerous, as any active coverage algorithm yields an exploration strategy that can be readily plugged in to tackle different problems. In our specific case, we shall see in this chapter how to apply it to design an algorithm for RFE. Then, in Chapter 4, we will present an algorithm for ε-BPI based on our solution to the coverage problem.

3.2 Definition of Active Coverage

3.2.1 Preliminaries

We consider the setting of Episodic MDPs, see Section 1.3.2. Denoting by $\mathbb{P}^\pi$ (resp. $\mathbb{E}^\pi$) the probability (resp. expectation) operator induced by the execution of a policy $\pi \in \Pi_S$ for an episode on $\mathcal{M}$, we define, for each (h, s, a), $p_h^\pi(s, a) := \mathbb{P}^\pi(s_h = s, a_h = a)$ and $p_h^\pi(s) := \mathbb{P}^\pi(s_h = s)$. We let $\Omega(\mathcal{M}) := \{[p_h^\pi(s, a)]_{h\in[H],s\in\mathcal{S},a\in\mathcal{A}} : \pi \in \Pi_S\}$ denote the set of all valid state-action distributions. It is well known (e.g., Puterman, 1994) that $\Omega(\mathcal{M})$ is convex and that

$$\Omega(\mathcal{M}) = \left\{\rho \in \mathbb{R}_+^{SAH} : \sum_{a\in\mathcal{A}} \rho_1(s, a) = 1, \sum_a \rho_h(s, a) = \sum_{(s',a')} \rho_{h-1}(s', a')p_{h-1}(s|s', a') \; \forall (h, s)\right\}.$$

We also recall that from every vector $\rho \in \Omega$ we can extract the corresponding policy π^ρ by normalization:

$$\forall (h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}, \;\; \pi_h^\rho(a|s) := \begin{cases} \dfrac{\rho_h(s, a)}{\sum_{b\in\mathcal{A}} \rho_h(s, b)} & \text{if } \sum_{b\in\mathcal{A}} \rho_h(s, b) > 0, \\ 1/A & \text{otherwise.} \end{cases} \tag{3.1}$$

Throughout the paper, we use $\mathbb{1}_\mathcal{X}$ to denote an indicator function over some set $\mathcal{X}$, i.e., $\mathbb{1}_\mathcal{X}(h, s, a) := \mathbb{1}\{(h, s, a) \in \mathcal{X}\}$ for all h, s, a. We shall hide $\mathcal{X}$ whenever $\mathcal{X} = [H] \times \mathcal{S} \times \mathcal{A}$. We make the following assumption to ensure that the whole state-space can be navigated.

Assumption 3.1 — Reachability. Each state $s \in \mathcal{S}$ is reachable at any stage $h \in \{2, \dots, H\}$ by some policy, i.e., $\max_{\pi\in\Pi_S} p_h^\pi(s) > 0$.

Reachability conditions like Assumption 3.1 are standard in prior work. In non-episodic reset-free MDPs (e.g., Jaksch et al., 2010), the MDP is often required to be communicating to ensure learnability, i.e., any two states are reachable from each other by some policy. Assumption 3.1 is the analogue for episodic MDPs, where we only need reachability from the initial state. In episodic MDPs, reachability conditions have been used in different settings, including model-free learning (Modi et al., 2021) and reward-free exploration (Zanette et al., 2020).

3.2.2 Learning problem

The learner interacts with an MDP $\mathcal{M}$ with unknown transition probabilities in order to fulfill some given *sampling requirements*. In particular, it is given a *target function* $c : [H] \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, where $c_h(s, a)$ denotes the minimum number of samples that must be gathered from (s, a) at stage h. In each episode of interaction $t \in \mathbf{N}^*$, the learner plays a policy $\pi^t \in \Pi_S$ and observes a corresponding trajectory $\{(s_h^t, a_h^t)\}_{h\in[H]}$. Let $n_h^t(s, a) := \sum_{j=1}^t \mathbb{1}(s_h^j = s, a_h^j = a)$ denote the number of times (s, a) has been visited at stage h up to episode t.

Definition 3.1 — (δ-correct c-coverage algorithm). Fix $\delta \in (0, 1)$ and a target function c. An algorithm is called δ-correct c-coverage if, with probability at least $1 - \delta$, it stops after interacting with $\mathcal{M}$ for a (possibly random) number of episodes τ and returns a

dataset of transitions with visitation counts guaranteeing

$$\forall (h,s,a),\ n_h^\tau(s,a) \geq c_h(s,a).$$

The goal in active coverage is to *minimize* the number of episodes required to collect at least $c_h(s,a)$ samples from each h,s,a with high probability.

Examples While the definition of the active coverage problem gives complete freedom in choosing the target function c, for our applications we shall mostly be interested in two specific instances. In *uniform coverage*, we have $c_h(s,a) = N\mathbb{1}\left((h,s,a) \in \mathcal{X}\right)$ for some given set $\mathcal{X}$ and $N \in \mathbf{N}$. Intuitively, this requires collecting at least N samples from each state-action-stage triplet in $\mathcal{X}$, and the name suggests that the learner should explore $\mathcal{X}$ as uniformly as possible. Possible applications include estimating the transition model uniformly well across the state-action space (Tarbouriech et al., 2020) and discovering sparse rewards. In our applications to PAC RL, we will further explore the benefits of performing *proportional coverage*, which corresponds to setting $c_h(s,a) = N \max_\pi p_h^\pi(s,a)\mathbb{1}\left((h,s,a) \in \mathcal{X}\right)$ [1]. This requires collecting a number of samples from each $(h,s,a) \in \mathcal{X}$ that scales proportionally to its reachability.

3.3 Lower Bound on the Complexity of Active Coverage

Minimizing the sample complexity needed to solve the active coverage problem requires the learner to properly plan how to distribute its exploration throughout the state-action space, hence accounting for the complex interplay between the MDP dynamics p and the target function c. The following theorem gives a precise characterization of the complexity of this problem. Its proof is deferred to Section 3.3.2

Theorem 3.1 For any target function c and $\delta \in (0,1)$, the stopping time τ of any δ-correct c-coverage algorithm satisfies $\mathbb{E}[\tau] \geq (1-\delta)\varphi^\star(c)$, where

$$\varphi^\star(c) = \inf_{\rho \in \Omega(\mathcal{M})} \max_{(s,a,h)\in\mathcal{X}} \frac{c_h(s,a)}{\rho_h(s,a)},$$

with $\mathcal{X} := \{(h,s,a) : c_h(s,a) > 0\}$.

The quantity $\varphi^\star(c)$ of Theorem 3.1 provides an *instance-dependent* complexity measure for the active coverage problem. In particular, it depends on both the MDP $\mathcal{M}$ through the set of valid state-action distributions $\Omega(\mathcal{M})$ and on the target function c. It can be interpreted as follows. Imagine that a learner repeatedly plays a policy that induces a state-action distribution $\rho \in \Omega(\mathcal{M})$. Then, for any (h,s,a), the quantity $1/\rho_h(s,a)$ is the expected number of episodes the learner takes to collect a single sample from (h,s,a). This implies that $\max_{(s,a,h)\in\mathcal{X}} \frac{c_h(s,a)}{\rho_h(s,a)}$ is roughly the expected number of episodes needed to satisfy the sampling requirements across all (h,s,a) when playing distribution ρ. Then, the complexity measure is intuitively the minimum of this quantity across all possible state-action distributions. In other words, any distribution $\rho^\star$ attaining the minimum in $\varphi^\star(c)$ denotes an *optimal* c-coverage distribution, i.e., generating data from $\rho^\star$ provably minimizes the time to satisfy all sampling requirements, in expectation.

Remark 3.1 Observe that the lower bound of Theorem 3.1 holds for any δ-correct algorithm, even for an oracle that knows the transition probabilities. In general, we do not believe it to be exactly matchable since (i) any algorithm must work with sample

[1] To cope with unknown transitions, we will use an upper bound of $p_h^\pi(s,a)$ in the definition of proportional coverage.

counts rather the expectations, (ii) the transition probabilities are unknown. However, $\varphi^\star(c)$ will appear as the leading order terms in the sample complexity of our algorithm, while these learning costs will be absorbed into lower order terms. ■

3.3.1 Links to other measures of coverage

3.3.1.1 Stochastic minimum flows

We begin by presenting an equivalent linear programming formulation of the optimal coverage problem of Theorem 3.1 that we call *stochastic minimum flow*. It is a direct extension to stochastic MDPs of the minimum flows for directed acyclic graphs in deterministic MDPs, which we presented in Section 1.6.3. We define a *flow* as a non-negative function $\eta : \mathcal{S} \times \mathcal{A} \times [H] \to [0, \infty)$ such that

$$\sum_{a \in \mathcal{A}} \eta_h(s,a) = \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p_{h-1}(s|s',a') \eta_{h-1}(s',a') \quad \forall s \in \mathcal{S}, h > 1, \tag{3.2}$$

$$\eta_1(s,a) = 0 \quad \forall s \in \mathcal{S} \setminus \{s_1\}, a \in \mathcal{A}. \tag{3.3}$$

That is, a flow η is a vector of visits to each state-action-stage triplet which satisfies the *navigation constraints* of the MDP. Note that the second constraint ensures that flow can only be created in the initial state s_1. The value of η is the total amount of flow leaving the initial state, i.e.,

$$\varphi(\eta) := \sum_{a \in \mathcal{A}} \eta_1(s_1, a).$$

We say that a flow η is *feasible* for a target function c if

$$\eta_h(s,a) \geq c_h(s,a) \quad \forall h \in [H], s \in \mathcal{S}, a \in \mathcal{A}.$$

The *stochastic minimum flow* problem consists in finding a feasible flow of minimum value. It can be clearly solved as a linear program,

$$\underset{\eta \in \mathbb{R}^{SAH}}{\text{minimize}} \sum_{a \in \mathcal{A}} \eta_1(s_1, a),$$

subject to

$$\begin{aligned} &\sum_{a \in \mathcal{A}} \eta_h(s,a) = \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} p_{h-1}(s|s',a') \eta_{h-1}(s',a') \quad \forall s \in \mathcal{S}, h > 1, \\ &\eta_1(s,a) = 0 \quad \forall s \in \mathcal{S} \setminus \{s_1\}, a \in \mathcal{A}, \\ &\eta_h(s,a) \geq c_h(s,a) \quad \forall h \in [H], s \in \mathcal{S}, a \in \mathcal{A}. \end{aligned} \tag{3.4}$$

We now prove that the optimal value of (3.4) is equal to $\varphi^\star(c)$, the optimal coverage complexity introduced in Section 3.3.

Lemma 3.1 If there exists a feasible flow for the target function c, the optimal value of (3.4) is exactly $\varphi^\star(c) = \min_{\rho \in \Omega(\mathcal{M})} \max_{h,s,a} \frac{c_h(s,a)}{\rho_h(s,a)}$.

Proof. Let us start from the linear programming formulation (3.4) and perform the change of variables $\rho_h(s,a) \leftarrow \frac{\eta_h(s,a)}{Z}$ and $Z \leftarrow \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \eta_h(s',a')$ for all h, s, a. Note that Z is the value of the original flow η (and thus it does not depend on the stage), while $\rho_h(s,a)$ is a probability distribution over the state-action space for each $h \in [H]$. We obtain the

following optimization problem (no longer a linear program due to the presence of a bilinear constraint):

$$
\begin{aligned}
&\underset{Z\geq 0,\rho\in\mathbb{R}^{SAH}}{\text{minimize}}\ Z,\\
&\text{subject to}\\
&\sum_{a\in\mathcal{A}}\rho_h(s,a)=\sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}}p_{h-1}(s|s',a')\rho_{h-1}(s',a') \quad \forall s\in\mathcal{S}, h>1,\\
&\rho_1(s,a)=0 \quad \forall s\in\mathcal{S}\setminus\{s_1\}, a\in\mathcal{A},\\
&\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\rho_h(s,a)=1 \quad \forall h\in[H],\\
&\rho_h(s,a)\geq 0 \quad \forall h\in[H], s\in\mathcal{S}, a\in\mathcal{A},\\
&Z\geq\frac{c_h(s,a)}{\rho_h(s,a)} \quad \forall h\in[H], s\in\mathcal{S}, a\in\mathcal{A}.
\end{aligned}
$$

The optimal solution for Z is clearly $Z=\max_{h,s,a}\frac{c_h(s,a)}{\rho_h(s,a)}$, while the first four constraints define exactly the set of valid state-action distributions $\Omega(\mathcal{M})$. This proves the statement. ■

3.3.2 Proof of Theorem 3.1

Define the coverage event $\mathcal{E}_{\text{cov}}=\big(\forall(h,s,a)\in\mathcal{X},\ n_h^\tau(s,a)\geq c_h(s,a)\big)$. We have that for any δ-correct algorithm $\mathbb{P}_{\mathcal{M},\mathcal{A}}\big(\mathcal{E}_{\text{cov}}\big)\geq 1-\delta$. Therefore, for any triplet $(h,s,a)\in\mathcal{X}$, we have that

$$
\mathbb{E}_{\mathcal{M},\mathcal{A}}[n_h^\tau(s,a)]\geq\mathbb{E}_{\mathcal{M},\mathcal{A}}[n_h^\tau(s,a)\mathbb{1}\left(\mathcal{E}_{\text{cov}}\right)]\geq c_h(s,a)\mathbb{P}_{\mathcal{M},\mathcal{A}}\big(\mathcal{E}_{\text{cov}}\big)\geq(1-\delta)c_h(s,a). \tag{3.5}
$$

Now consider the function $\eta_h(s,a):=\mathbb{E}_{\mathcal{M},\mathcal{A}}[n_h^\tau(s,a)]$ for all h,s,a. It is known that η satisfies the navigation constraints (3.2)[2]. Hence η is a flow vector. Moreover, it satisfies the constraint (3.5). By definition of stochastic minimum flow, this means that

$$
\mathbb{E}_{\mathcal{M},\mathcal{A}}[\tau]=\sum_{a\in\mathcal{A}}\mathbb{E}_{\mathcal{M},\mathcal{A}}[n_h^\tau(s_1,a)]=\varphi(\eta)\geq\varphi^\star\left([(1-\delta)c_h(s,a)]_{h,s,a}\right)=(1-\delta)\varphi^\star(c),
$$

where in the last line we used that for any constant $\alpha,\varphi^\star(\alpha c)=\alpha\varphi^\star(c)$. ■

3.3.3 Bounding the minimum flow

Lemma 3.2 Suppose there exists a feasible flow for the target function c. Then,

$$
\underbrace{\max_h\sum_{s,a}c_h(s,a)}_{\text{❶}}\leq\varphi^\star(c)\leq\underbrace{\sum_h\inf_{\rho\in\Omega}\max_{s,a}\frac{c_h(s,a)}{\rho_h(s,a)}}_{\text{❷}}\leq\underbrace{\sum_{h,s,a}\frac{c_h(s,a)}{\max_\pi p_h^\pi(s,a)}}_{\text{❸}}.
$$

Proof. The proof of the lower bound is trivial by noting that the value of any flow η can be written as $\varphi(\eta)=\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\eta_h(s,a)$ for all $h\in[H]$ and that any optimal flow satisfies $\eta_h^\star(s,a)\geq c_h(s,a)$ for all h,s,a. Let us prove the upper bound. Fix any $h\in[H]$ and let ρ^h denote a solution to the optimization problem $\min_{\rho\in\Omega}\max_{s,a}\frac{c_h(s,a)}{\rho_h(s,a)}$.

Further define the mixed distribution $\tilde{\rho}:=\sum_{l=1}^H\frac{Z_l}{Z}\rho^l$, where $Z_l:=\min_{\rho\in\Omega}\max_{s,a}\frac{c_h(s,a)}{\rho_h(s,a)}$ and $Z:=\sum_{l=1}^H Z_l$. Then, $\tilde{\rho}\in\Omega(\mathcal{M})$ is a convex combination of state-action distributions.

[2]For instance, by following the same steps in our proof of (1.41) and adapting it to episodic MDPs

Hence,

$$\begin{aligned}\varphi^\star(c) &\le \max_{h,s,a} \frac{c_h(s,a)}{\tilde{\rho}_h(s,a)} \\ &\overset{(a)}{\le} \max_h \frac{Z}{Z_h} \max_{s,a} \frac{c_h(s,a)}{\rho^h_h(s,a)} \\ &= \max_h \frac{Z}{Z_h} \min_{\rho\in\Omega} \max_{s,a} \frac{c_h(s,a)}{\rho_h(s,a)} = Z = \sum_{h\in[H]} \min_{\rho\in\Omega} \max_{s,a} \frac{c_h(s,a)}{\rho_h(s,a)},\end{aligned}$$

where (a) uses that $\tilde{\rho} \ge \frac{Z_h}{Z}\rho^h$ entry-wise. For the second upper bound, we define $w_h(s,a) := \frac{c_h(s,a)}{\max_{\pi\in\Pi} p^\pi_h(s,a)}$, with the convention that $w(s,a) = 0$ if $c_h(s,a) = 0$ regardless of the value of the denominator [3]. For any reachable (s,a,h), let $\pi_{s,a,h} \in \arg\max_{\pi\in\Pi_{\mathrm{D}}} p^\pi_h(s,a)$. For any unreachable (s,a,h), let $\pi_{s,a,h}$ be an arbitrary deterministic policy. Let us define the following mixed state-action distribution:

$$\forall(h,s,a): \quad \tilde{p}_h(s,a) := \sum_{s'\in\mathcal{S}} \sum_{a'\in\mathcal{A}} \frac{w_h(s',a')}{Z} p_h^{\pi_{s',a',l}}(s,a),$$

where $Z_h := \sum_{s'\in\mathcal{S}} \sum_{a'\in\mathcal{A}} w(s',a')$. Since this is a convex combination of state-action distributions, $\tilde{p} \in \Omega(\mathcal{M})$. Then,

$$\begin{aligned}\min_{\rho\in\Omega} \max_{s,a} \frac{c_h(s,a)}{\rho_h(s,a)} \le \max_{s,a} \frac{c_h(s,a)}{\tilde{p}_h(s,a)} &\le Z_h \max_{s,a} \frac{c_h(s,a)}{w_h(s,a) p_h^{\pi_{s,a,h}}(s,a)} \\ &= Z_h \max_{s,a} \frac{c_h(s,a)}{w_h(s,a) \sup_{\pi\in\Pi_{\mathrm{D}}} p^\pi_h(s,a)} \\ &= \sum_{h\in[H]} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \frac{c_h(s,a)}{\max_\pi p^\pi_h(s,a)}.\end{aligned}$$

■

Interestingly, each of the terms in the lemma above relates to a complexity measure that appeared in previous works. The term ❶ is the complexity for covering a tree-based deterministic MDP (Tirinzoni et al., 2022), perhaps the easiest MDP topology to navigate. As $\varphi^\star(c)$ reduces to the complexity of (Tirinzoni et al., 2022) in deterministic MDPs, we attain the equality $\varphi^\star(c) =$ ❶ in this specific tree structure. For a specific choice of c, ❷ can be shown to be exactly the "gap visitation" complexity measure introduced by (Wagenmaker et al., 2022a) for ε-BPI. As a component of their ε-BPI algorithm MOCA, (Wagenmaker et al., 2022a) introduced Learn2Explore, a strategy that learns policies to reach all states in the MDP. While it may be possible to adapt Learn2Explore for our active coverage problem, one limitation is that it learns how to reach each layer independently, and this is reflected in the fact that ❷ is only a loose upper bound (up to a factor H larger) to the optimal complexity $\varphi^\star(c)$. Finally, ❸ can be related to the sample complexity for active coverage obtained by the GOSPRL algorithm of (Tarbouriech et al., 2021)[4]. It can be interpreted as the complexity for learning how to reach each h,s,a independently, which makes it an even looser upper bound to $\varphi^\star(c)$.

[3] Note that, if $\max_{\pi\in\Pi} p^\pi_h(s,a) = 0$, then (s,a,h) is unreachable and it must be that $c_h(s,a) = 0$ since we assumed the minimum flow problem to be feasible.

[4] Since (Tarbouriech et al., 2021) consider reset-free MDPs, their complexity actually scales as $\sum_{s,a} D_{s,a} c(s,a)$, where $D_{s,a}$ is the minimum expected time to reach s,a from any state. In episodic MDPs, the minimum expected number of episodes to reach some (h,s,a) is exactly $1/\max_\pi p^\pi_h(s,a)$, hence yielding ❸.

3.3.4 Concentrability and coverability

A definition of the *concentrability coefficient* for a distribution $\rho \in \mathcal{P}(\mathcal{S} \times \mathcal{A} \times [H])$ is $C_{\mathrm{conc}}(\rho) := \max_{s,a,h} \frac{\max_\pi p_h^\pi(s,a)}{\rho_h(s,a)}$. This plays a fundamental role in characterizing the efficiency of offline RL methods (see, e.g., (Chen & Jiang, 2019; Xie et al., 2022) and references therein). It is easy to see that $\varphi^\star(c) = \inf_{\rho \in \Omega} C_{\mathrm{conc}}(\rho)$ for the target function c of proportional coverage. That is, our coverage complexity is equivalent to the minimum concentrability coefficient achievable by any distribution generated by some stochastic policy. Under a similar perspective, (Xie et al., 2022) introduced the *coverability coefficient* $C_{\mathrm{cov}} := \inf_{\rho_1,\dots,\rho_H \in \mathcal{P}(\mathcal{X}\times\mathcal{A})} \max_{s,a,h} \frac{\max_\pi p_h^\pi(s,a)}{\rho_h(s,a)}$ to characterize to what extent the best data distribution covers all policies. Noting that the infimum is taken across all probability distributions rather than valid state-action distributions, the optimal data distribution in C_{cov} may not be attained by the execution of any stochastic policy. This means that C_{cov} is not a valid complexity measure for active coverage in general, and it reduces exactly to ❶ for proportional coverage (see their Lemma 3), i.e., to a loose lower bound on $\varphi^\star(c)$.

3.4 Near-Optimal Active Coverage by Solving Games

3.4.1 Intuition and pseudo-code of CovGame

We propose CovGame (Algorithm 9), which adopts a game-based perspective inspired by the bandit literature (Degenne et al., 2019b). We first observe that the complexity $\varphi^\star(c)$ can be interpreted as a zero-sum game between a learner trying to produce the best sampling distribution $\rho \in \Omega(\mathcal{M})$ and an adversary trying to challenge it with the tuple (h,s,a) whose sampling requirement is the hardest to meet under ρ. CovGame does not directly solve the game in the definition of $\varphi^\star(c)$ but rather an equivalent formulation that simplifies learning. Recall that $\mathcal{P}(\mathcal{X})$ denotes the set of probability distributions with support in cX. Thanks to the min-max theorem, we can write

$$\begin{aligned}
\frac{1}{\varphi^\star(c)} = \sup_{\rho\in\Omega(\mathcal{M})} \min_{(s,a,h)\in\mathcal{X}} \frac{\rho_h(s,a)}{c_h(s,a)} &= \sup_{\rho\in\Omega(\mathcal{M})} \inf_{\lambda\in\mathcal{P}(\mathcal{X})} \sum_{(h,s,a)\in\mathcal{X}} \lambda_h(s,a)\frac{\rho_h(s,a)}{c_h(s,a)} \\
&= \inf_{\lambda\in\mathcal{P}(\mathcal{X})} \sup_{\rho\in\Omega(\mathcal{M})} \sum_{(h,s,a)\in\mathcal{X}} \lambda_h(s,a)\frac{\rho_h(s,a)}{c_h(s,a)} \\
&= \inf_{\lambda\in\mathcal{P}(\mathcal{X})} \max_{\pi\in\Pi_{\mathrm{D}}} \sum_{(h,s,a)\in\mathcal{X}} p_h^\pi(s,a)\frac{\lambda_h(s,a)}{c_h(s,a)},
\end{aligned}$$

where in the last equation we used that the inner maximization is a standard RL problem with reward function given by $\frac{\lambda_h(s,a)}{c_h(s,a)}\mathbb{1}\left((h,s,a)\in\mathcal{X}\right)$ and its optimum is known to be attained by a deterministic policy (e.g., Puterman, 1994).

CovGame solves a variant of this min-max game that does not involve the target function c directly. The idea is to cluster the state-action pairs in $\mathcal{X}$ based on their sampling requirement. To this end, we define the sequence of sets $\{\mathcal{X}_k\}_{k\in\mathbf{N}}$ as $\mathcal{X}_0 := \mathcal{X}$ and $\mathcal{X}_k := \{(h,s,a) : c_h(s,a) > c_{\min}^+ 2^k\}$ for all $k \in \mathbb{N}^*$, where $c_{\min}^+ = \min_{(h,s,a)\in\mathcal{X}} c_h(s,a) \vee 1$. At each round $t \in \mathbf{N}^*$, CovGame tries to solve the game $\inf_{\lambda\in\mathcal{P}(\mathcal{X}_{k_t})} \max_{\pi\in\Pi_{\mathrm{D}}} \sum_{h,s,a} p_h^\pi(s,a)\lambda_h(s,a)$, where k_t is the largest index such that all state-action pairs in $\mathcal{X}\setminus\mathcal{X}_{k_t} = \{(h,s,a)\in\mathcal{X} : c_h(s,a) \le c_{\min}^+ 2^{k_t}\}$ have been already covered. Intuitively, CovGame progressively focuses on covering state-action pairs with larger sampling requirements, while ignoring those that have already been covered. The main advantage over solving the initial formulation of $\varphi^\star(c)$ is two-fold. First, the learner is allowed to play only deterministic policies, each being the solution to an RL problem.

Algorithm 9 CovGame

Input: Target function c, RL algorithm $\mathcal{A}^{\Pi}$, online learning algorithm $\mathcal{A}^{\lambda}$, risk $\delta \in (0,1)$.

Let $\mathcal{X}_0 := \mathcal{X}$ and $\mathcal{X}_k := \{(h,s,a) : c_h(s,a) > c^+_{\min} 2^k\}$ for all $k \in \mathbb{N}^*$
Initialize counts $n^0_h(s,a) = 0$ for all h,s,a
Reset $\mathcal{A}^{\lambda}$ on $\mathcal{P}(\mathcal{X})$, set $\lambda^1_h(s,a) \leftarrow \mathbb{1}((h,s,a) \in \mathcal{X})/|\mathcal{X}|$ for all h,s,a
Initialize $k_1 \leftarrow 0$
for $t = 1, 2, \dots$ **do**
  Get π^t from $\mathcal{A}^{\Pi}$ given reward function λ^t and confidence $1 - \delta/2$
  Generate a trajectory $\{(s^t_h, a^t_h)\}_{h \in [H]}$ using policy π^t and update counts n^t
  if $n^t_h(s,a) \geq c_h(s,a)$ for all h,s,a **then**
    Stop and return all sampled trajectories
  Update $k_{t+1} \leftarrow \max\{j \in \mathbb{N} : n^t_h(s,a) \geq c_h(s,a)\ \forall (h,s,a) \in \mathcal{X} \setminus \mathcal{X}_j\}$
  if $k_{t+1} \neq k_t$ **then**
    Reset $\mathcal{A}^{\lambda}$ on $\mathcal{P}(\mathcal{X}_{k_{t+1}})$, set $\lambda^{t+1}_h(s,a) \leftarrow \mathbb{1}((h,s,a) \in \mathcal{X}_{k_{t+1}})/|\mathcal{X}_{k_{t+1}}|$ for all h,s,a
  else
    Feed $\mathcal{A}^{\lambda}$ with loss $\ell^t(\lambda) = \sum_{(h,s,a) \in \mathcal{X}_{k_t}} \lambda_h(s,a) \mathbb{1}(s^t_h = s, a^t_h = a)$, get weight λ^{t+1}

Second, in the sequence of games that we consider, the objective function is independent of the scale of c, which avoids undesired dependencies (e.g., on the inverse of the minimum value of c) when the target function is unbalanced.

CovGame approximately solves the sequence of games above by leveraging two online learning algorithms, $\mathcal{A}^{\lambda}$ and $\mathcal{A}^{\Pi}$. The one for the adversary ($\mathcal{A}^{\lambda}$) can be any method for online convex optimization on the simplex with linear losses. The one for the learner ($\mathcal{A}^{\Pi}$) can be any regret minimizer for RL that handles reward functions changing at each round (but observed at the beginning of the round). A simple approach like UCBVI (Azar et al., 2017) can be adapted to this purpose.

The final intuition behind CovGame is quite simple: at each round t, the adversary produces a reward function λ^t supported over $\mathcal{X}_{k_t}$ (the current set to be covered) and the learner tries to find a good policy for maximizing it. This encourages the learner to visit uncovered state-action pairs, eventually meeting the sampling requirements.

3.4.2 Sample complexity of CovGame

In order to analyze the sample complexity of CovGame, we make the following assumption on the adopted online learning algorithms, which will be satisfied by our specific instance.

Assumption 3.2 — First-order regret. There exists a non-decreasing function $\mathcal{R}^{\lambda}(T)$ such that, if $\mathcal{A}^{\lambda}$ is instantiated on $\mathcal{P}(\mathcal{X}_k)$ for some k on a sequence of linear losses $\{\ell^t\}_{t \geq 1}$ bounded in $[0,1]$,

$$\forall T \in \mathbb{N}^*,\ \sum_{t=1}^{T} \ell^t(\lambda^t) - \min_{\lambda \in \Delta_{\mathcal{X}_k}} \sum_{t=1}^{T} \ell^t(\lambda) \leq \sqrt{\mathcal{R}^{\lambda}(T) \sum_{t=1}^{T} \ell^t(\lambda^t)} + \mathcal{R}^{\lambda}(T). \tag{3.6}$$

There exists a non-decreasing function $\mathcal{R}^{\Pi}_{\delta}(T)$ such that, if $\mathcal{A}^{\Pi}$ is run with confidence $1 - \delta$ on a sequence of rewards $\{\lambda^t\}_{t \geq 1}$ with $\lambda^t \in \mathcal{P}(\mathcal{X})$ for all t, with probability $1 - \delta$,

for all $T \in \mathbf{N}^*$,

$$\sum_{t=1}^{T} V_1^{\star}\left(s_1;\lambda^t\right) - \sum_{t=1}^{T} V_1^{\pi_t}\left(s_1;\lambda^t\right) \le \sqrt{\mathcal{R}_\delta^{\Pi}(T)\sum_{t=1}^{T} V_1^{\pi^t}\left(s_1;\lambda^t\right)} + \mathcal{R}_\delta^{\Pi}(T), \tag{3.7}$$

where $V_1^{\pi}(s_1;\lambda) := \sum_{h,s,a} p_h^{\pi}(s,a)\lambda_h(s,a)$ and $V_1^{\star}(s_1;\lambda) := \max_\pi V_1^{\pi}\left(s_1;\lambda\right)$.

Theorem 3.2 — Sample complexity of CovGame. Under Assumption 3.1 and 3.2, with probability at least $1-\delta$, CovGame satisfies $n_h^\tau(s,a) \ge c_h(s,a)$ for all h,s,a and its stopping time τ satisfies $\tau \le 64m\varphi^\star(c) + T_1$, with $m := \lceil \log_2(c_{\max}/c_{\min}^+)\rceil \vee 1$, $c_{\max} := \max_{h,s,a} c_h(s,a)$ and

$$T_1 = \inf\left\{T \in \mathbb{N}^* : \frac{T}{2} \ge m\varphi^\star(\mathbb{1}_{\mathcal{X}})\left(3\mathcal{R}_{\delta/2}^{\Pi}(T) + 12\mathcal{R}^{\lambda}(T) + 24\log(4T/\delta)\right) + 1\right\}.$$

Remark 3.2 While we require both learners to have first-order regret bounds (i.e., depending on the sum of observed losses), standard $\widetilde{O}(\sqrt{T})$ bounds can also be used at the cost of a larger second-order term T_1 in Theorem 3.2, from $T_1 = \widetilde{O}(\varphi^\star(\mathbb{1}_{\mathcal{X}}))$ as in our instantiation to $T_1 = \widetilde{O}(\varphi^\star(\mathbb{1}_{\mathcal{X}})^2)$. The key step in our proof is to show that first-order regret implies convergence to the value $\varphi^\star(c)$ of the game at a rate $\widetilde{O}(1/T)$ instead of the slower $\widetilde{O}(1/\sqrt{T})$ achieved with $\widetilde{O}(\sqrt{T})$ regret. As $\varphi^\star(\mathbb{1}_{\mathcal{X}})$ depends on the inverse visitation probabilities (see Theorem 3.1), this $\varphi^\star(\mathbb{1}_{\mathcal{X}})$ versus $\varphi^\star(\mathbb{1}_{\mathcal{X}})^2$ improvement will be crucial to avoid undesired scaling with these quantities in our applications to PAC RL. ■

3.4.2.1 Our instanciation

For $\mathcal{A}^\lambda$ we propose to use the weighted majority forecaster (WMF, Littlestone & Warmuth, 1994) with variance-dependent learning rate for which, for any sequence of losses bounded in $[0,1]$, we have by Theorem 5 of (Cesa-Bianchi et al., 2005) that Assumption 3.2 is satisfied with

$$\mathcal{R}^{\lambda}(T) = 16\log(SAH). \tag{3.8}$$

For $\mathcal{A}^\Pi$ we propose to use a variant of UCBVI (Azar et al., 2017) that can cope with varying reward functions. The idea is that since the reward function λ^t is revealed to $\mathcal{A}^\Pi$ at the beginning of round t, we can build an upper confidence bound $\overline{Q}_h^{t-1}(s,a;\lambda^t)$ to the optimal action-value function $Q_h^\star(s,a;\lambda^t)$ by estimating the transition probabilities with the data collected up to round $t-1$. Then, we play $\pi_h^t(s) = \arg\max_a \overline{Q}_h^{t-1}(s,a;\lambda^t)$, the greedy policy w.r.t. $\overline{Q}_h^{t-1}$. We build the UCBs by leveraging the "monotonic value propagation" trick from (Zhang et al., 2021d) and prove that Assumption 3.2 is satisfied with

$$\mathcal{R}_\delta^{\Pi}(T) = 65536SAH^2(\log(2SAH/\delta) + 6S)\log(T+1)^2. \tag{3.9}$$

See Appendix C of (Al-Marjani et al., 2023) for details. Notably, we managed to prove a similar first-order regret bound as the one derived by (Jin et al., 2020) for EULER (Zanette & Brunskill, 2019b) with a remarkably simple analysis, without using any correction factor in the bonuses, and with improved dependences on H (from H^4 to H^2) and δ (from $\log(1/\delta)^3$ to $\log(1/\delta)$).

As compared to the minimax regret rate (Azar et al., 2017), our resulting bound in (3.7) features a dependence on S instead of $\sqrt{S}$ in its leading-order term. This is the cost of

handling changing rewards, which prevents us from building tight UCBs as commonly done for a fixed reward function. Instead, we build UCBs that hold for all rewards simultaneously using techniques from reward-free exploration (Ménard et al., 2021), a setting where an extra dependence on S is unavoidable in the worst case (Jin et al., 2020). Time-varying rewards, albeit under a weaker notion of regret, have also been studied in an adversarial setting in which the reward λ^t is not revealed prior to round t (Rosenberg & Mansour, 2019).

Corollary 3.1 — Sample complexity of CovGame with WMF and UCBVI. With probability at least $1-\delta$, the stopping time of CovGame with $\mathcal{A}^\lambda$ =WMF and $\mathcal{A}^\Pi$ =UCBVI is bounded by

$$\tau \le 64 m \varphi^\star(c) + \widetilde{O}(m\varphi^\star(\mathbb{1}_\mathcal{X}) SAH^2(\log(1/\delta) + S)),$$

where $m := \lceil \log_2(c_{\max}/c^+_{\min}) \rceil \vee 1$ and $\widetilde{O}$ hides poly-logarithmic factors in $S, A, H, \varphi^\star(\mathbb{1}_\mathcal{X}), \log(1/\delta)$.

The second term in the bound above can be interpreted as the cost incurred for learning the optimal coverage complexity $\varphi^\star(c)$ under *unknown* transition probabilities p. Still, this *learning cost* depends at most logarithmically on the total sampling requirement $\|c\|_1 = \sum_{h,s,a} c_h(s,a)$. This implies that, for large $\|c\|_1$, this cost becomes negligible as compared to the first term and $\tau \le \widetilde{O}(\varphi^\star(c))$, which matches the lower bound of Theorem 3.1 up to numerical constants and logarithmic terms.

Remark 3.3 If the transition kernel p is known, by replacing UCBVI with the computation of the optimal policy w.r.t. to λ^t, we have $\mathcal{R}^\Pi_{\delta/2}(T) = 0$. In this case, we get a smaller additive cost $\widetilde{O}(m\varphi^\star(\mathbb{1}_\mathcal{X}) \log(SAH) \log(1/\delta))$ which is only due to the randomness in the collection of trajectories. ∎

3.4.3 Proof of Theorem 3.2

Note that, at the beginning of any round $t \ge 1$, the learner $\mathcal{A}^\lambda$ works over the simplex $\mathcal{P}(\mathcal{X}_{k_t})$, hence $\lambda^t \in \mathcal{P}(\mathcal{X}_{k_t})$. Let $\tau_0 := 1$ and, for $i \in [m]$, let τ_i be the round at the beginning of which k_t has changed for the i-th time (i.e., $k_{\tau_i} \neq k_{\tau_i - 1}$). Note that, for any $i \ge 0$ and $t \in \{\tau_i, \dots, \tau_{i+1} - 1\}$, $k_t = k_{\tau_i}$.

Lemma 3.3 Under Assumption 3.1 and 3.2, with probability at least $1-\delta$, for any $i \in \{0, \dots, m-1\}$,

$$\min_{(h,s,a) \in \mathcal{X}_{k_{\tau_i}}} n_h^{\tau_{i+1}-1}(s,a) \ge \frac{1}{8} \frac{\tau_{i+1} - \tau_i}{\varphi^\star(\mathbb{1}_{\mathcal{X}_{k_{\tau_i}}})} - \frac{3}{8}\mathcal{R}^\Pi_\delta(\tau_{i+1}) - \frac{3}{2}\mathcal{R}^\lambda(\tau_{i+1}) - 3\log(4\tau_{i+1}/\delta).$$

Proof. Take any $i \in \{0, \dots, m-1\}$. Note that

$$\begin{aligned}
\min_{(h,s,a)\in\mathcal{X}_{k_{\tau_i}}} n_h^{\tau_{i+1}-1}(s,a) &= \min_{(h,s,a)\in\mathcal{X}_{k_{\tau_i}}} \sum_{t=1}^{\tau_{i+1}-1} \mathbb{1}\left(s_h^t = s, a_h^t = a\right) && \text{(definition of counts)} \\
&= \min_{(h,s,a)\in\mathcal{X}_{k_{\tau_i}}} \sum_{j=0}^{i} \sum_{t=\tau_j}^{\tau_{j+1}-1} \mathbb{1}\left(s_h^t = s, a_h^t = a\right) \\
& && \text{(definition of } \{\tau_j\}_{j\geq 0}) \\
&\geq \sum_{j=0}^{i} \min_{(h,s,a)\in\mathcal{X}_{k_{\tau_j}}} \sum_{t=\tau_j}^{\tau_{j+1}-1} \mathbb{1}\left(s_h^t = s, a_h^t = a\right) \\
& && (\mathcal{X}_{k_{\tau_i}} \subseteq \mathcal{X}_{k_{\tau_j}} \text{ for all } j \leq i) \\
&= \sum_{j=0}^{i} \min_{\lambda\in\mathcal{P}(\mathcal{X}_{k_{\tau_j}})} \sum_{(h,s,a)\in\mathcal{X}_{k_{\tau_j}}} \lambda_h(s,a) \sum_{t=\tau_j}^{\tau_{j+1}-1} \mathbb{1}\left(s_h^t = s, a_h^t = a\right) \\
&= \sum_{j=0}^{i} \min_{\lambda\in\mathcal{P}(\mathcal{X}_{k_{\tau_j}})} \sum_{t=\tau_j}^{\tau_{j+1}-1} \ell^t(\lambda). && \text{(definition of } \ell_t(\lambda)) \\
&\geq \min_{\lambda\in\mathcal{P}(\mathcal{X}_{k_{\tau_i}})} \sum_{t=\tau_i}^{\tau_{i+1}-1} \ell^t(\lambda)
\end{aligned}$$

For each i, by the regret bound of the λ player (Assumption 3.2),

$$\begin{aligned}
\min_{\lambda\in\mathcal{P}(\mathcal{X}_{k_{\tau_i}})} \sum_{t=\tau_i}^{\tau_{i+1}-1} \ell^t(\lambda) &\geq \sum_{t=\tau_i}^{\tau_{i+1}-1} \ell^t(\lambda^t) - \sqrt{\mathcal{R}^\lambda(\tau_{i+1}-\tau_i) \sum_{t=\tau_i}^{\tau_{i+1}-1} \ell^t(\lambda^t)} - \mathcal{R}^\lambda(\tau_{i+1}-\tau_i) \\
&\overset{(a)}{\geq} \frac{1}{2} \sum_{t=\tau_j}^{\tau_{i+1}-1} \ell^t(\lambda^t) - \frac{3}{2}\mathcal{R}^\lambda(\tau_{i+1}-\tau_i) \\
&\overset{(b)}{\geq} \frac{1}{2} \sum_{t=\tau_j}^{\tau_{i+1}-1} \ell^t(\lambda^t) - \frac{3}{2}\mathcal{R}^\lambda(\tau_{i+1}),
\end{aligned}$$

where in (a) we used the AM-GM inequality $\sqrt{xy} \leq \frac{x+y}{2}$ for $x, y \geq 0$ and in (b) we used that $\mathcal{R}^\lambda(\tau_{i+1}-\tau_i) \leq \mathcal{R}^\lambda(\tau_{i+1})$ by monotonicity of $T \mapsto \mathcal{R}^\lambda(T)$. Let us now bound $\sum_{t=1}^{\tau_{i+1}-1} \ell^t(\lambda^t)$. Note that $\ell^t(\lambda^t) = \sum_{h,s,a} \lambda_h^t(s,a) \mathbb{1}\left(s_h^t = s, a_h^t = a\right)$ for all for all $t \in \{\tau_j, \dots, \tau_{j+1}-1\}$ since λ^t is equal to zero outside $\mathcal{X}_{k_{\tau_j}}$. Then,

$$\begin{aligned}
\sum_{t=1}^{\tau_{i+1}-1} \ell^t(\lambda^t) &= \sum_{t=1}^{\tau_{i+1}-1} \sum_{h,s,a} \lambda_h^t(s,a) \Big(\mathbb{1}\left(s_h^t = s, a_h^t = a\right) \pm p_h^{\pi^t}(s,a) \Big) \\
&= \sum_{t=1}^{\tau_{i+1}-1} V_1^{\pi_t}\left(s_1; \lambda^t\right) + \underbrace{\sum_{t=1}^{\tau_{i+1}-1} \sum_{h,s,a} \lambda_h^t(s,a) \Big(\mathbb{1}\left(s_h^t = s, a_h^t = a\right) - p_h^{\pi^t}(s,a) \Big)}_{:= M_{\tau_{i+1}-1}}.
\end{aligned}$$

Since both λ^t and π^t are $\mathcal{F}_{t-1}$-measurable, $M_{\tau_{i+1}-1}$ is a martingale with differences bounded by 1 in absolute value. Therefore, by Freedman's inequality (e.g., Lemma 26 of Papini

et al., 2021), with probability at least $1-\delta/2$,

$$\forall T \geq 1, \quad |M_T| \leq \sqrt{\sum_{t=1}^{T} V_t \times 4\log(4T/\delta) + 4\log(4T/\delta)} \leq \sqrt{\sum_{t=1}^{T} V_1^{\pi_t}\left(s_1;\lambda^t\right) \times 4\log(4T/\delta) + 4\log(4T/\delta)},$$

where we defined $V_t := \mathrm{Var}[\sum_{h,s,a} \lambda_h^t(s,a)\mathbb{1}\left(s_h^t = s, a_h^t = a\right) \mid \mathcal{F}_{t-1}]$ and used the simple bound $V_t \leq \mathbb{E}[\sum_{h,s,a} \lambda_h^t(s,a)\mathbb{1}\left(s_h^t = s, a_h^t = a\right) \mid \mathcal{F}_{t-1}] = V_1^{\pi_t}\left(s_1;\lambda^t\right)$, which holds since $\sum_{h,s,a} \lambda_h^t(s,a)\mathbb{1}\left(s_h^t = s, a_h^t = a\right) \leq 1$ almost surely by definition of λ^t. Plugging this into the initial decomposition of $\sum_{t=1}^{\tau_{i+1}-1} \ell^t(\lambda^t)$ and using the AM-GM inequality $\sqrt{xy} \leq \frac{x+y}{2}$ for $x, y \geq 0$,

$$\sum_{t=1}^{\tau_{i+1}-1} \ell^t(\lambda^t) \geq \sum_{t=1}^{\tau_{i+1}-1} V_1^{\pi_t}\left(s_1;\lambda^t\right) - \sqrt{\sum_{t=1}^{\tau_{i+1}-1} V_1^{\pi_t}\left(s_1;\lambda^t\right) \times 4\log(4\tau_{i+1}/\delta)} - 4\log(4\tau_{i+1}/\delta) \geq \frac{1}{2}\sum_{t=1}^{\tau_{i+1}-1} V_1^{\pi_t}\left(s_1;\lambda^t\right) - 6\log(4\tau_{i+1}/\delta).$$

We finally bound $\sum_{t=1}^{T} V_1^{\pi_t}\left(s_1;\lambda^t\right)$ for any T. For all $T \geq 1$, with probability at least $1-\delta/2$ from Assumption 3.2,

$$\sum_{t=1}^{T} V_1^{\pi_t}\left(s_1;\lambda^t\right) \geq \sum_{t=1}^{T} V_1^{\star}\left(s_1;\lambda^t\right) - \sqrt{\mathcal{R}_\delta^{\Pi}(T)\sum_{t=1}^{T} V_1^{\star}\left(s_1;\lambda^t\right)} - \mathcal{R}_\delta^{\Pi}(T).$$

Applying once again the AM-GM inequality yields

$$\sum_{t=1}^{T} V_1^{\pi_t}\left(s_1;\lambda^t\right) \geq \frac{1}{2}\sum_{t=1}^{T} V_1^{\star}\left(s_1;\lambda^t\right) - \frac{3}{2}\mathcal{R}_\delta^{\Pi}(T) = \frac{1}{2}\sum_{t=1}^{T}\sup_{\rho\in\Omega}\sum_{h,s,a}\rho_h(s,a)\lambda_h^t(s,a) - \frac{3}{2}\mathcal{R}_\delta^{\Pi}(T).$$

Now note that, since λ^t is supported on $\mathcal{X}_{k_{\tau_j}}$ for any $t \in \{\tau_j, \ldots, \tau_{j+1}-1\}$,

$$\begin{aligned}\sum_{t=1}^{\tau_{i+1}-1}\sup_{\rho\in\Omega}\sum_{h,s,a}\rho_h(s,a)\lambda_h^t(s,a) &= \sum_{j=0}^{i}\sum_{t=\tau_j}^{\tau_{j+1}-1}\sup_{\rho\in\Omega}\sum_{h,s,a}\rho_h(s,a)\lambda_h^t(s,a)\\ &\geq \sum_{j=0}^{i}\sum_{t=\tau_j}^{\tau_{j+1}-1}\sup_{\rho\in\Omega}\min_{(h,s,a)\in\mathcal{X}_{k_{\tau_j}}}\rho_h(s,a)\\ &= \sum_{j=0}^{i}\frac{\tau_{j+1}-\tau_j}{\varphi^\star(\mathbb{1}_{\mathcal{X}_{k_{\tau_j}}})}\\ &\geq \frac{\tau_{i+1}-\tau_i}{\varphi^\star(\mathbb{1}_{\mathcal{X}_{k_{\tau_i}}})}.\end{aligned}$$

Plugging everything together proves the statement.

■

Let m denote the number of times k_t changes value through the execution of the algorithm, that is $m = |\{t \le \tau : k_t \neq k_{t+1}\}|$. We provide a bound on m.

Lemma 3.4 It holds that $m \le \lceil \log_2(c_{\max}/c^+_{\min}) \rceil \vee 1$. Moreover, for any $i \in \{0, \dots, m-1\}$, we have $\min_{(h,s,a)\in\mathcal{X}_{k_{\tau_i}}} n_h^{\tau_{i+1}-1}(s,a) \le c^+_{\min} 2^{k_{\tau_i}+2}$.

Proof. By definition of the update rule, we have that $k_{t+1} \ge k_t$ for all $t \ge 1$. Now take any time t in which k_t has changed value m times. Since $k_1 \ge 0$, this means that $k_t \ge m$. By definition of k_t, we know that $n_h^{t-1}(s,a) \ge c_h(s,a)$ for all $(h,s,a) \in \mathcal{X} \setminus \mathcal{X}_j$ for some $j \ge m$. However, if $m \ge \lceil \log_2(c_{\max}/c^+_{\min}) \rceil \vee 1$, $\mathcal{X}_j = \emptyset$ and thus the algorithm must have stopped. This proves that $m \le \lceil \log_2(c_{\max}/c^+_{\min}) \rceil \vee 1$.

To prove the second statement, we note that for any $i < m$, we have $k_{\tau_{i+1}-1} = k_{\tau_i}$ and $n_h^{\tau_{i+1}-2}(s,a) \ge c_h(s,a)$ for all $(h,s,a) \in \mathcal{X} \setminus \mathcal{X}_{k_{\tau_i}}$. Moreover, there must be some $(h,s,a) \in \mathcal{X} \setminus \mathcal{X}_{k_{\tau_i}+1}$ such that $n_h^{\tau_{i+1}-2}(s,a) < c_h(s,a)$. Indeed, if this was not the case, we would have an update of k at the end of round $\tau_{i+1}-2$ instead of $\tau_{i+1}-1$. Since all the triplets in $\mathcal{X}_{k_{\tau_i}}$ have been covered, the uncovered triplet must be in $\mathcal{X}_{k_{\tau_i}} \cap \mathcal{X} \setminus \mathcal{X}_{k_{\tau_i}+1} = \mathcal{X}_{k_{\tau_i}} \setminus \mathcal{X}_{k_{\tau_i}+1}$. By definition, all $(h,s,a) \in \mathcal{X}_{k_{\tau_i}} \setminus \mathcal{X}_{k_{\tau_i}+1}$ satisfy $c_h(s,a) \le c^+_{\min} 2^{k_{\tau_i}+1}$. Hence,

$$\min_{(h,s,a)\in\mathcal{X}_{k_{\tau_i}}} n_h^{\tau_{i+1}-1}(s,a) \le \min_{(h,s,a)\in\mathcal{X}_{k_{\tau_i}}} n_h^{\tau_{i+1}-2}(s,a) + 1 < c^+_{\min} 2^{k_{\tau_i}+1} + 1 \le c^+_{\min} 2^{k_{\tau_i}+2}$$

where we use that $c^+_{\min} \ge 1$. ∎

We are now ready to prove Theorem 3.2

Proof of Theorem 3.2. Let m be the number of times k_t has changed throughout the execution of the algorithm. Note that, in the round τ in which the algorithm stops the last change must occur, thus $\tau_m = \tau + 1$, and $k_{\tau+1}$ is set to any value such that $\mathcal{X}_{k_{\tau+1}} = \emptyset$. Then,

$$\tau = \tau_m - 1 = \sum_{i=0}^{m-1} (\tau_{i+1} - \tau_i)\,.$$

By combining Lemma 3.4 with Lemma 3.3 and rearranging, with probability at least $1-\delta$, for any $i \in \{0, \dots, m-1\}$,

$$\begin{aligned}\tau_{i+1} - \tau_i &\le 8\varphi^\star(\mathbb{1}_{\mathcal{X}_{k_{\tau_i}}}) c^+_{\min} 2^{k_{\tau_i}+2} + 8\varphi^\star(\mathbb{1}_{\mathcal{X}_{k_{\tau_i}}}) \left(\frac{3}{8}\mathcal{R}^\Pi_\delta(\tau_{i+1}) + \frac{3}{2}\mathcal{R}^\lambda(\tau_{i+1}) + 3\log(4\tau_{i+1}/\delta)\right) \\ &\le 8\varphi^\star(\mathbb{1}_{\mathcal{X}_{k_{\tau_i}}}) c^+_{\min} 2^{k_{\tau_i}+2} + \varphi^\star(\mathbb{1}_{\mathcal{X}}) \left(3\mathcal{R}^\Pi_\delta(\tau_m) + 12\mathcal{R}^\lambda(\tau_m) + 24\log(4\tau_m/\delta)\right),\end{aligned}$$

where the second inequality is due to $\mathcal{X}_k \subseteq \mathcal{X}$ for all $k \in \mathbb{N}$ and $\tau_{i+1} \le \tau_m$ for $i \le m-1$. Then,

$$\tau_m \le 8 \sum_{i=0}^{m-1} c^+_{\min} \varphi^\star(\mathbb{1}_{\mathcal{X}_{k_{\tau_i}}}) 2^{k_{\tau_i}+2} + m\varphi^\star(\mathbb{1}_{\mathcal{X}}) \left(3\mathcal{R}^\Pi_\delta(\tau_m) + 12\mathcal{R}^\lambda(\tau_m) + 24\log(4\tau_m/\delta)\right) + 1.$$

The first term can be bounded by

$$
\begin{aligned}
8\sum_{i=0}^{m-1} c_{\min}^{+}\varphi^{\star}(\mathbb{1}_{\mathcal{X}_{k_{\tau_i}}})2^{k_{\tau_i}+2} &= 8\sum_{i=0}^{m-1} c_{\min}^{+}2^{k_{\tau_i}+2}\min_{\rho\in\Omega}\max_{s,a,h}\frac{\mathbb{1}((h,s,a)\in\mathcal{X}_{k_{\tau_i}})}{\rho_h(s,a)} \\
&\le 32\sum_{i=0}^{m-1} c_{\min}^{+}2^{k_{\tau_i}}\min_{\rho\in\Omega}\max_{s,a,h}\frac{\mathbb{1}(c_{\min}^{+}2^{k_{\tau_i}} < c_h(s,a))}{\rho_h(s,a)} \\
&\le 32\sum_{i=0}^{m-1}\min_{\rho\in\Omega}\max_{s,a,h}\frac{c_h(s,a)}{\rho_h(s,a)} = 32m\varphi^{\star}(c).
\end{aligned}
$$

Plugging this into the bound on τ_m, we obtain the inequality,

$$
\tau_m \le 32m\varphi^{\star}(c) + m\varphi^{\star}(\mathbb{1}_{\mathcal{X}})\left(3\mathcal{R}_{\delta}^{\Pi}(\tau_m) + 12\mathcal{R}^{\lambda}(\tau_m) + 24\log(4\tau_m/\delta)\right) + 1.
$$

Thus, for $\tau_m \ge T_1$, we get that the sample complexity is bounded by $\tau \le 64m\varphi^{\star}(c)$. Thus, we conclude that $\tau \le \tau_m \le \max\{T_1, 64m\varphi^{\star}(c)\} \le 64m\varphi^{\star}(c) + T_1$. The proof is concluded by using Lemma 3.4 to bound m. ■

3.4.4 Comparison with prior work

While inspired by an original game perspective which is crucial in our analysis, the actual algorithmic approach of CovGame has a similar flavor as existing algorithms for different exploration tasks: it runs a regret minimizer on different reward functions enforcing the visitation of uncovered states. Using WMF as the λ-learner, the reward function in round t is

$$
\lambda_h^{t+1}(s,a) = \frac{\exp\left(-\xi_{t-i_t}\left(n_h^t(s,a) - n_h^{i_t}(s,a)\right)\right)\mathbb{1}\left((h,s,a)\in\mathcal{X}_{k_t}\right)}{\sum_{(h',s',a')\in\mathcal{X}_{k_t}}\exp\left(-\xi_{t-i_t}\left(n_{h'}^t(s',a') - n_{h'}^{i_t}(s',a')\right)\right)},
$$

where i_t is the last restart of WMF that happened before t and ξ_t is the variance-dependent learning rate defined by (Cesa-Bianchi et al., 2005). Our reward function is related to the number of prior visits and smoothly evolves over time, which is in contrast with most prior approaches that rely on rewards of the form $r_h^{\mathcal{Y}}(s,a) = \mathbb{1}((h,s,a)\in\mathcal{Y})$ for some set $\mathcal{Y}$, For example, GOSPRL translated to our episodic setting would use $r_h^{t+1}(s,a) = \mathbb{1}\left(n_h^t(s,a) < c_h^t(s,a)\right)$. The Learn2Explore strategy (Wagenmaker et al., 2022a) uses a subroutine to visit N times some of the state-action pairs in $\mathcal{Y}$: it runs EULER (Zanette & Brunskill, 2019a) on $r^{\mathcal{Y}}$ and restarts the algorithm with a reward function with reduced support whenever some new state-action pair has reached N visits. Several algorithms for RFE (Jin et al., 2020; Zhang et al., 2021a) also collect data using regret minimizers on top of indicator-based rewards.

3.5 Application to Reward-Free Exploration

A strategy for RFE should return an estimate of the transition kernel $\widehat{p}$ from which a planning agent can compute a near-optimal policy for any reward function. To be robust to any possible reward in the test phase, we intuitively need to gather sufficient samples everywhere in the MDP, which we propose to do explicitly by relying on CovGame with proportional coverage (Section 3.5.1). The resulting algorithm is called Proportional Coverage Exploration (PCE). PCE takes as input two parameters ε, δ and returns an estimate of the transition probabilities $\widehat{p}$ that, with probability $1-\delta$, yields an ε-optimal policy for any reward function bounded in $[0,1]$.

3.5.1 PCE: Intuition and pseudo-code

The first observation in the design of PCE is that, it does not really matter which action the planner plays at the pairs $(h,s) \in [H] \times \mathcal{S}$ that are hard to reach. More precisely, denoting by $V_1^\pi(s_1;r) := \sum_{h,s,a} p_h^\pi(s,a) r_h(s,a)$ the expected return of π under the reward function r, it holds that

$$\forall \pi \in \Pi_{\mathrm{D}}, \quad \sum_{(h,s):\ \sup_\pi p_h^\pi(s) \le \varepsilon/2SH} \ \sum_{a \in \mathcal{A}} p_h^\pi(s,a) r_h(s,a) \le \varepsilon/2.$$

In other words, even if the planner selects sub-optimal actions in the step-state pairs (h,s) such that $\sup_\pi p_h^\pi(s) \le \varepsilon/2SH$, she will at most incur a loss of $\varepsilon/2$ in the value function. Therefore, we do not need to explore states whose *reacheability* is low.

This leads us to the second ingredient which motivates the choice of proportional coverage: a novel *ellipsoid-shaped confidence region* for the value functions of all policies under any reward. Let $\widehat{p}^t$ denote the maximum likelihood estimator of p after observing t episodes. Denote by $\widehat{V}_1^{\pi,t}(s_1;r) := \sum_{h,s,a} \widehat{p}_h^{\pi,t}(s,a) r_h(s,a)$ the expected return of π in the empirical MDP with transitions $\widehat{p}^t$ and reward function r. Theorem 3.4 in Appendix 3.9 gives that, with probability $1-\delta$, jointly over all episodes t,

$$\forall r \in [0,1]^{SAH},\ \forall \pi \in \Pi^D,\quad \left|V_1^\pi(s_1;r) - \widehat{V}_1^{\pi,t}(s_1;r)\right| \le \sqrt{\beta^{\mathrm{RF}}(t,\delta) \sum_{(h,s,a)\in\mathcal{X}_\varepsilon} \frac{p_h^\pi(s,a)^2}{n_h^t(s,a)}} + \frac{\varepsilon}{4}, \tag{3.10}$$

where $\beta^{\mathrm{RF}}(t,\delta) \propto H^2 \log(1/\delta) + SH^3 \log(A(1+t))$ and $\mathcal{X}_\varepsilon$ is a subset of triplets that are not too hard to reach: $\mathcal{X}_\varepsilon \subseteq \{(h,s,a) : \max_\pi p_h^\pi(s,a) \ge \frac{\varepsilon}{4SH^2}\}$. Hence, if we gather $c_h(s,a) = \mathcal{O}(H\beta^{\mathrm{RF}}(t,\delta)\sup_\pi p_h^\pi(s,a)/\varepsilon^2)$ visits from every $(h,s,a) \in \mathcal{X}_\varepsilon$, the confidence interval above will satisfy

$$\begin{aligned}\sqrt{\beta^{\mathrm{RF}}(t,\delta) \sum_{(h,s,a)\in\mathcal{X}_\varepsilon} \frac{p_h^\pi(s,a)^2}{n_h^t(s,a)}} &\le \varepsilon \sqrt{\beta^{\mathrm{RF}}(t,\delta) \sum_{(h,s,a)\in\mathcal{X}_\varepsilon} \frac{p_h^\pi(s,a)^2}{c_1 H \beta^{\mathrm{RF}}(t,\delta) \sup_\pi p_h^\pi(s,a)}} \\ &\le \varepsilon \sqrt{\frac{\sum_{(h,s,a)\in\mathcal{X}_\varepsilon} p_h^\pi(s,a)}{c_1 H}} \le \varepsilon/c_1,\end{aligned}$$

for some constant $c_1 > 0$. The last inequality above is due to the fact that for each step h, the probabilities $(p_h^\pi(s,a))_{s,a}$ sum to one. Hence, for a good choice of c_1, the estimation error of $V_1^\pi(s_1;r)$ for any π and r will be below $\varepsilon/2$, which was demonstrated to be sufficient for solving RFE (Jin et al., 2020).

Yet as the visitation probabilities are unknown, neither $\mathcal{X}_\varepsilon$ nor $c_h(s,a)$ can actually be computed. To solve this issue, we rely on an initialization phase based on the EstimateReachability subroutine (line 2 of Algorithm 10), described in Appendix 3.10. This procedure, which is similar to the initialization phase in MOCA (Wagenmaker et al., 2022a), outputs for each (h,s) an interval $[\underline{W}_h(s), \overline{W}_h(s)]$ to which $\max_\pi p_h^\pi(s)$ belongs with high probability using a low-order number of episodes of $\widetilde{O}(S^3AH^4/\varepsilon)$. The lower confidence bound is then used to build a set $\widehat{\mathcal{X}}$ that satisfies the requirements for $\mathcal{X}_\varepsilon$ and the upper bound is used to define the target function that is given as input to CovGame in phase k of the algorithm: $c_h^k(s,a) := 2^k \overline{W}_h(s) \mathbb{1}\big((h,s,a) \in \widehat{\mathcal{X}}\big)$. The pseudo-code of PCE is presented in Algorithm 10.

Algorithm 10 PCE (Proportional Coverage Exploration)

Input: Precision ε, Risk $\delta \in (0,1)$.
For each (h,s), run EstimateReachability$((h,s); \frac{\varepsilon}{4SH^2}, \frac{\delta}{3SH})$ to get confidence intervals $\left[\underline{W}_h(s), \overline{W}_h(s)\right]$ on $\max_\pi p_h^\pi(s)$ (see Appendix 3.10)
Define $\widehat{\mathcal{X}} := \{(h,s,a) : \underline{W}_h(s) \geq \frac{\varepsilon}{32SH^2}\}$
Define target function $c_h^0(s,a) = \mathbb{1}\big((h,s,a) \in \widehat{\mathcal{X}}\big)$ for all (h,s,a)
Execute CovGame$\big(c^0,\ \delta/6\big)$ to get a dataset $\mathcal{D}_0$ of d_0 episodes // Burn-in phase
Initialize episode count $t_0 \leftarrow d_0$ and statistics $n_h^0(s,a), \widehat{p}_h^0(.|s,a)$ using $\mathcal{D}_0$
for $k = 1, \ldots$ **do**
  // Proportional Coverage
  Compute targets $c_h^k(s,a) := 2^k \overline{W}_h(s) \mathbb{1}\big((h,s,a) \in \widehat{\mathcal{X}}\big)$ for all (h,s,a)
  Execute CovGame$\big(c^k,\ \delta/6(k+1)^2\big)$ to get dataset $\mathcal{D}_k$ and number of episodes d_k
  Update episode count $t_k \leftarrow t_{k-1} + d_k$ and statistics $n_h^k(s,a), \widehat{p}_h^k(.|s,a)$ using $\mathcal{D}_k$
  if $\sqrt{H\beta^{RF}(t_k, \delta/3)2^{4-k}} \leq \varepsilon$ **then** stop and return $\widehat{p}^k$
end for

Remark 3.4 We remark that PCE is computationally efficient as it inherits the complexity of CovGame and EstimateReachability, both of which require solving one dynamic program in every round to compute the optimistic policy used by UCBVI. We now present its theoretical properties. ■

Remark 3.5 — Reachability. Thanks to its initialization phase, PCE can be used even when Assumption 3.1 is violated. All triplets that have zero probability to be reached are filtered out from the set $\widehat{\mathcal{X}}$ (line 3 of Algorithm 10), and CovGame always targets reachable states. ■

3.5.2 Sample Complexity of PCE

Theorem 3.3 Let $\widehat{p}$ be the estimate of the transition probabilities that PCE outputs. For any reward function r, let $\hat{\pi}_r$ be an optimal policy in the MDP $(\widehat{p}, r)$. Then,

$$\mathbb{P}\left(\forall r \in [0,1]^{SAH}, |V_1^{\hat{\pi}_r}(s_1; r) - V_1^\star(s_1; r)| \leq \varepsilon\right) \geq 1 - \delta.$$

Furthermore, with probability at least $1-\delta$, the total sample complexity of PCE satisfies

$$\tau \leq \widetilde{\mathcal{O}}\Bigg((H^3\log(1/\delta) + SH^4)\varphi^\star\Bigg(\left[\frac{\sup_\pi p_h^\pi(s)\mathbb{1}(\sup_\pi p_h^\pi(s) \geq \frac{\varepsilon}{32SH^2})}{\varepsilon^2}\right]_{h,s,a}\Bigg) + \frac{S^3A^2H^5(\log(1/\delta)+S)}{\varepsilon}\Bigg),$$

where $\widetilde{\mathcal{O}}$ hides poly-logarithmic factors in $S, A, H, 1/\varepsilon$ and $\log(1/\delta)$.

Perhaps the most interesting feature of this bound is that thanks to Lemma 3.2, the $\varphi^\star$ term is at most SAH/ε^2. As a result, a worst-case bound can be directly extracted from Theorem 3.3:

$$\tau = \widetilde{O}\left(\frac{SAH^4}{\varepsilon^2}\log(1/\delta) + \frac{S^2AH^5}{\varepsilon^2} + \frac{S^3A^2H^5}{\varepsilon}(\log(1/\delta)+S)\right).$$

Given that the minimax rate of RFE is of order $\Omega\big(\frac{SAH^3\log(1/\delta)+S^2AH^3}{\varepsilon^2}\big)$ (Jin et al., 2020; Kaufmann et al., 2021), we conclude that PCE is *minimax-optimal* up to an H^2 factor

and low-order terms scaling in $1/\varepsilon$. More interestingly, the next Lemmas provide benign MDP instances where the complexity of PCE can be much smaller than the minimax rate in terms of the dependence on the number of states.

3.5.3 Adaptive reward-free exploration

We define the simplified complexity

$$\mathrm{PCE}(\mathcal{M}, \varepsilon) := \varphi^\star([\sup_\pi p_h^\pi(s,a)]_{h,s,a})/\varepsilon^2, \tag{3.11}$$

which is an upper bound on the $\varphi^\star$ term of Theorem 3.3. We start by considering the case where the MDP is actually a contextual bandit, but this fact is unknown to the learner.

Lemma 3.5 — Disguised contextual bandits. Suppose that $\mathcal{M}$ is a "disguised" contextual bandit, i.e.,

$$\forall (h, s, a, s'),\ p_h(s'|s,a) = p_h(s'|s).$$

Then $\mathrm{PCE}(\mathcal{M}, \varepsilon) = A/\varepsilon^2$.

Plugging the Lemma above into Theorem 3.3, we get a reduced sample complexity for PCE of order:

$$\tau = \widetilde{O}\left(\frac{AH^3}{\varepsilon^2}\log(1/\delta) + \frac{SAH^4}{\varepsilon^2} + \frac{S^3A^2H^5}{\varepsilon}(\log(1/\delta) + S)\right).$$

For ε small enough, the term in $1/\varepsilon$ becomes negligible and we save an S factor compared to the minimax rate.

Proof. In this case for any (h, s) and any policy π, $p_h^\pi(s) = p_h(s)$ is independent of the policy. Thanks to the one-to-one correspondence between vectors in $\Omega(\mathcal{M})$ and Markovian stochastic policies (see Section 3.2.1) we may write

$$\begin{aligned}
\varphi^\star([\sup_\pi p_h^\pi(s,a)]_{h,s,a}) &= \inf_{\pi^{exp}\in\Pi^S} \max_{s,a,h} \frac{\sup_\pi p_h^\pi(s,a)}{p_h^{\pi^{exp}}(s,a)} \\
&= \inf_{\pi^{exp}\in\Pi^S} \max_{s,a,h} \frac{p_h(s)\sup_\pi \pi_h(a|s)}{p_h(s)\pi_h^{exp}(a|s)} \\
&= \inf_{\pi^{exp}\in\Pi^S} \max_{s,h} \frac{1}{\min_a \pi_h^{exp}(a|s)} \\
&= A,
\end{aligned}$$

where the last equality is because $(\min_a \pi_h^{exp}(a|s))^{-1} \geq A$ and the infimum over Π^S is achieved by the uniform policy. ∎

3.5.3.1 Ergodic MDPs

Let $\alpha, \beta \in (0,1)$ such that $\alpha > \beta$. Further, define the set of probability vectors such that

$$\mathcal{P}_{\alpha,\beta} = \left\{ q \in \mathbb{R}_+^S : \sum_{i=1}^S q_i = 1,\ \max_i q_i \leq S^{\alpha-1}, \min_i q_i \geq \frac{1 - S^{\beta-1}}{S-1} \right\}.$$

Note that such set is never empty since the vector $(S^{\beta-1}, \frac{1-S^{\beta-1}}{S-1}, \ldots, \frac{1-S^{\beta-1}}{S-1})$ always satisfies the inequalities in its definition. We define the class of MDPs $\mathfrak{M}_{erg}$ such that their transition kernel satisfies

$$\forall (h, s, a),\ p_h(.|s,a) \in \mathcal{P}_{\alpha,\beta}.$$

Lemma 3.6 Assume that $\mathcal{M} \in \mathfrak{M}_{erg}$, then $\mathrm{PCE}(\mathcal{M}, \varepsilon) \leq S^\alpha AH/\varepsilon^2$.

Therefore when the MDP is ergodic, the sample complexity of PCE is at most

$$\tau = \widetilde{O}\left(\frac{S^\alpha AH^4}{\varepsilon^2}\log(1/\delta) + \frac{S^\alpha AH^5}{\varepsilon^2} + \frac{S^3A^2H^5}{\varepsilon}\left(\log(1/\delta) + S\right)\right),$$

where $\alpha \in (0,1)$. In other words, we gain a $S^{1-\alpha}$ factor and the dependence on S is no longer quadratic despite the fact that we estimate a transition kernel p of dimension S^2AH.

Remark 3.6 Note that the "ergodicity" of MDPs in $\mathfrak{M}_{erg}$ can be as small as one wishes: by taking the limit $\beta \to 1$, the constraint $\min_{s'} p_h(s'|s,a) \geq \frac{1-S^{\beta-1}}{S-1}$ becomes vacuous so the MDP can be non-ergodic. In that regime, $\alpha = 1$ and we recover the minimax rate (up to an H factor) SAH^3/ε^2. ∎

Proof. First of all, we note that

$$\forall \pi \in \Pi_{\mathrm{D}}\ \forall s \in \mathcal{S},\ p_h^\pi(s) = \sum_{s' \in \mathcal{S}} p_{h-1}^\pi(s) p_h(s|s', \pi_{h-1}(s')) \leq \sum_{s' \in \mathcal{S}} p_{h-1}^\pi(s) S^{\alpha-1} = S^{\alpha-1}. \tag{3.12}$$

Similarly

$$\forall \pi \in \Pi_{\mathrm{D}}\ \forall s \in \mathcal{S},\ p_h^\pi(s) \geq \frac{1 - S^{\beta-1}}{S-1}. \tag{3.13}$$

Now using Lemma 3.2 we have that

$$\begin{aligned}
\varphi^\star([\sup_\pi p_h^\pi(s,a)]_{h,s,a}) &\leq \sum_{h=1}^H \inf_{\pi^{exp} \in \Pi^S} \max_{s,a} \frac{\sup_\pi p_h^\pi(s,a)}{p_h^{\pi^{exp}}(s)\pi_h^{exp}(a|s)} \\
&\overset{(a)}{=} \sum_{h=1}^H \inf_{\pi^{exp} \in \Pi^S} \max_s \frac{1}{p_h^{\pi^{exp}}(s)} \times \inf_{(\pi_h^{exp}(.|s)) \in \mathcal{P}(\mathcal{A})} \max_a \frac{\sup_\pi p_h^\pi(s,a)}{\pi_h^{exp}(a|s)} \\
&\overset{(b)}{=} \sum_{h=1}^H \inf_{\pi^{exp} \in \Pi^S} \max_s \frac{1}{p_h^{\pi^{exp}}(s)} \times \Big(\sum_a \sup_\pi p_h^\pi(s,a)\Big) \\
&\overset{(c)}{=} \sum_{h=1}^H \inf_{\pi^{exp} \in \Pi^S} \max_s \frac{A \sup_\pi p_h^\pi(s)}{p_h^{\pi^{exp}}(s)} \\
&= A \sum_{h=1}^H \underbrace{\inf_{\pi^{exp} \in \Pi^S} \max_s \frac{\sup_\pi p_h^\pi(s)}{p_h^{\pi^{exp}}(s)}}_{:=\mathcal{C}_h},
\end{aligned} \tag{3.14}$$

where (a) uses that for $h \in [H]$ and any policy $\pi^{\exp}$, $(p_h^{\pi^{exp}}(s))_s$ and $(\pi_h^{\exp}(a|s))_{a,s}$ are independent[5], (b) solves the right-hand side minimization problem in $(\pi_h^{exp}(.|s))$ for a fixed state s and (c) uses that $\sup_\pi p_h^\pi(s,a) = \sup_\pi p_h^\pi(s)$ (the equality is achieved by playing the policy that maximizes $p_h^\pi(s)$ then playing action a at (h,s)). Now fix $h \in [H]$ and denote by π^s any policy in $\arg\max_{\pi \in \Pi_{\mathrm{D}}} p_h^\pi(s)$. Further define the stochastic policy $\widetilde{\pi}$ such that

$$p^{\widetilde{\pi}} = \frac{\sum_{s \in \mathcal{S}} p^{\pi^s}}{S}.$$

[5]The actions that a policy π plays at step h have no impact on the probabilities of reaching states $\mathbb{P}^\pi(s_h = s)$ at that step.

Using (3.13) we have that for all $s \in \mathcal{S}$,

$$p_h^{\widetilde{\pi}}(s) = \frac{\sum_{s' \in \mathcal{S}} p_h^{\pi^{s'}}(s)}{S} \geq \frac{\sup_{\pi \in \Pi} p_h^{\pi}(s) + (S-1)\frac{1-S^{\beta-1}}{S-1}}{S} = \frac{\sup_{\pi \in \Pi} p_h^{\pi}(s) + 1 - S^{\beta-1}}{S}. \tag{3.15}$$

Therefore

$$\begin{aligned}
\mathcal{C}_h = \inf_{\pi^{exp} \in \Pi^S} \max_s \frac{\sup_\pi p_h^\pi(s)}{p_h^{\pi^{exp}}(s)} \leq \max_s \frac{\sup_\pi p_h^\pi(s)}{p_h^{\widetilde{\pi}}(s)} &\overset{(a)}{\leq} \max_s \frac{S \sup_\pi p_h^\pi(s)}{\sup_{\pi \in \Pi_D} p_h^\pi(s) + 1 - S^{\beta-1}} \\
&= \max_s \frac{S}{1 + \frac{1-S^{\beta-1}}{\sup_\pi p_h^\pi(s)}} \\
&\overset{(b)}{\leq} \max_s \frac{S}{1 + S^{1-\alpha}(1 - S^{\beta-1})} \\
&= \frac{S}{1 + S^{1-\alpha} - S^{\beta-\alpha}} \leq S^\alpha,
\end{aligned}$$

where (a) uses (3.15) and (b) uses (3.12). Combining (3.14) with the previous inequality yields that $\varphi^\star([\sup_\pi p_h^\pi(s,a)]_{h,s,a}) \leq S^\alpha AH$. ■

The examples above suggest that, while RFE is by essence a worst-case problem where one has to be robust to any reward at test time, there is still hope to adapt to the "explorability" of the MDP.

3.6 Analysis of PCE

In this final section, we provide the full analysis leading to the proof of Theorem 3.4. To simplify the presentation of the algorithm and the analysis, we index the counts as well as the empirical estimates of transitions and rewards by their phase number. Hence, for each triplet (h,s,a), $n_h^k(s,a)$ and $\widehat{p}_h^k(.|s,a)$ will refer to the number of visits and the empirical transition kernel respectively after t_k episodes, i.e. at the end of the k-th phase. Finally, for a dataset of episodes $\mathcal{D}$, $n_h(s,a;\mathcal{D})$ denotes the number of visits of (h,s,a) in the episodes stored in $\mathcal{D}$.

Good event

We introduce the following events

$$\begin{aligned}
\mathcal{E}_{vis} := \Big(&\text{The set built using } \textsc{EstimateReachability}\left((h,s); \tfrac{\varepsilon}{4SH^2}, \tfrac{\delta}{3SH}\right) \text{ for all } (h,s) \\
&\text{satisfies } \left\{(h,s) : \sup_\pi p_h^\pi(s) \geq \tfrac{\varepsilon}{4SH^2}\right\} \subseteq \widehat{\mathcal{X}} \subseteq \left\{(h,s) : \sup_\pi p_h^\pi(s) \geq \tfrac{\varepsilon}{32SH^2}\right\} \\
&\text{and } \forall (h,s) \in \widehat{\mathcal{X}},\ \sup_\pi p_h^\pi(s) \leq \overline{W}_h(s) \leq 36 \sup_\pi p_h^\pi(s)\Big), \\
\mathcal{E}_p^{RF} := \Big(&\forall k \in \mathbb{N}^\star, \forall \pi \in \Pi^D, \forall r \in [0,1]^{SAH}, \\
&\left|\textstyle\sum_{s,a,h} \left(\widehat{p}_h^{\pi,k}(s,a) - p_h^\pi(s,a)\right) r_h(s,a)\right| \leq \sqrt{\beta^{RF}(t_k, \delta/3) \textstyle\sum_{(s,a,h) \in \widehat{\mathcal{X}}} \frac{p_h^\pi(s,a)^2}{n_h^k(s,a)}} + \tfrac{\varepsilon}{4}\Big), \\
\mathcal{E}_{cov} := \Big(&\forall k \in \mathbb{N},\ \text{CovGame run with inputs } (c^k, \delta/6(k+1)^2) \text{ terminates after at most} \\
&64 m_k \varphi^\star(c^k) + \widetilde{\mathcal{O}}\big(m_k \varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}}) SAH^2 (\log(6(k+1)^2/\delta) + S)\big) \text{ episodes and returns a dataset } \mathcal{D}_k \\
&\text{such that for all } (h,s,a) \in \widehat{\mathcal{X}}, n_h(s,a;\mathcal{D}_k) \geq c_h^k(s,a)\Big),
\end{aligned}$$

where $m_k = \log_2\left(\frac{\max_{s,a,h} c_h^k(s,a)}{\min_{s,a,h} c_h^k(s,a)\vee 1}\right)\vee 1$ and β^{RF} is defined in appendix 3.9.2. Then our good event is defined as the intersection

$$\mathcal{E}_{good}^{RF} := \mathcal{E}_{vis} \cap \mathcal{E}_p^{RF} \cap \mathcal{E}_{cov}.$$

Lemma 3.7 We have that $\mathbb{P}_{\mathcal{M}}(\mathcal{E}_{good}^{RF}) \geq 1-\delta$.

Proof. Let $\overline{\mathcal{E}}$ denote the complementary event of $\mathcal{E}$. We start by the following decomposition

$$\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{good}^{RF}}) \leq \mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{vis}}) + \mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{cov}}) + \mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_p^{RF}} \cap \mathcal{E}_{vis} \cap \mathcal{E}_{cov}).$$

Now we bound each term separately. First observe that applying Theorem 3.6 with parameter $\varepsilon_0 = \varepsilon/4SH^2$ yields $\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{vis}}) \leq \delta/3$. Second, using Corollary 3.1 we have

$$\begin{aligned}\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{cov}}) &\leq \sum_{k=0}^{\infty} \mathbb{P}_{\mathcal{M}}(\text{CovGame with inputs } (c^k, \delta/6(k+1)^2) \text{ fails}) \\ &\leq \sum_{k=0}^{\infty} \frac{\delta}{6(k+1)^2} = \frac{\delta\pi^2}{36} \leq \delta/3.\end{aligned}$$

Next, note that by design of PCE $n_h^0(s,a) = n_h(s,a;\widetilde{\mathcal{D}}_0)$ and $c^0 = \mathbb{1}_{\widehat{\mathcal{X}}}$ so that $\mathcal{E}_{cov} \subset \big(\forall (h,s,a)\in\widehat{\mathcal{X}},\ n_h^0(s,a)\geq 1\big)$. Therefore we have

$$\begin{aligned}\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_p^{RF}} \cap \mathcal{E}_{vis} \cap \mathcal{E}_{cov}) &\leq \mathbb{P}_{\mathcal{M}}\big(\overline{\mathcal{E}_p^{RF}},\ \big\{(h,s) : \sup_\pi p_h^\pi(s) \geq \frac{\varepsilon}{4SH^2}\big\} \subseteq \widehat{\mathcal{X}},\ \forall (h,s,a)\in\widehat{\mathcal{X}}\ n_h^0(s,a)\geq 1\big) \\ &= \mathbb{P}_{\mathcal{M}}\Bigg(\big\{(h,s) : \sup_\pi p_h^\pi(s) \geq \frac{\varepsilon}{4SH^2}\big\} \subseteq \widehat{\mathcal{X}},\ \exists k\geq 0\ \exists \pi\in\Pi^D\ \exists r\in[0,1]^{SAH} : \\ &\qquad \Big|\sum_{s,a,h}\big(\widehat{p}_h^{\pi,k}(s,a) - p_h^\pi(s,a)\big) r_h(s,a)\Big| > \sqrt{\beta^{RF}(t_k,\delta/3)\sum_{(s,a,h)\in\widehat{\mathcal{X}}}\frac{p_h^\pi(s,a)^2}{n_h^k(s,a)}} + \frac{\varepsilon}{4}\Bigg) \\ &\overset{(a)}{\leq} \mathbb{P}_{\mathcal{M}}\Bigg(\big\{(h,s) : \sup_\pi p_h^\pi(s) \geq \frac{\varepsilon}{4SH^2}\big\} \subseteq \widehat{\mathcal{X}},\ \exists t\geq t_0\ \exists \pi\in\Pi^D\ \exists r\in[0,1]^{SAH} : \\ &\qquad \Big|\sum_{s,a,h}\big(\widehat{p}_h^{\pi,t}(s,a) - p_h^\pi(s,a)\big) r_h(s,a)\Big| > \sqrt{\beta^{RF}(t,\delta/3)\sum_{(s,a,h)\in\widehat{\mathcal{X}}}\frac{p_h^\pi(s,a)^2}{n_h^t(s,a)}} + \frac{\varepsilon}{4}\Bigg) \\ &\overset{(b)}{\leq} \delta/3,\end{aligned}$$

where in (a) we introduced $t_0 = \inf\{t\geq 1 : n_h^t(s,a)\geq 1, \forall (h,s,a)\in\widehat{\mathcal{X}}\}$ and switched back to indexing counts and estimates by the episode number (instead of the phase) in order to apply Theorem 3.4 in (b) with $\mathcal{Z} = \{(h,s,a) : (h,s)\in\widehat{\mathcal{X}}\}$ and $\varepsilon_0 = \varepsilon/4SH^2$. Combining the four inequalities above yields the desired result. ■

Low concentrability / Good coverage of all policies

The next lemma shows that PCE achieves proportional coverage.

Lemma 3.8 Under the good event, for all phases $k\geq 0$, we have that

$$n_h^k(s,a) \geq 2^k \sup_\pi p_h^\pi(s,a) \quad \forall (h,s,a)\in\widehat{\mathcal{X}}.$$

Proof. First of all, note that for any triplet $(h,s,a) \in \widehat{\mathcal{X}}, \sup_\pi p_h^\pi(s,a)$ is always attained by some deterministic policy. Therefore, it is sufficient to prove that, given a fixed deterministic policy $\pi \in \Pi^D$,

$$\forall k \geq 0, \forall (h,s,a) \in \widehat{\mathcal{X}}, \quad n_h^k(s,a) \geq 2^k p_h^\pi(s,a) \ .$$

We do this by induction over k. For $k=0$ the result is trivial since, under the good event, we have that for all $(h,s,a) \in \widehat{\mathcal{X}}$, $n_h^0(s,a) \geq c_h^0(s,a) = 1 \geq 2^0 p_h^\pi(s,a)$. Now suppose that the property holds for phase k. Then under the good event we know that for all $(h,s,a), n_h^{k+1}(s,a) - n_h^k(s,a) = n_h(s,a,\mathcal{D}_{k+1}) \geq c_h^{k+1}(s,a)$. Plugging the definition of c^{k+1} (Line 9 of Algorithm 10) we get that for any $(h,s,a) \in \widehat{\mathcal{X}}$,

$$\begin{aligned} n_h^{k+1}(s,a) &\geq c_h^{k+1}(s,a) \\ &= 2^{k+1} \overline{W}_h(s) \\ &\geq 2^{k+1} \sup_\pi p_h^\pi(s) \\ &= 2^{k+1} \sup_\pi p_h^\pi(s,a), \end{aligned} \tag{3.16}$$

where the second inequality uses the event $\mathcal{E}_{vis}$. ■

Correctness

Lemma 3.9 Let $\widehat{p}$ be the estimate of the transition probabilities that PCE outputs. For any reward function r, let $\hat{\pi}_r$ be an optimal policy in the MDP $(\widehat{p}, r)$. Then

$$\mathbb{P}\left(\forall r \in [0,1]^{SAH}, V_1^{\hat{\pi}_r}(s_1;r) \geq V_1^\star(s_1;r) - \varepsilon\right) \geq 1-\delta.$$

In other words, PCE is (ε,δ)-PAC for reward-free exploration.

Proof. Assume that PCE stops as phase k and let $\widehat{p}^k$ denote the empirical transition estimates that it returns. Fix any reward function $r = [r_h(s,a)]_{h,s,a} \in [0,1]^{SAH}$ and let $\widehat{\pi} \in \arg\max_{\pi \in \Pi^D} (\widehat{p}^{\pi,k})^\top r$ be the policy obtained when planning for reward function r under the transition model $\widehat{p}^k$. Further define $\pi^\star \in \arg\max_{\pi \in \Pi^D} (p^\pi)^\top r, V_1^\star := (p^{\pi^\star})^\top r$, and $V_1^{\widehat{\pi}} := (p^{\widehat{\pi}})^\top r$. Note that both $\widehat{\pi}$ and $\pi^\star$ are deterministic. Therefore under the good event $\mathcal{E}_{good}^{RF}$ we have

$$\begin{aligned} V_1^{\widehat{\pi}} &= (p^{\widehat{\pi}})^\top r \\ &\overset{(a)}{\geq} (\widehat{p}^{\widehat{\pi},k})^\top r - \sqrt{\beta^{RF}(t_k,\delta/3) \sum_{(s,a,h)\in\widehat{\mathcal{X}}} \frac{p_h^{\widehat{\pi}}(s,a)^2}{n_h^k(s,a)}} - \frac{\varepsilon}{4} \\ &\overset{(b)}{\geq} (\widehat{p}^{\pi^\star,k})^\top r - \sqrt{\beta^{RF}(t_k,\delta/3) \sum_{(s,a,h)\in\widehat{\mathcal{X}}} \frac{p_h^{\widehat{\pi}}(s,a)^2}{n_h^k(s,a)}} - \frac{\varepsilon}{4} \\ &\overset{(c)}{\geq} (p^{\pi^\star})^\top r - \sqrt{\beta^{RF}(t_k,\delta/3) \sum_{(s,a,h)\in\widehat{\mathcal{X}}} \frac{p_h^{\pi^\star}(s,a)^2}{n_h^k(s,a)}} - \sqrt{\beta^{RF}(t_k,\delta/3) \sum_{(s,a,h)\in\widehat{\mathcal{X}}} \frac{p_h^{\widehat{\pi}}(s,a)^2}{n_h^k(s,a)}} - \frac{\varepsilon}{2} \\ &\overset{(d)}{\geq} V_1^\star - 2\sqrt{H\beta^{RF}(t_k,\delta/3)2^{-k}} - \frac{\varepsilon}{2} \\ &\overset{(e)}{\geq} V_1^\star - \varepsilon, \end{aligned}$$

where (a) and (c) use the good event $\mathcal{E}_p^{RF}$ for policies $\widehat{\pi}$ and $\pi^\star$ respectively, (b) uses the definition of $\widehat{\pi}$, (d) uses Lemma 3.8 and (e) uses the stopping condition of PCE (Line 10 in Algorithm 10). Note that the inequality above holds, under the good event $\mathcal{E}_{good}$, jointly for all reward functions r. Since $\mathbb{P}_{\mathcal{M}}(\mathcal{E}_{good}) \geq 1-\delta$, we have just proved that PCE is (ε, δ)-PAC for reward-free exploration. ■

Upper bound on the number of phases

Lemma 3.10 Define the index of the final phase of PCE,

$$\kappa_f := \inf\left\{k \in \mathbf{N}_+ : \sqrt{H\beta^{RF}(t_k, \delta/3)2^{4-k}} \leq \varepsilon\right\}.$$

Further let τ denote the number of episodes played by the algorithm. Then under the good event, it holds that $\kappa_f < \infty$ and

$$2^{\kappa_f} \leq \frac{32H\beta^{RF}(\tau, \delta/3)}{\varepsilon^2}.$$

Proof. First, we prove that κ_f is finite. Under the good event, we have

$$\begin{aligned} t_k &= \sum_{j=0}^{k} d_j \\ &\leq \sum_{j=0}^{k} \left[64 m_j \varphi^\star(c^j) + \widetilde{\mathcal{O}}\big(m_j \varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}}) SAH^2(\log(6(j+1)^2/\delta) + S)\big)\right], \end{aligned}$$

where we recall that $m_j = \log_2\left(\frac{\max_{s,a,h} c_h^j(s,a)}{\min_{s,a,h} c_h^j(s,a) \vee 1}\right) \vee 1$. Now using the fact that $c_h^j(s,a) \leq 2^j \mathbb{1}((h,s,a) \in \widehat{\mathcal{X}})$ for $j \geq 0$ we deduce that $m_0 = 1$ and $m_j \leq j \ \forall j \geq 1$ so that

$$\begin{aligned} t_k &\leq \sum_{j=0}^{k} \left[8(j+1)2^j \varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}}) + \widetilde{\mathcal{O}}\big((j+1)\varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}}) SAH^2(\log(4(j+1)^2/\delta) + S)\big)\right] \\ &= \mathcal{O}_{k\to\infty}\big(k^2 2^k\big). \end{aligned} \tag{3.17}$$

Now recall that the threshold β^{RF} was defined in Appendix 3.9 as

$$\beta^{RF}(t, \delta) := 4H^2 \log(1/\delta) + 24SH^3 \log(A(1+t)) \tag{3.18}$$

Combining (3.17) and (3.18) gives that

$$\beta^{RF}(t_k, \delta/3) = o_{k\to\infty}\big(2^k\big).$$

Therefore $\kappa_f = \inf\left\{k \in \mathbf{N}_+ : \sqrt{H\beta^{RF}(t_k, \delta/3)2^{4-k}} \leq \varepsilon\right\}$ is indeed finite. The proof of the second statement is straightforward by noting that $\kappa_f - 1$ does not satisfy the stopping condition (Line 12 in Algorithm 10) and using the (crude) upper bound $t_{\kappa_f - 1} \leq \tau$. ■

Upper bound on the phase length

Lemma 3.11 Let $k \geq 1$ be such that PCE did not stop before phase k. Under the good

event, the number of episodes played by PCE during phase k satisfies

$$d_k \leq c_1 k H \beta^{RF}(\tau, \delta/3)\varphi^\star\left(\left[\frac{\sup_\pi p_h^\pi(s)\mathbb{1}\left(\sup_\pi p_h^\pi(s) \geq \frac{\varepsilon}{32SH^2}\right)}{\varepsilon^2}\right]_{h,s,a}\right) + \widetilde{\mathcal{O}}\left(k\frac{S^3A^2H^5(\log(6(k+1)^2/\delta)+S)}{\varepsilon}\right),$$

where $c_1 = 73728$. Furthermore, the duration of the initial phase is upper-bounded as

$$d_0 \leq \widetilde{\mathcal{O}}\left(\frac{S^3A^2H^5(\log(6/\delta)+S)}{\varepsilon}\right).$$

Proof. Using the good event and the definition of c^k we write

$$\begin{aligned} d_k &\leq 64 m_k \varphi^\star\left(\left[2^k \overline{W}_h(s)\mathbb{1}\left((h,s,a)\in\widehat{\mathcal{X}}\right)\right]_{h,s,a}\right) + \widetilde{\mathcal{O}}\big(m_k\varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}})SAH^2(\log(6(k+1)^2/\delta)+S)\big) \\ &\overset{(a)}{\leq} 64k\varphi^\star\left(\left[2^k\overline{W}_h(s)\mathbb{1}\left((h,s,a)\in\widehat{\mathcal{X}}\right)\right]_{h,s,a}\right) + \widetilde{\mathcal{O}}\big(k\varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}})SAH^2(\log(6(k+1)^2/\delta)+S)\big), \end{aligned} \tag{3.19}$$

where (a) uses that $m_k = \log_2\left(\frac{\max_{s,a,h} c_h^k(s,a)}{\min_{s,a,h} c_h^k(s,a)\vee 1}\right) \vee 1 \leq k$. Now by definition of the good event we have that for any triplet $(h,s,a)\in\widehat{\mathcal{X}}$, $\overline{W}_h(s) \leq 36\sup_\pi p_h^\pi(s)$. Therefore

$$\begin{aligned} \varphi^\star\left(\left[2^k\overline{W}_h(s)\mathbb{1}\left((h,s,a)\in\widehat{\mathcal{X}}\right)\right]_{h,s,a}\right) &\overset{(a)}{\leq} \varphi^\star\left(\left[36\times 2^k\sup_\pi p_h^\pi(s)\mathbb{1}\left((h,s,a)\in\widehat{\mathcal{X}}\right)\right]_{h,s,a}\right) \\ &\overset{(b)}{\leq} \varphi^\star\left(\left[\frac{1152H\beta^{RF}(\tau,\delta/3)\sup_\pi p_h^\pi(s)\mathbb{1}\left((h,s,a)\in\widehat{\mathcal{X}}\right)}{\varepsilon^2}\right]_{h,s,a}\right) \\ &\overset{(c)}{\leq} 1152H\beta^{RF}(\tau,\delta/3)\varphi^\star\left(\left[\frac{\sup_\pi p_h^\pi(s)\mathbb{1}\left(\sup_\pi p_h^\pi(s)\geq\frac{\varepsilon}{32SH^2}\right)}{\varepsilon^2}\right]_{h,s,a}\right), \end{aligned} \tag{3.20}$$

where (a) uses that $\varphi^\star(c) \leq \varphi^\star(c')$ if $\forall(h,s,a)\ c_h(s,a) \leq c'_h(s,a)$, (b) uses Lemma 3.10 and the fact that $k \leq \kappa_f$ since PCE did not stop before phase k and (c) uses Lemma 3.12 and the fact that $\widehat{\mathcal{X}} \subseteq \left\{(h,s,a) : \sup_\pi p_h^\pi(s) \geq \frac{\varepsilon}{32SH^2}\right\}$ on the good event. Using again this last property yields

$$\begin{aligned} \varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}}) &\leq \sum_{h,s,a}\frac{\mathbb{1}\left((h,s,a)\in\hat{\mathcal{X}}\right)}{\sup_\pi p_h^\pi(s,a)} \\ &= \sum_{(h,s,a)\in\widehat{\mathcal{X}}}\frac{1}{\sup_\pi p_h^\pi(s)} \leq \frac{32H^3S^2A}{\varepsilon}, \end{aligned} \tag{3.21}$$

where the first inequality uses Lemma 3.2. Combining (3.19), (3.20) and (3.21) proves the statement for $k \geq 1$. Now it remains to upper bound the duration of the burn-in phase. To that end, we write that by definition of the good event

$$d_0 \leq 64 m_0 \varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}}) + \widetilde{\mathcal{O}}\big(\varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}})SAH^2(\log(6/\delta)+S)\big),$$

where $m_0 = \log_2\left(\frac{\max_{s,a,h} c_h^0(s,a)}{\min_{s,a,h} c_h^0(s,a)\vee 1}\right) \vee 1 = 1$. Therefore

$$\begin{aligned} d_0 &\leq \widetilde{\mathcal{O}}\big(\varphi^\star(\mathbb{1}_{\widehat{\mathcal{X}}})SAH^2(\log(6/\delta)+S)\big) \\ &\leq \widetilde{\mathcal{O}}\left(\frac{S^3A^2H^5(\log(6/\delta)+S)}{\varepsilon}\right), \end{aligned}$$

where the last inequality uses (3.21). ■

Proof of Theorem 3.3

Proof. Denoting by T_{vis} the number of episodes used by the EstimateReachability sub-routine in line 2 of the algorithm, we write

$$\begin{aligned} \tau &= T_{vis} + \sum_{k=0}^{\kappa_f} d_k \\ &\leq T_{vis} + \widetilde{\mathcal{O}}\left(\frac{S^3A^2H^5(\log(6/\delta)+S)}{\varepsilon}\right) \\ &\quad + \sum_{k=1}^{\kappa_f}\left[c_1 kH\beta^{RF}(\tau,\delta/3)\varphi^\star\left(\left[\frac{\sup_\pi p_h^\pi(s,a)\mathbb{1}\left(\sup_\pi p_h^\pi(s)\geq \frac{\varepsilon}{32SH^2}\right)}{\varepsilon^2}\right]_{h,s,a}\right)\right. \\ &\quad \left. + \widetilde{\mathcal{O}}\left(k\frac{S^3A^2H^5(\log(6(k+1)^2/\delta)+S)}{\varepsilon}\right)\right] \\ &\leq T_{vis} + c_1\kappa_f^2 H\beta^{RF}(\tau,\delta/3)\varphi^\star\left(\left[\frac{\sup_\pi p_h^\pi(s)\mathbb{1}\left(\sup_\pi p_h^\pi(s)\geq \frac{\varepsilon}{32SH^2}\right)}{\varepsilon^2}\right]_{h,s,a}\right) \\ &\quad + \widetilde{\mathcal{O}}\left(\kappa_f^2\frac{S^3A^2H^5(\log(6(\kappa_f+1)^2/\delta)+S)}{\varepsilon}\right), \end{aligned} \tag{3.22}$$

where we used Lemma 3.11 to upper bound $(d_k)_{k\geq 0}$. From Theorem 3.6, we know that T_{vis} is deterministic and satisfies

$$T_{vis} = \widetilde{\mathcal{O}}\left(\frac{S^3AH^4\left(\log\left(\frac{SAH}{\delta}\right)+S\right)}{\varepsilon}\right) = \widetilde{\mathcal{O}}\left(\kappa_f^2\frac{S^3A^2H^5(\log(6(\kappa_f+1)^2/\delta)+S)}{\varepsilon}\right). \tag{3.23}$$

Combining inequalities (3.22) and (3.23) with the definition of the threshold $\beta^{RF}(t,\delta) = 4H^2\log(1/\delta) + 24SH^3\log(A(1+t))$ we get

$$\begin{aligned} \tau &\leq c_1\kappa_f^2 H\beta^{RF}(\tau,\delta/3)\varphi^\star\left(\left[\frac{\sup_\pi p_h^\pi(s)\mathbb{1}\left(\sup_\pi p_h^\pi(s)\geq \frac{\varepsilon}{32SH^2}\right)}{\varepsilon^2}\right]_{h,s,a}\right) \\ &\quad + \widetilde{\mathcal{O}}\left(\kappa_f^2\frac{S^3A^2H^5(\log(6(\kappa_f+1)^2/\delta)+S)}{\varepsilon}\right) \\ &\leq c_2\kappa_f^2\left(H^3\log(1/\delta) + SH^4\log(A(1+\tau))\right)\varphi^\star\left(\left[\frac{\sup_\pi p_h^\pi(s)\mathbb{1}\left(\sup_\pi p_h^\pi(s)\geq \frac{\varepsilon}{32SH^2}\right)}{\varepsilon^2}\right]_{h,s,a}\right) \\ &\quad + \widetilde{\mathcal{O}}\left(\kappa_f^2\frac{S^3A^2H^5(\log(6(\kappa_f+1)^2/\delta)+S)}{\varepsilon}\right), \end{aligned} \tag{3.24}$$

where $c_2 = 24c_1$ On the other hand, thanks to Lemma 3.10 and the definition of the threshold β^{RF} we have that

$$\kappa_f \leq \log_2\left(\frac{128H^3\log(1/\delta) + 768SH^4\log(A(1+\tau))}{\varepsilon^2}\right). \tag{3.25}$$

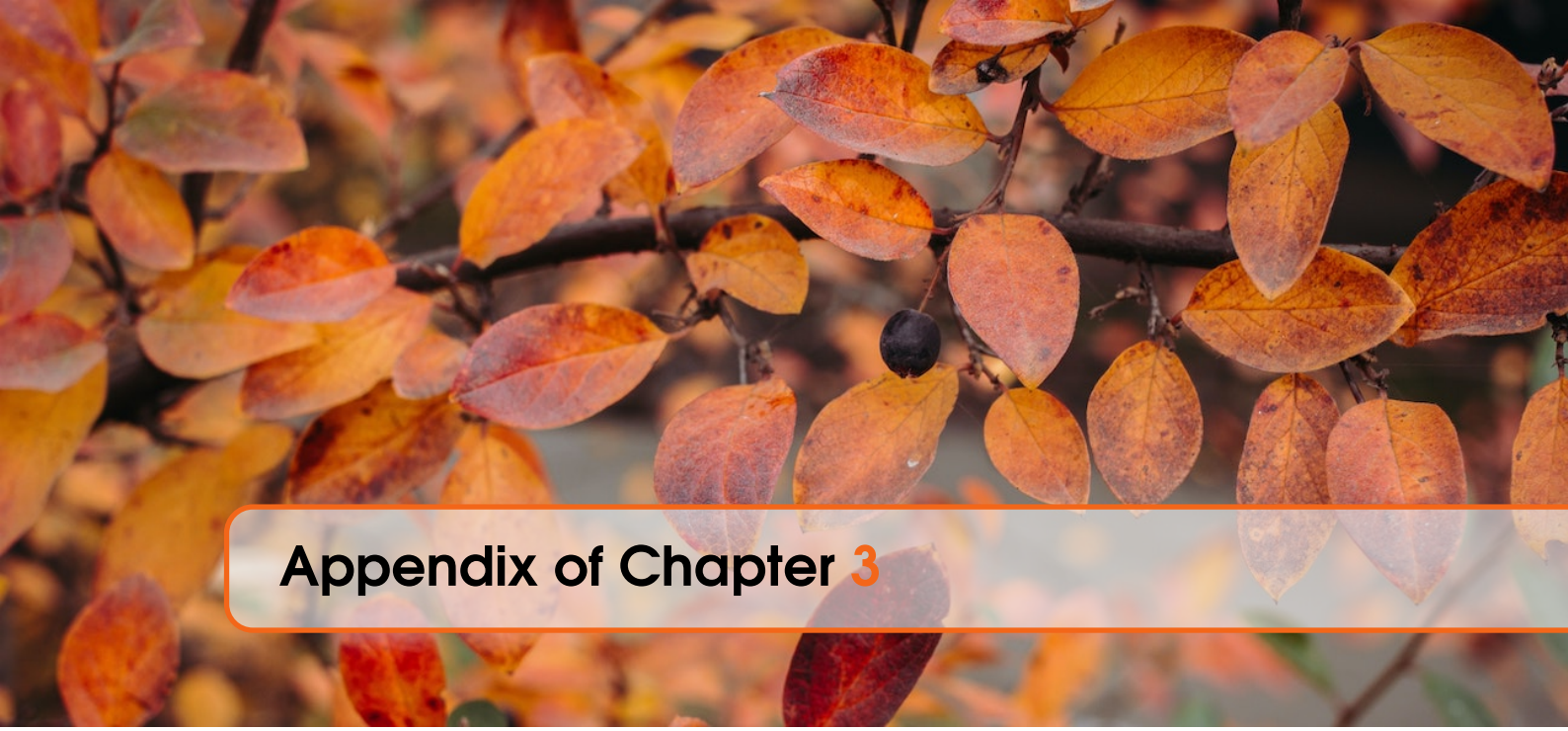
Combining (3.24) with (3.25) and solving for τ we get that

$$\tau \leq \widetilde{\mathcal{O}}\left((H^3 \log(1/\delta) + SH^4)\varphi^\star\left(\left[\frac{\sup_\pi p_h^\pi(s)\mathbb{1}\left(\sup_\pi p_h^\pi(s) \geq \frac{\varepsilon}{32SH^2}\right)}{\varepsilon^2}\right]_{h,s,a}\right) + \frac{S^3A^2H^5(\log(1/\delta)+S)}{\varepsilon}\right),$$

where $\widetilde{\mathcal{O}}$ hides poly-logarithmic factors in S, A, H, ε and $\log(1/\delta)$. ■

3.7 Conclusion

We proposed CovGame, a simple algorithm that adaptively collects episodes in an MDP to explicitly gather a required number of samples $c_h(s,a)$ from each triplet (h,s,a). We proved that its sample complexity scales with a new notion of optimal coverage $\varphi^\star(c)$, which is an instance-dependent lower bound on the sample complexity of *any* adaptive coverage algorithm. We then illustrated the use of CovGame as a building block for reward-free exploration. By relying on (an optimistic variant of) proportional coverage, we proposed PCE, an algorithm for RFE with an instance-dependent sample complexity bound that improves over the minimax rate in several classes of "easy-to-navigate" MDPs.



Appendix of Chapter 3

3.8 Properties of the Minimum Flow

Lemma 3.12 For any $\alpha, \beta \geq 0$ and target functions c_1, c_2, $\varphi^\star(\alpha c_1 + \beta c_2) \leq \alpha\varphi^\star(c_1) + \beta\varphi^\star(c_2)$.

Proof. Clearly, $\varphi^\star(\alpha c_1) = \alpha\varphi^\star(c_1)$ by definition for any $\alpha \geq 0, c_1$. From the LP formulation, we note that if $\eta_1^\star$ (resp. $\eta_2^\star$) is an optimal flow for c_1 (resp. c_2), then $\eta_1^\star + \eta_2^\star$ is a feasible flow for $c_1 + c_2$. This implies that $\varphi^\star(c_1 + c_2) \leq \varphi^\star(c_1) + \varphi^\star(c_2)$ for any c_1, c_2, which proves the statement. ■

3.9 Concentration of Value Functions

In this appendix, we derive the concentration bounds on value functions needed for our PAC RL algorithms. We shall assume that rewards lie in $[0, 1]$ almost surely.

3.9.1 General results

Lemma 3.13 [Concentration of $\widehat{p}^T V$] Let $\mathcal{Z} \subseteq [H] \times \mathcal{S} \times \mathcal{A}$, $Z := |\mathcal{Z}|$, and $\{V_h : \mathcal{S} \to [0, H]\}_{h \in [H+1]}$ be a collection of bounded functions. With probability at least $1 - \delta$, for any $t \geq t_0 := \inf\{t : n_h^t(s, a) \geq 1, \forall (h, s, a) \in \mathcal{Z}\}$,

$$\sum_{(h,s,a) \in \mathcal{Z}} n_h^t(s, a) \left| (\widehat{p}_h^t(s, a) - p_h(s, a))^T V_{h+1} \right|^2 \leq 4H^2 \log(1/\delta) + 2ZH^2 \log(1 + t).$$

Proof. We start by building a suitable stochastic process to apply Theorem 1 of Abbasi-Yadkori et al., 2011. Let $\mathcal{F}_{t,h}$ denote the filtration up to stage h of round t. For any $h \in [H], t \geq 1$, the random variable $\eta_h^t := V_{h+1}(s_{h+1}^t) - p_h(s_h^t, a_h^t)^T V_{h+1}$ is zero-mean and H^2-subgaussian conditionally on $\mathcal{F}_{t,h}$ due to the boundedness of the functions $\{V_h\}_{h \in [H]}$. Let X_h^t be a Z-dimensional vector containing a value 1 at position (h, s_h^t, a_h^t) if $(h, s_h^t, a_h^t) \in \mathcal{Z}$, and zero at all other positions. Note that X_h^t is $\mathcal{F}_{t,h}$-measurable, while η_h^t is $\mathcal{F}_{t,h+1}$-

measurable. Let $Y_t := \sum_{j=1}^{t} \sum_{h=1}^{H} X_h^t \eta_h^t$. For all $(h,s,a) \in \mathcal{Z}$, we have

$$\begin{aligned}[Y_t]_{h,s,a} &= \sum_{j=1}^{t} \mathbb{1}\left(s_h^j = s, a_h^j = a\right)\left(V_{h+1}(s_{h+1}^j) - p_h(s_h^j, a_h^j)^T V_{h+1}\right) \\ &= n_h^t(s,a)(\widehat{p}_h^t(s,a) - p_h(s,a))^T V_{h+1}.\end{aligned}$$

Let $D_t := \sum_{j=1}^{t} \sum_{h=1}^{H} X_h^t (X_h^t)^T = \mathrm{diag}([n_h^t(s,a)]_{(h,s,a)\in\mathcal{Z}})$. Theorem 1 of Abbasi-Yadkori et al., 2011 combined with Equation 20.9 from Lattimore and Szepesvari, 2019 yield that

$$\mathbb{P}\left(\forall t \geq 1,\ \left\|Y^t\right\|^2_{(I+D_t)^{-1}} \leq 2H^2 \log(1/\delta) + ZH^2 \log(1 + t/Z)\right) \geq 1 - \delta.$$

Since $n_h^t(s,a) \geq 1$ for any $t \geq t_0$ and $(h,s,a) \in \mathcal{Z}$, following Corollary 3 in Réda et al., 2021,

$$D_t = \mathrm{diag}\big([n_h^t(s,a)]_{(h,s,a)\in\mathcal{Z}}\big) \succeq (I + D_t)/2,$$

which implies $\left\|Y^t\right\|^2_{D_t^{-1}} \leq 2\left\|Y^t\right\|^2_{(I+D_t)^{-1}}$ for any $t \geq t_0$. Plugging this into the probability above and using that $\left\|Y^t\right\|^2_{D_t^{-1}}$ is exactly the left-hand side of the statement concludes the proof. ■

Lemma 3.14 [Concentration of $\widehat{p}^T V$ for all V] Let $\mathcal{Z} \subseteq [H] \times \mathcal{S} \times \mathcal{A}$, $Z := |\mathcal{Z}|$, and $\mathcal{V} := \{V : \mathcal{S} \to [0,H]\}$ be the set of all bounded functions mapping $\mathcal{S}$ into $[0,H]$. With probability at least $1-\delta$, for any functions $\{V_h \in \mathcal{V}\}_{h=2}^{H+1}$ and $t \geq t_0 := \inf\{t : n_h^t(s,a) \geq 1, \forall (h,s,a) \in \mathcal{Z}\}$,

$$\sum_{(h,s,a)\in\mathcal{Z}} n_h^t(s,a)\left|(\widehat{p}_h^t(s,a) - p_h(s,a))^T V_{h+1}\right|^2 \leq 4H^2 \log(1/\delta) + 12(SH + Z)H^2 \log(1+t).$$

Proof. Let $Y_t(V_2, \ldots, V_{H+1}) := \sum_{(h,s,a)\in\mathcal{Z}} n_h^t(s,a)\left|(\widehat{p}_h^t(s,a) - p_h(s,a))^T V_{h+1}\right|^2$ denote the quantity to be bounded for fixed functions $V_h \in \mathcal{V}$ for all $2 \leq h \leq H+1$. Let $\{\xi_t\}_{t\geq 1}$ be a sequence of positive values to be specified later. For all t, let $\Xi_t := \{\xi_t, 2\xi_t, \ldots \lfloor H/\xi_t \rfloor \xi_t\}$. Note that $|\Xi_t| = \lfloor H/\xi_t \rfloor$ and, for all $x \in [0,H]$, there exists $y \in \Xi_t$ s.t. $|x - y| \leq \xi_t$. For all t, we build a discrete cover $\overline{\mathcal{V}}_t$ of $\mathcal{V}$ as $\overline{\mathcal{V}}_t := \{V : \mathcal{S} \to [0,H] \mid \forall s : V(s) \in \Xi_t\}$. For any t, $\{V_h \in \mathcal{V}\}_{h=2}^{H+1}$, and $\{\overline{V}_h \in \overline{\mathcal{V}}_t\}_{h=2}^{H+1}$, using $x^2 - y^2 = (x+y)(x-y)$ and abbreviating $p_h(s,a)$ and $\widehat{p}_h^t(s,a)$ respectively as $p_{h,s,a}$ and $\widehat{p}_{h,s,a}^t$,

$$\begin{aligned}&\left|Y_t(V_2, \ldots, V_{H+1}) - Y_t(\overline{V}_2, \ldots, \overline{V}_{H+1})\right| \\ &\quad= \left|\sum_{(h,s,a)\in\mathcal{Z}} n_h^t(s,a)(\widehat{p}_{h,s,a}^t - p_{h,s,a})^T (V_{h+1} + \overline{V}_{h+1})(\widehat{p}_{h,s,a}^t - p_{h,s,a})^T (V_{h+1} - \overline{V}_{h+1})\right| \\ &\quad\leq 2H \sum_{(h,s,a)\in\mathcal{Z}} n_h^t(s,a)\left|(\widehat{p}_{h,s,a}^t - p_{h,s,a})^T (V_{h+1} - \overline{V}_{h+1})\right| \\ &\quad\leq 4Ht\|V_{h+1} - \overline{V}_{h+1}\|_\infty.\end{aligned}$$

Therefore,

$$\min_{\{\overline{V}_h \in \overline{\mathcal{V}}_t\}_{h=2}^{H+1}} \left|Y_t(V_2, \ldots, V_{H+1}) - Y_t(\overline{V}_2, \ldots, \overline{V}_{H+1})\right| \leq 4H\xi_t t. \tag{3.26}$$

Now let $\alpha_t := 4H^2 \log(1/\delta_t) + 2ZH^2 \log(1+t) + 4H\xi_t t$ for a sequence $\{\delta_t\}_t$ of values in $(0,1)$ to be defined. We have

$$\begin{aligned}
&\mathbb{P}\left(\exists t \geq t_0, \{V_h \in \mathcal{V}\}_{h=2}^{H+1} : Y_t(V_2, \dots, V_{H+1}) \geq \alpha_t\right) \\
&\quad \leq \mathbb{P}\left(\exists t \geq t_0, \{\overline{V}_h \in \overline{\mathcal{V}}_t\}_{h=2}^{H+1} : Y_t(\overline{V}_2, \dots, \overline{V}_{H+1}) \geq \alpha_t - 4H\xi_t t\right) \\
&\quad \leq \sum_{t=t_0}^{\infty} \sum_{\{\overline{V}_h \in \overline{\mathcal{V}}_t\}_{h=2}^{H+1}} \mathbb{P}\left(Y_t(\overline{V}_2, \dots, \overline{V}_{H+1}) \geq 4H^2 \log(1/\delta_t) + 2ZH^2 \log(1+t)\right) \\
&\quad \leq \sum_{t=t_0}^{\infty} \sum_{\{\overline{V}_h \in \overline{\mathcal{V}}_t\}_{h=2}^{H+1}} \delta_t = \sum_{t=t_0}^{\infty} \delta_t \lfloor H/\xi_t \rfloor^{SH},
\end{aligned}$$

where the first inequality uses (3.26), the second one uses a union bound and the definition of α_t, the third one uses Lemma 3.13, and the equality uses the sizes of the two sets in the sums. Setting $\xi_t = H/t$ and $\delta_t = \frac{\delta}{2t^{SH+2}}$,

$$\sum_{t=t_0}^{\infty} \delta_t \lfloor H/\xi_t \rfloor^{SH} \leq \frac{\delta}{2} \sum_{t=t_0}^{\infty} \frac{1}{t^2} \leq \delta.$$

Finally, with these choices we have

$$\begin{aligned}
\alpha_t &= 4H^2 \log(1/\delta) + 4H^2 \log(2) + 4H^2 \log(t^{SH+2}) + 2ZH^2 \log(1+t) + 4H^2 \\
&\leq 4H^2 \log(1/\delta) + 4H^2 \log(2) + 12SH^3 \log(t) + 2ZH^2 \log(1+t) + 4H^2 \\
&\leq 4H^2 \log(1/\delta) + 12SH^3 \log(t) + 12ZH^2 \log(1+t).
\end{aligned}$$

This implies the statement. ■

Lemma 3.15 [Concentration of $\widehat{r}$] Let $\mathcal{Z} \subseteq [H] \times \mathcal{S} \times \mathcal{A}$ and $Z := |\mathcal{Z}|$. With probability at least $1-\delta$, for any $t \geq t_0 := \inf\{t : n_h^t(s,a) \geq 1, \forall (h,s,a) \in \mathcal{Z}\}$,

$$\sum_{(h,s,a) \in \mathcal{Z}} n_h^t(s,a) \left(\widehat{r}_h^t(s,a) - r_h(s,a)\right)^2 \leq 4 \log(1/\delta) + 2Z \log(1+t).$$

Proof. Following the proof of Lemma 3.13, we build a suitable stochastic process to apply Theorem 1 of Abbasi-Yadkori et al., 2011. We define $\mathcal{F}_{t,h}, X_h^t, Y_t, D_t$ exactly as in the proof of Lemma 3.13, while we redefine $\eta_h^t := r_h^t - r_h(s_h^t, a_h^t)$, with r_h^t the random reward sample observed at stage h of episode t. Since rewards lie in $[0,1]$ almost surely, η_h^t is zero-mean and 1-subgaussian conditionally on $\mathcal{F}_{t,h}$. Moreover, it is easy to see that, for all $(h,s,a) \in \mathcal{Z}$,

$$[Y_t]_{h,s,a} = n_h^t(s,a)(\widehat{r}_h^t(s,a) - r_h(s,a)).$$

Theorem 1 of Abbasi-Yadkori et al., 2011 combined with Equation 20.9 from Lattimore and Szepesvari, 2019 yield that

$$\mathbb{P}\left(\forall t \geq 1,\ \left\|Y^t\right\|^2_{(I+D_t)^{-1}} \leq 2\log(1/\delta) + Z\log(1+t/Z)\right) \geq 1-\delta.$$

We can then conclude exactly as in Lemma 3.13 by showing that $\left\|Y^t\right\|^2_{D_t^{-1}} \leq 2\left\|Y^t\right\|^2_{(I+D_t)^{-1}}$ for any $t \geq t_0$, which implies the statement. ■

3.9.2 Concentration results for RFE

For reward-free exploration, it is sufficient to concentrate the values of all *deterministic* policies. Our concentration result stated below features the threshold function

$$\beta^{RF}(t,\delta) := 4H^2 \log(1/\delta) + 24SH^3 \log(A(1+t)).$$

Theorem 3.4 Let $\mathcal{Z} \subseteq [H] \times \mathcal{S} \times \mathcal{A}$ and $Z := |\mathcal{Z}|$. Suppose that, for some $\varepsilon_0 > 0$, $\max_\pi p_h^\pi(s,a) \le \varepsilon_0$ for all $(h,s,a) \notin \mathcal{Z}$. With probability at least $1-\delta$, for any $t \ge t_0 := \inf\{t : n_h^t(s,a) \ge 1, \forall (h,s,a) \in \mathcal{Z}\}$, $\pi \in \Pi_{\mathrm{D}}$, and reward function $r \in [0,1]^{SAH}$,

$$\Big|\sum_{h,s,a} \big(\widehat{p}_h^{\pi,t}(s,a) - p_h^\pi(s,a)\big) r_h(s,a)\Big| \le \sqrt{\beta^{RF}(t,\delta) \sum_{(h,s,a)\in\mathcal{Z}} \frac{p_h^\pi(s,a)^2}{n_h^t(s,a)}} + (SH - Z_\pi) H \varepsilon_0,$$

where $Z_\pi := |\mathcal{Z} \cap \{(h,s,\pi_h(s)) : h \in [H], s \in \mathcal{S}\}|$.

Proof. Fix any reward r and deterministic policy π. Let V_h^π and $\widehat{V}_h^{\pi,t}$ denote the value functions of π under (p,r) and $(\widehat{p}^t, r)$, respectively. By Lemma 3.16 and the assumption on the set $\mathcal{Z}$,

$$\begin{aligned}\Big|\sum_{h,s,a} \big(\widehat{p}_h^{\pi,t}(s,a) - p_h^\pi(s,a)\big) r_h(s,a)\Big| &\le \sum_{h,s,a} p_h^\pi(s,a) \big|(\widehat{p}_h^t(s,a) - p_h(s,a))^T \widehat{V}_{h+1}^{\pi,t}\big| \\ &\le \sum_{(h,s,a)\in\mathcal{Z}} p_h^\pi(s,a) \big|(\widehat{p}_h^t(s,a) - p_h(s,a))^T \widehat{V}_{h+1}^{\pi,t}\big| + (SH - Z_\pi) H \varepsilon_0.\end{aligned}$$

By applying Lemma 3.14 on the set $\mathcal{Z}_\pi = \mathcal{Z} \cap \{(h,s,\pi_h(s)) : h \in [H], s \in \mathcal{S}\}$, whose cardinality is at most SH, and union bounding over all A^{SH} deterministic policies, with probability at least $1-\delta$, the following holds for all $t \ge t_0$, $\pi \in \Pi_{\mathrm{D}}$, and value functions bounded in $[0,H]$:

$$\sum_{(h,s,\pi_h(s))\in\mathcal{Z}} n_h^t(s,\pi_h(s)) \big|(\widehat{p}_h^t(s,\pi_h(s)) - p_h(s,\pi_h(s)))^T V_{h+1}\big|^2 \le \beta^{RF}(t,\delta).$$

Thus, by Lemma 3.17,

$$\begin{aligned}\sum_{(h,s,a)\in\mathcal{Z}} p_h^\pi(s,a) \big|(\widehat{p}_h^t(s,a) - p_h(s,a))^T \widehat{V}_{h+1}^{\pi,t}\big| &= \sum_{(s,\pi_h(s),h)\in\mathcal{Z}} p_h^\pi(s) \big|(\widehat{p}_h^t(s,\pi_h(s)) - p_h(s,\pi_h(s)))^T \widehat{V}_{h+1}^{\pi,t}\big| \\ &\le \sup_{\substack{u\in\mathbb{R}^{SH}, \\ \sum_{(s,\pi_h(s),h)\in\mathcal{Z}} n_h^t(s,\pi_h(s)) u_{s,h}^2 \le \beta^{RF}(t,\delta)}} \sum_{(s,\pi_h(s),h)\in\mathcal{Z}} p_h^\pi(s) u_{s,h} \\ &= \sqrt{\beta^{RF}(t,\delta) \sum_{(h,s,a)\in\mathcal{Z}} \frac{p_h^\pi(s,a)^2}{n_h^t(s,a)}}.\end{aligned}$$

■

3.9.3 Concentration results for BPI

For BPI, we need concentration bounds on $\big|\widehat{V}_1^{\pi,t} - V_1^\pi\big|$ that hold uniformly across all time steps and *stochastic* policies. Here $\widehat{V}_1^{\pi,t} := \sum_{h,s,a} \widehat{p}_h^{\pi,t}(s,a) \widehat{r}_h^t(s,a)$, where $\widehat{r}_h^t(s,a)$ is

the MLE of $r_h(s,a)$ and $\widehat{p}_h^{\pi,t}(s,a)$ is an estimator of $p_h^\pi(s,a)$ computed from the MLEs $\{\widehat{p}_h(s'|s,a)\}_{h,s,a,s'}$ of the transition probabilities. To this end, we shall define the thresholds

$$\begin{aligned}\beta^r(t,\delta) &:= 4\log(2/\delta) + 2SAH\log(1+t),\\ \beta^p(t,\delta) &:= 4H^2\log(2/\delta) + 24SAH^3\log(1+t),\\ \beta^{bpi}(t,\delta) &:= 16H^2\log(2/\delta) + 96SAH^3\log(1+t).\end{aligned}$$

Compared to $\beta^{RF}(t,\delta)$, we note that $\beta^{bpi}(t,\delta)$ features larger multiplicative constants but also a dependency in A instead of $\log(A)$ in its second term which comes from the need to concentrate the values of all stochastic policies.

Theorem 3.5 With probability at least $1-\delta$, for any $t \ge t_0 := \inf\{t : n_h^t(s,a) \ge 1, \forall(h,s,a)\}$ and $\pi \in \Pi_{\mathrm{S}}$, the following holds:

$$\left|\widehat{V}_1^{\pi,t} - V_1^\pi\right| \le \sqrt{\beta^{bpi}(t,\delta)\min\Big(\sum_{h,s,a}\frac{p_h^\pi(s,a)^2}{n_h^t(s,a)}, \sum_{h,s,a}\frac{\widehat{p}_h^{\pi,t}(s,a)^2}{n_h^t(s,a)}\Big)}.$$

Moreover, for any $\widetilde{r} \in [0,1]^{SAH}$,

$$\Big|\sum_{h,s,a}\big(\widehat{p}_h^{\pi,t}(s,a) - p_h^\pi(s,a)\big)\widetilde{r}_h(s,a)\Big| \le \sqrt{\beta^p(t,\delta)\sum_{h,s,a}\frac{p_h^\pi(s,a)^2}{n_h^t(s,a)}}\,.$$

Proof. Fix any stochastic policy π. By Lemma 3.16,

$$\left|\widehat{V}_1^{\pi,t} - V_1^\pi\right| \le \sum_{h,s,a} p_h^\pi(s,a)\left|\widehat{r}_h^t(s,a) - r_h(s,a)\right| + \sum_{h,s,a} p_h^\pi(s,a)\left|(\widehat{p}_h^t(s,a) - p_h(s,a))^T\widehat{V}_{h+1}^{\pi,t}\right|.$$

By applying Lemma 3.15 and Lemma 3.14 for the set $\mathcal{Z} = \{(h,s,a) : h \in [H], s \in \mathcal{S}, a \in \mathcal{A}\}$, which is of cardinality SAH, with probability at least $1-\delta$, the following hold for all $t \ge t_0$ and for all value functions $(V_h)_{h\in[H]}$ supported in $[0,H]$:

$$\begin{aligned}\sum_{h,s,a} n_h^t(s,a)\left|\widehat{r}_h^t(s,a) - r_h(s,a)\right|^2 &\le \beta^r(t,\delta),\\ \sum_{h,s,a} n_h^t(s,a)\left|(\widehat{p}_h^t(s,a) - p_h(s,a))^T V_{h+1}^{\pi,t}\right|^2 &\le \beta^p(t,\delta).\end{aligned} \tag{3.27}$$

Thus, by Lemma 3.17, optimizing over the deviations as in the proof of Lemma 3.4,

$$\left|\widehat{V}_1^{\pi,t} - V_1^\pi\right| \le \sqrt{\beta^r(t,\delta)\sum_{h,s,a}\frac{p_h^\pi(s,a)^2}{n_h^t(s,a)}} + \sqrt{\beta^p(t,\delta)\sum_{h,s,a}\frac{p_h^\pi(s,a)^2}{n_h^t(s,a)}}.$$

Using that $\beta^r(t,\delta) \le \beta^p(t,\delta)$ and noting that $\beta^{bpi}(t,\delta) = 4\beta^p(t,\delta)$ proves the first statement with the first term in the minimum only. To prove it with the second term as well, it is enough to use Lemma 3.16 with the roles of the two value functions swapped and repeat the same steps as above.

To prove the second statement, we proceed as in the proof of Theorem 3.4 and write

$$\begin{aligned}\Big|\sum_{h,s,a}\big(\widehat{p}_h^{\pi,t}(s,a)-p_h^\pi(s,a)\big)\widetilde{r}_h(s,a)\Big| &\le \sum_{h,s,a} p_h^\pi(s,a)\big|(\widehat{p}_h^t(s,a)-p_h(s,a))^T\widehat{V}_{h+1}^{\pi,t}\big| \\ &\le \sup_{\substack{u\in\mathbb{R}^{SH},\\ \sum_{h,s,a} n_h^t(s,a)u_{h,s,a}^2\le\beta^p(t,\delta)}} \sum_{h,s,a} p_h^\pi(s,a)u_{h,s,a} \\ &= \sqrt{\beta^p(t,\delta)\sum_{h,s,a}\frac{p_h^\pi(s,a)^2}{n_h^t(s,a)}},\end{aligned}$$

where we used Lemma 3.17 and together with inequality (3.27). ■

3.9.4 Auxiliary results

Lemma 3.16 — (Lemma E.15 of Dann et al., 2017). Consider two MDPs with transitions $p,\widehat{p}$ and rewards $r,\widehat{r}$, respectively. Let $V_h^\pi,\widehat{V}_h^\pi$ denote the value function of a (possibly stochastic) policy π in these two MDPs. Then, for any s,h,

$$V_h^\pi(s)-\widehat{V}_h^\pi(s)=\widehat{\mathbb{E}}^\pi\left[\sum_{\ell=h}^H\Big(r_\ell(s_\ell,a_\ell)-\widehat{r}_\ell(s_\ell,a_\ell)+\big(p_\ell(s_\ell,a_\ell)-\widehat{p}_\ell(s_\ell,a_\ell)\big)^T V_{\ell+1}^\pi\Big)\,\Big|\,s_h=s\right].$$

Lemma 3.17 Let $n\in\mathbb{N}$, $p,b\in\mathbb{R}^n$ with b having strictly positive entries, and $c\in\mathbb{R}_{\ge 0}$. Then,

$$\sup_{\substack{x\in\mathbb{R}^n:\\ \sum_{i=1}^n b_i x_i^2\le c}}\sum_{i=1}^n p_i x_i=\sqrt{c\sum_{i=1}^n\frac{p_i^2}{b_i}}.$$

Proof. Let v be the value of the optimization program. Then we know that

$$-v=\inf_{\substack{x\in\mathbb{R}^n:\\ \sum_{i=1}^n b_i x_i^2\le c}} -\sum_{i=1}^n p_i x_i. \tag{3.28}$$

The Lagrangian of the quadratic program above writes as

$$\mathcal{L}(x,\lambda)=-\sum_{i=1}^n p_i x_i+\lambda\Big(\sum_{i=1}^n b_i x_i^2-c\Big),$$

where $\lambda\ge 0$. The KKT conditions then yield that the optimal solution satisfies that

$$\forall i\in[|1,n|],\quad x_i=-\frac{p_i}{2\lambda b_i}$$
$$\sum_{i=1}^n b_i x_i^2=c$$

Solving this system yields that the optimal Lagrange multiplier $\lambda=\sqrt{\frac{c}{\sum_{i=1}^n\frac{p_i^2}{b_i}}}$ which implies that the value of (3.28) is $-\sqrt{c\sum_{i=1}^n\frac{p_i^2}{b_i}}$. ■

3.10 Estimating State Reachability

Let $\mathcal{A}^{\Pi}$ be a regret minimizer that has a small regret for a (fixed) reward function r. If we set this reward function to $r_{h'}^{(h,s)}(s',a') = \mathbb{1}((s'=s, h'=h))$ for a target pair (h,s) intuitively the regret minimizer will visit as much as possible state s in step h and the total reward collected by the algorithm, $n_h^t(s) = \sum_{a\in\mathcal{A}} n_h^t(s,a)$, will be close to $t \times W_h(s)$, where the maximum visitation probability $W_h(s) = \max_\pi p_h^\pi(s)$ is actually the optimal value function in the MDP with reward function $r^{(h,s)}$. The empirical number of visitations can thus be used to estimate the unknown visitation probability.

This idea is already at the heart of the initialization phase of the MOCA algorithm (Wagenmaker et al., 2022a), which relies on repeatedly running the Euler algorithm. We propose a slightly simpler version below, that doesn't need any restart and relies on a generic algorithm $\mathcal{A}^{\Pi}$ satisfying some first-order regret bound scaling with a quantity $\mathcal{R}_\delta^{\Pi}(T)$, as specified in the following theorem. EstimateReachability $((h,s);\varepsilon_0,\delta)$ outputs a valid confidence interval $[\underline{W}_h(s), \overline{W}_h(s)]$ on the value of $W_h(s)$, which can be further used to eliminate all (h,s) whose maximum visitation probability is smaller than a target ε_0.

Algorithm 11 EstimateReachability $((h,s);\varepsilon_0,\delta)$

1: **Input:** Step h, state s, threshold $\varepsilon_0 > 0$, failure probability $\delta \in (0,1)$, regret minimizer $\mathcal{A}^{\Pi}$
2: **Output:** An interval $[\underline{W}_h(s), \overline{W}_h(s)]$
3: Compute $T = T(\varepsilon_0,\delta) = \inf\left\{T \in \mathbb{N} : 4\mathcal{R}_{\delta/2}^{\Pi}(T) + 6\log\left(\frac{4}{\delta}\right) \le \frac{\varepsilon_0}{4}T\right\}$
4: Collect T episodes $\{(s_1^t, a_1^t, \dots, s_H^t, a_H^t)\}_{t\le T}$ using $\mathcal{A}^{\Pi}$ with reward $\widetilde{r}_{h'}(s',a') = \mathbb{1}((s'=s, h'=h))$ and confidence $1-\delta/2$
5: Let $n_h^T(s) = \sum_{t=1}^T \mathbb{1}(s_h^t = s)$ be the number of visits of (h,s)
6: Define $\underline{W}_h(s) = \left(\frac{n_h^T(s)}{2T} - \frac{\varepsilon_0}{16}\right) \vee 0$ and $\overline{W}_h(s) = \left(\frac{2n_h^T(s)}{T} + \frac{\varepsilon_0}{4}\right) \wedge 1$

Theorem 3.6 Assume that, for all (h,s), when $\mathcal{A}^{\Pi}$ is run for the reward function $r = r^{(h,s)}$ and confidence $1-\delta$ up to some horizon $T \in \mathbb{N}$, with probability larger than $1-\delta$,

$$\sum_{t=1}^T V_1^\star(s_1;r) - \sum_{t=1}^T V_1^{\pi^t}(s_1;r) \le \sqrt{\mathcal{R}_\delta^{\Pi}(T) T V^\star(s_1;r)} + \mathcal{R}_\delta^{\Pi}(T). \tag{3.29}$$

For all (h,s), let $[\underline{W}_h(s), \overline{W}_h(s)]$ be the output of EstimateReachability $(h,s;\varepsilon_0,\delta/(SH))$ and define

$$\widehat{\mathcal{X}} = \left\{(h,s) : \underline{W}_h(s) \ge \frac{\varepsilon_0}{8}\right\}.$$

With probability $1-\delta$, the following holds:

- For all (h,s), $W_h(s) \in \left[\underline{W}_h(s), \overline{W}_h(s)\right]$
- $\{(h,s) : W_h(s) \ge \varepsilon_0\} \subseteq \widehat{\mathcal{X}} \subseteq \left\{(h,s) : W_h(s) \ge \frac{\varepsilon_0}{8}\right\}$
- For all $(h,s) \in \widehat{\mathcal{X}}$, $\overline{W}_h(s) \le 36 W_h(s)$.

Moreover, the (deterministic) sample complexity necessary to construct $\widehat{\mathcal{X}}$ is

$$T_{\varepsilon_0}(\delta) := SH \times \inf\left\{T \in \mathbb{N}^\star : T \in \mathbb{N} : 4\mathcal{R}_{\delta/(2SH)}^{\Pi}(T) + 6\log\left(\frac{4}{\delta}\right) \le \frac{\varepsilon_0}{4}T\right\}.$$

In particular, using UCBVI as the regret minimizer, we have $T_{\varepsilon_0}(\delta) = \widetilde{\mathcal{O}}\left(\frac{S^2AH^2\left(\log\left(\frac{SAH}{\delta}\right)+S\right)}{\varepsilon_0}\right)$.

Proof. Let $T = T(\varepsilon_0, \delta)$ be the (deterministic) number of episodes of EstimateReachability $((h,s);\varepsilon_0,\delta)$, which satisfies

$$4\mathcal{R}^{\Pi}_{\delta/2}(T) + 6\log\left(\frac{4}{\delta}\right) \le \alpha\varepsilon_0 T \quad \text{for} \quad \alpha := \frac{1}{4}. \tag{3.30}$$

The analysis relies on the first-order bound on the regret of $\mathcal{A}^{\Pi}$ assumed in (3.29) and on a tight control of the martingale

$$M_T = \sum_{t=1}^{T} \left[\mathbb{1}(s_h^t = s) - p_h^{\pi_t}(s)\right],$$

where $p_h^{\pi}(s) = p_h^{\pi}(s,\pi(s))$ is the probability to reach s under policy π. Observing that the increment of this martingale is bounded in $[-1,1]$ and that its variance is upper bounded by $W_h(s)$, we can use Bernstein's inequality to get that

$$\mathbb{P}\left(|M_T| \le \sqrt{2TW_h(s)\log\left(\frac{4}{\delta}\right)} + \frac{2}{3}\log\left(\frac{4}{\delta}\right)\right) \ge 1 - \frac{\delta}{2}.$$

Remarking that the regret of $\mathcal{A}^{\Pi}$ for the reward function $r = r^{(h,s)}$ can be written

$$\sum_{t=1}^{T} V_1^{\star}(s_1;r) - \sum_{t=1}^{T} V_1^{\pi^t}(s_1;r) = TW_h(s) - \sum_{t=1}^{T} p_h^{\pi^t}(s) = TW_h(s) - n_h^T(s) + M_T$$

and that $n_h^T(s) \le TW_h(s) + M_T$, we obtain that with probability larger than $1-\delta$, the following two inequalities hold:

$$\begin{aligned} n_h^T(s) &\ge TW_h(s) - \left[\sqrt{\mathcal{R}_{\delta/2}(T)TW_h(s)} + \mathcal{R}_{\delta/2}(T) + \sqrt{2\log\left(\frac{4}{\delta}\right)TW_h(s)} + \frac{2}{3}\log\left(\frac{4}{\delta}\right)\right] \\ TW_h(s) &\ge n_h^T(s) - \left[\sqrt{2\log\left(\frac{4}{\delta}\right)TW_h(s)} + \frac{2}{3}\log\left(\frac{4}{\delta}\right)\right] \end{aligned}$$

Using the AM-GM inequality above, this first yields

$$n_h^T(s)/2 - g(\delta) \le TW_h(s) \le 2n_h^T(s) + f(T,\delta),$$

where $f(T,\delta) := 4\mathcal{R}_{\delta/2}(T) + \frac{16}{3}\log\left(\frac{4}{\delta}\right)$ and $g(\delta) := \frac{7}{6}\log\left(\frac{4}{\delta}\right)$. Observing that $g(\delta) \le \frac{1}{4}f(T,\delta)$ and $f(T,\delta) \le \alpha\varepsilon_0 T$ by inequality (3.30), we get

$$\frac{n_h^T(s)}{2T} - \frac{\alpha\varepsilon_0}{4} \le W_h(s) \le \frac{2n_h^T(s)}{T} + \alpha\varepsilon_0,$$

which also implies

$$\frac{W_h(s)}{2} - \frac{\alpha\varepsilon_0}{2} \le \frac{n_h^T(s)}{T} \le 2W_h(s) + \frac{\alpha\varepsilon_0}{2}.$$

As the output of EstimateReachability $((h,s);\varepsilon_0,\delta)$ can be written

$$\left[\underline{W}_h(s) = \left(\frac{n_h^T(s)}{2T} - \frac{\alpha\varepsilon_0}{4}\right) \vee 0, \overline{W}_h(s) = \left(\frac{2n_h^T(s)}{T} + \alpha\varepsilon_0\right) \wedge 1\right]$$

and we get that with probability larger than $1-\delta$:

1. For any value of $W_h(s)$,

$$\frac{W_h(s)}{4} - \frac{\alpha\varepsilon_0}{2} \le \underline{W}_h(s) \le W_h(s) \le \overline{W}_h(s) \le 4W_h(s) + 2\alpha\varepsilon_0.$$

2. If $W_h(s) \geq \varepsilon_0$, then $W_h(s) \in [\underline{W}_h(s), \overline{W}_h(s)] \in [\frac{1-2\alpha}{4} W_h(s), (4+2\alpha)\, W_h(s)]$.
3. If $W_h(s) < \varepsilon_0$, then $W_h(s) \in [\underline{W}_h(s), \overline{W}_h(s)] \in [0, (4+2\alpha)\, \varepsilon_0]$.

Now if $[\underline{W}_h(s), \overline{W}_h(s)]$ is the output of EstimateReachability $((h,s); \varepsilon, \delta/SH)$ and

$$\widehat{\mathcal{X}} = \left\{ (h,s) : \underline{W}_h(s) \geq \frac{1-2\alpha}{4} \varepsilon_0 \right\}$$

we deduce that, with probability $1-\delta$:

- (h,s) with $W_h(s) \geq \varepsilon_0$ are all in $\widehat{\mathcal{X}}$.
- Since $\underline{W}_h(s) \leq W_h(s)$, any (h,s) with $W_h(s) < \frac{1-2\alpha}{4}\varepsilon_0$ does not belong to $\widehat{\mathcal{X}}$.

This proves that $\{(h,s) : W_h(s) \geq \varepsilon_0\} \subseteq \hat{\mathcal{X}} \subseteq \{(h,s) : W_h(s) \geq \frac{1-2\alpha}{4}\varepsilon_0\}$. To prove the last statement we remark that for $(h,s) \in \widehat{\mathcal{X}}$, if $W_h(s) \geq \varepsilon_0$, we have by 2. that $\overline{W}_h(s) \leq (4+2\alpha)\, W_h(s)$ while if $W_h(s) \in \left[\frac{1-2\alpha}{4}\varepsilon_0, \varepsilon_0\right)$ we have by 3. that

$$\overline{W}_h(s) \leq (4+2\alpha)\, \varepsilon_0 \leq 4\frac{4+2\alpha}{1-2\alpha} W_h(s)$$

Plugging the value $\alpha = 1/4$ yields $\overline{W}_h(s) \leq 36 W_h(s)$ in both cases.

To get an upper bound on the number of episodes used by an instance of EstimateReachability, we need to find a T that satisfies

$$T - 1 \leq \frac{16}{\varepsilon_0} \mathcal{R}^{\Pi}_{\delta/(2SH)}(T) + \frac{24}{\varepsilon_0} \log\left(\frac{SAH}{\delta}\right). \tag{3.31}$$

For UCBVI, Theorem 19 of (Al-Marjani et al., 2023) yields a regret bound with $\mathcal{R}_\delta(T) = 256^2 SAH \left(\log\left(\frac{2SAH}{\delta}\right) + 6S\right) \log^2(T+1)$. Using the inequality $\log^2(x) \leq 4\sqrt{x}$ we get a first crude upper bound on T by solving a quadratic equation which gives the final scaling by plugging back this crude bound in (3.31). ■

4. Implicit Policy Eliminations for Efficient ε-BPI

In this chapter, we present an asymptotic instance-dependent lower bound for the sample complexity of ε-Best Policy Identification (ε-BPI) in episodic MDPs, see Section 1.4.1. Then we design PRINCIPLE, an algorithm for this problem based on the idea of *implicit policy eliminations*. The lower bound is based on some (yet) unpublished results, while the algorithm comes from Appendix F of the conference paper:

Aymen Al Marjani, Andrea Tirinzoni, and Emilie Kaufmann. **Active Coverage for PAC Reinforcement Learning**. In *Proceedings of the 36th Conference On Learning Theory* (COLT), 2023.

Contents

4.1 Instance Dependent Lower Bounds

In this section we consider the class $\mathfrak{M}_1$ of stochastic MDPs with *Gaussian rewards* of unit variance, in which $\nu_h(s,a) = \mathcal{N}(r_h(s,a), 1)$. While this setting differs from the standard assumption that the rewards are almost surely in $[0,1]$, there are two reasons which, in our opinion, justify its study. First, this setting has proved useful in previous works to derive *closed-form lower bounds that scale with intuitive quantities* such as the *return gaps*, see (Dann et al., 2021) and (Tirinzoni et al., 2022). Second, as we will see shortly, the resulting lower bound is nearly matched by an algorithm that assumes that the reward distributions are sub-gaussian.

Notation: $\Pi^\varepsilon := \{\pi \in \Pi_{\mathrm{D}} : V_1^\pi(s_1) \geq V_1^\star(s_1) - \varepsilon\}$ refers to the set of all deterministic ε-optimal policies. Denoting by $\mathbb{P}^\pi$ (resp. $\mathbb{E}^\pi$) the probability (resp. expectation) operator induced by the execution of a Markovian policy $\pi \in \Pi_{\mathrm{S}}$ for an episode on $\mathcal{M}$, we let $V_1^\pi := \mathbb{E}^\pi\left[\sum_{h=1}^H R_h | s_1\right]$ be the value function of π at the initial state[1]. The *policy gap* of π is then defined as $\Delta(\pi) := V_1^\star - V_1^\pi$, where $V_1^\star := \max_{\pi\in\Pi_{\mathrm{D}}} V_1^\pi$ is the optimal value function at s_1. We further define the minimum policy gap $\Delta_{\min}(\Pi_{\mathrm{D}}) := \min_{\pi\in\Pi_{\mathrm{D}}\setminus\{\pi^\star\}} \Delta(\pi)$, where $\pi^\star$ is an arbitrary optimal policy (i.e., $V_1^{\pi^\star} = V_1^\star$). Note that $\Delta_{\min}(\Pi_{\mathrm{D}}) = 0$ whenever multiple optimal policies exist. Moreover, we denote the visitation probability of (h,s,a) under π as $p_h^\pi(s,a) := \mathbb{P}^\pi(s_h = s, a_h = a)$ and $p_h^\pi(s) := \mathbb{P}^\pi(s_h = s)$. We let $\Omega := \left\{\left(p_h^\pi(s,a)\right)_{h,s,a} : \pi \in \Pi_{\mathrm{S}}\right\}$ the set of state-action distributions generated by stochastic policies. We recall that

$$\Omega(\mathcal{M}) := \left\{\rho \in \mathbb{R}_+^{SAH} : \textstyle\sum_{a\in\mathcal{A}} \rho_1(s,a) = 1, \sum_a \rho_h(s,a) = \sum_{(s',a')} \rho_{h-1}(s',a') p_{h-1}(s|s',a') \ \forall (h,s)\right\}.$$

4.1.1 General lower bound for near-optimal policy identification

Our first result is a general bound that holds for any $\varepsilon \geq 0$ in the regime $\delta \to 0$. Its proof, which follows the same steps as the proof of the lower bound for ε-Best Arm Identification (and other pure exploration problems) of (Degenne & Koolen, 2019), is deferred to Appendix 4.6.

Theorem 4.1 Any ε-BPI algorithm that is (ε,δ)-PAC for all instances in $\mathfrak{M}_1$ satisfies, for any $\mathcal{M} \in \mathfrak{M}_1$,

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_{\mathcal{M}}[\tau]}{\log(1/\delta)} \geq LB(\mathcal{M},\varepsilon)$$

where

$$LB(\mathcal{M},\varepsilon) := 2 \min_{\pi^\varepsilon\in\Pi^\varepsilon} \min_{\rho\in\Omega(\mathcal{M})} \max_{\pi\in\Pi_{\mathrm{D}}} \sum_{s,a,h} \frac{\left(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a)\right)^2}{\rho_h(s,a)(\Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon)^2}.$$

Theorem 4.1 states that no matter how adaptive an ε-BPI algorithm is, there is a minimal cost in terms of episodes that it must pay in order to learn an ε-optimal policy of $\mathcal{M}$. This cost is problem-dependent since it is a functional of $\mathcal{M}$, the MDP to be learned. A more detailed interpretation of the complexity $LB(\mathcal{M},\varepsilon)$ is further provided in Section 4.1.3. We note that one can get rid of the assumption of unit variance simply by multiplying each term of the sum that appears in $LB(\mathcal{M},\varepsilon)$ with σ_{hsa}^2. This gives a lower bound for MDPs with Gaussian rewards where the (known) variances may vary across triplets (h,s,a).

[1] Since we consider episodic MDPs where the initial state s_1 is fixed, we drop it from the notation of value functions.

4.1.2 Finite-risk bound for exact identification

In the case of exact identification (i.e. $\varepsilon = 0$), we derive a lower bound which is valid for any $\delta \in (0,1)$ under the assumption that the optimal state-action distribution is unique.

Assumption 4.1 We assume that there exists $p^\star \in \Omega(\mathcal{M})$ s.t. for any optimal policy $\pi^\star$ (i.e., with $V_1^{\pi^\star} = V_1^\star$) we have $p^{\pi^\star} = p^\star$.

Note that this is assumption was considered in (Tirinzoni et al., 2021). As shown in that paper, it implies that there is a unique optimal action in states visited with positive probability by some optimal policy, but there can be arbitrarily many optimal actions in all other states.

Theorem 4.2 Fix any MDP $\mathcal{M} \in \mathfrak{M}_1$ s.t. the optimal state-action distribution $p^\star$ is unique. Then, for any $(0,\delta)$-correct ε-BPI algorithm,

$$\mathbb{E}_{\mathcal{M}}[\tau] \geq 2 \min_{\rho \in \Omega} \max_{\substack{\pi \in \Pi_{\mathrm{D}}: \\ \Delta(\pi) > 0}} \sum_{s,a,h} \frac{(p_h^\pi(s,a) - p_h^\star(s,a))^2}{\rho_h(s,a)\Delta(\pi)^2} \log\left(\frac{1}{2.4\delta}\right).$$

Remark 4.1 When $S = H = 1$ and the optimal action $a^\star$ is unique, the bound above reduces to $2\sum_{a \in [K]} \frac{1}{\Delta_a^2}$, where $\Delta_{a^\star} := \min_{a \neq a^\star} \Delta_a$ and $\Delta_a := \mu^\star - \mu_a$ for $a \neq a^\star$. This is, up to a universal constant, equal to the lower bound for best-arm identification in Gaussian multi-armed bandits, see Lemma (Garivier & Kaufmann, 2016). ■

Proof. The idea is to compute, in closed form, the smallest KL divergence between the distribution of the observation under the MDP $\mathcal{M}$ and under an alternative $\widetilde{\mathcal{M}}$ that has the same transitions but a different mean reward function $r_h^{\widetilde{\mathcal{M}}}$. Let $n_h^\tau(s,a) := \sum_{t=1}^\tau \mathbb{1}(s_h = s, a_h = a)$ denote the number of visits to the triplet (h,s,a) until the last episode. By an analogue of (1.39), the KL divergence between distributions of observations under $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ takes the simple form

$$\mathrm{KL}(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) = \sum_{h,s,a} \mathbb{E}_{\mathcal{M}}[n_h^\tau(s,a)] \frac{\left(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\right)^2}{2},$$

where we used that within the class $\mathfrak{M}_1$ reward distributions are Gaussian and the transition kernel is the same as that of $\mathcal{M}$. Note that, since $p^\star$ is unique, any δ-correct algorithm satisfies $\mathbb{P}_{\mathcal{M}}(V_1^{\widehat{\pi}} = V_1^\star) = \mathbb{P}_{\mathcal{M}}(p^{\widehat{\pi}} = p^\star) \geq 1 - \delta$. Now fix a sub-optimal policy π (i.e., with $\Delta(\pi) > 0$). The closest alternative $\widetilde{\mathcal{M}}$ where π becomes better than any optimal policy of $\mathcal{M}$ can be computed by solving the quadratic program

$$\inf_{\widetilde{r}: \widetilde{r}^T p^\pi > \widetilde{r}^T p^\star} \sum_{s,a,h} \mathbb{E}[n_h^\tau(s,a)] \frac{(r_h(s,a) - \widetilde{r}_h(s,a))^2}{2} = \frac{\Delta(\pi)^2}{2\sum_{s,a,h} \frac{(p_h^\pi(s,a) - p_h^\star(s,a))^2}{\mathbb{E}[n_h^\tau(s,a)]}}.$$

By δ-correctness, in such closest alternative we have $\mathbb{P}_{\widetilde{\mathcal{M}}}(p^{\widehat{\pi}} = p^\star) \leq \delta$. Then, by an analogue of Lemma 1.1, for any π with $\Delta(\pi) > 0$,

$$\frac{\Delta(\pi)^2}{2\sum_{s,a,h} \frac{(p_h^\pi(s,a) - p_h^\star(s,a))^2}{\mathbb{E}[n_h^\tau(s,a)]}} \geq \log\left(\frac{1}{2.4\delta}\right).$$

The final lower bound is obtained by solving the optimization problem

$$\begin{aligned}
&\underset{\eta \in \mathbb{R}^{SAH}}{\text{minimize}} \sum_a \eta_1(s_1, a), \\
&\text{subject to} \\
&\quad \sum_a \eta_h(s,a) = \sum_{s',a'} p_{h-1}(s|s',a')\eta_{h-1}(s',a') \quad \forall s \in \mathcal{S}, h > 1, \\
&\quad \eta_1(s,a) = 0 \quad \forall s \in \mathcal{S} \setminus \{s_1\}, a \in \mathcal{A}, \\
&\quad 2\log\left(\frac{1}{2.4\delta}\right) \sum_{s,a,h} \frac{(p_h^\pi(s,a) - p_h^\star(s,a))^2}{\eta_h(s,a)} \le \Delta(\pi)^2 \quad \forall \pi : \Delta(\pi) > 0,
\end{aligned}$$

where we performed the change of variable $\eta_h(s,a) = \mathbb{E}[n_h^\tau(s,a)]$ and used that η must satisfy the navigation constraints. Note that the last constraint is equivalent to

$$2\log\left(\frac{1}{2.4\delta}\right) \max_{\pi:\Delta(\pi)>0} \sum_{s,a,h} \frac{\sum_{a'} \eta_1(s_1,a')}{\eta_h(s,a)} \frac{(p_h^\pi(s,a) - p_h^\star(s,a))^2}{\Delta(\pi)^2} \le \sum_a \eta_1(s_1,a) \tag{4.1}$$

Finally, we apply the change of variable

$$\forall (h,s,a),\ \rho_h(s,a) = \frac{\eta_h(s,a)}{\sum_a \eta_1(s_1,a)}.$$

It is straightforward that $\rho \in \Omega(\mathcal{M})$ is a valid state-action distribution. Replacing by ρ in the LHS of (4.1) and plugging back into the optimization program yields the stated lower bound. ■

4.1.3 Interpreting the lower bound

While the expression of the lower bound might seem mysterious at first glance, we provide below a possible interpretation in terms of the reduction of some confidence intervals, in the simpler setting of known transitions and unknown reward distributions. Our explanation hinges on the following concentration inequality, proved in Appendix 4.7.

Lemma 4.1 Assume that the rewards are in $[0,1]$ almost surely. For any policy $\pi \in \Pi_{\mathrm{D}}$, define the estimator $\widehat{V}_1^{\pi,t} := \sum_{h,s,a} p_h^\pi(s,a)\widehat{r}_h^t(s,a)$, where $\widehat{r}_h^t(s,a)$ is the MLE of $r_h(s,a)$ using samples gathered until episode t. Then the event

$$\mathcal{E} := \left(\forall t \ge t_0,\ \forall \pi, \pi' \in \Pi_{\mathrm{D}},\ \left|(\widehat{V}_1^{\pi,t} - \widehat{V}_1^{\pi',t}) - (V_1^\pi - V_1^{\pi'})\right| \le \sqrt{\beta^r(t,\delta) \sum_{h,s,a} \frac{\left(p_h^\pi(s,a) - p_h^{\pi'}(s,a)\right)^2}{n_h^t(s,a)}}\right)$$

holds with probability larger than $1-\delta$, where $t_0 := \inf\{t : n_h^t(s,a) \ge 1, \forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}\}$ and $\beta^r(t,\delta) := 4\log(1/\delta) + 4SH\log(A(1+t))$.

Suppose that a learner explores the MDP $\mathcal{M}$ using a fixed (stochastic) policy π^{exp} whose state-action distribution is ρ. Then, after playing π^{exp} for $K \ge 1$ episodes, $\mathbb{E}[n_h^K(s,a)] = K\rho_h(s,a)$ so that the size of the confidence interval on $V_1^{\pi^\varepsilon} - V_1^\pi$ should roughly be $\sqrt{\beta^r(t,\delta) \sum_{h,s,a} \frac{\left(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a)\right)^2}{K\rho_h(s,a)}}$. Now, if the learner wishes to test whether π^ε is ε-optimal it has to determine the sign of $V_1^{\pi^\varepsilon} - V_1^\pi + \varepsilon$ for all other policies π. To do that, it is sufficient to shrink the size of the confidence interval on $V_1^{\pi^\varepsilon} - V_1^\pi$ below $\frac{1}{2}|V_1^{\pi^\varepsilon} - V_1^\pi + \varepsilon| = \frac{1}{2}|\Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon|$ for all policies π. Solving for the minimal K that satisfies the previous conditions, we see that playing roughly

$$K(\pi^{\mathrm{exp}}, \pi^\varepsilon) \propto \log(1/\delta) \max_{\pi \in \Pi_{\mathrm{D}}} \sum_{s,a,h} \frac{\left(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a)\right)^2}{\rho_h(s,a)(\Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon)^2}$$

episodes using the exploration policy π^{exp} is enough to determine whether π^{ε} is ε-optimal. Since the learner has the liberty to return *any* ε-optimal policy using *any* exploration policy, the lower bound corresponds to the minimum of $K(\pi^{\text{exp}}, \pi^{\varepsilon})$ w.r.t to these two variables.

4.2 Towards a Matching Upper Bound

In this section, we review the existing problem-dependent upper bounds for the ε-BPI problem (with $\varepsilon > 0$). As we will see, a recent algorithm proposed by (Wagenmaker & Jamieson, 2022) nearly matches the lower bound of Theorem 4.1.

ε-BPI with a generative model (Zanette et al., 2019) were the first to propose an instance-dependent ε-BPI algorithm, called BESPOKE. In infinite-horizon tabular MDPs with a discount factor $\gamma \in [0, 1)$ and when the algorithm has access to a generative model, BESPOKE finds an ε-optimal policy with a sample complexity of at most

$$\widetilde{\mathcal{O}}\left(\left[\sum_{s,a} \min \left(\frac{1}{(1-\gamma)^3 \varepsilon^2}, \frac{\text{Var}[R(s,a)] + \gamma^2 \text{Var}_{s' \sim p(.|s,a)}[V^\star(s')]}{\max(\Delta(s,a), (1-\gamma)\varepsilon)^2} + \frac{1}{(1-\gamma)\max(\Delta(s,a), (1-\gamma)\varepsilon)} \right) \right] \log\left(\tfrac{1}{\delta}\right) \right),$$

where $\Delta(s,a) = V^\star_{\mathcal{M}}(s) - Q^\star_{\mathcal{M}}(s,a)$ is the value gap of state-action pair (s,a) and Var denotes the variance operator. This bound is always smaller than the conjectured minimax rate for this setting (Azar et al., 2013). For the setting of episodic linear MDPs (Jin et al., 2019), the GSS-E algorithm by (Taupin et al., 2022) solves a G-optimal design to determine the sampling frequencies of each state-action pair. The sample complexity of GSS-E is upper bounded by $\widetilde{\mathcal{O}}\left(\frac{dH^4}{(\Delta_{min}(\mathcal{M}) + \varepsilon)^2} (\log(1/\delta) + d) \right)$, where $\Delta_{min}(\mathcal{M}) = \min_{s,a \neq \pi^\star(s)} \Delta(s,a)$ is the minimum value gap in $\mathcal{M}$. Up to horizon factors, this result improves upon the $\Omega(d^2 H^2/\varepsilon^2)$ minimax bound for this setting (Wagenmaker et al., 2022b) whenever the minimum value gap in $\mathcal{M}$ is large.

Online ε-BPI On top of the sub-optimality gaps which characterized the bounds when a generative model is available, the problem-dependent complexities in online ε-BPI feature an additional component, namely visitation probabilities. These constitute the price that online ε-BPI algorithms pay in order to navigate the MDP and collect observations from distant states. Most existing results on the sample complexity are of the form $\mathbb{P}_{\mathcal{M},\text{Alg}}\left(\tau = \widetilde{\mathcal{O}}\left(\mathcal{C}_{\text{Alg}}(\mathcal{M}, \varepsilon) \log\left(\tfrac{1}{\delta}\right)\right)\right) \geq 1 - \delta$, where $\mathcal{C}_{\text{Alg}}(\mathcal{M}, \varepsilon)$ is a complexity measure corresponding to a given algorithm Alg and the $\widetilde{\mathcal{O}}$ notation is used to hide numerical constants and logarithmic factors in $S, A, H, 1/\varepsilon$ and $\log(1/\delta)$. For example, for the MOCA algorithm (Wagenmaker et al., 2022a) obtain

$$\mathcal{C}_{\text{MOCA}}(\mathcal{M}, \varepsilon) = H^2 \sum_{h=1}^{H} \min_{\pi^{\text{exp}} \in \Pi_S} \max_{s,a} \frac{1}{p_h^{\pi^{\text{exp}}}(s,a)} \min\left(\frac{1}{\Delta_h(s,a)^2}, \frac{W_h(s)^2}{\varepsilon^2} \right) + \frac{H^4 \left|\text{OPT}(\mathcal{M}, \varepsilon)\right|}{\varepsilon^2},$$

where $\Delta_h(s,a) := V^\star_h(s) - Q^\star_h(s,a)$ is the value gap of triplet (h,s,a), $W_h(s) = \sup_\pi p_h^\pi(s)$ is the maximum reachability of state s at step $h \in [H]$ and $\text{OPT}(\mathcal{M}, \varepsilon)$ is a set of near-optimal triplets (h,s,a). In the above bound, the contribution of a triplet (h,s,a) to the total complexity will be small when either (i) its value gap $\Delta_h(s,a)$ is large or (ii) it is hard to reach by any policy, that is $W_h(s) \ll \varepsilon$. This "local complexity" of (h,s,a) is weighted by $1/p_h^{\pi^{\text{exp}}}(s,a)$, which is the (expected) number of episodes that the algorithm needs to play in order to reach (h,s,a) when using π^{exp} as an exploration policy. Subsequent works have proposed alternative local complexity measures featuring policy gaps instead of value gaps (Tirinzoni et al., 2022; Wagenmaker & Jamieson, 2022). Policy gaps can be larger than value gaps, notably in deterministic MDPs (Tirinzoni et al., 2022).

4.2.1 PEDEL: A close to optimal algorithm

Among algorithms whose sample complexity is expressed with policy gaps, the PEDEL algorithm proposed by (Wagenmaker & Jamieson, 2022) has the complexity term which looks the most like the complexity measure in our lower bound. PEDEL can tackle the more general setting of identifying a near-optimal policy in linear MDPs (Jin et al., 2019). Its instantiation to the special case of tabular MDPs yields a sample complexity whose leading term is

$$\mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M},\varepsilon) = H^4 \sum_{h=1}^{H} \min_{\rho\in\Omega} \max_{\pi\in\Pi_{\mathrm{D}}} \sum_{s,a} \frac{p_h^\pi(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2},$$

ignoring some additive second-order term which is polynomial in $S, A, H, \log(1/\delta)$ and $\log(1/\varepsilon)$. The next proposition, proved in Appendix 4.8, compares this complexity measure to the lower bound.

Proposition 4.1 For any MDP $\mathcal{M}$, it holds that

$$\mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M},\varepsilon) \leq 8H^5 LB(\mathcal{M},\varepsilon) + \frac{4H^6}{(\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}.$$

This result shows that for MDPs in which the minimum policy gap is a constant w.r.t other problem parameters $\Delta_{\min}(\Pi_{\mathrm{D}}) = \Omega(1)$, the complexity $\mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M},\varepsilon)$ is only an H^5 factors away from the instance-dependent lower bound. The same conclusion holds when we are interested in the regime $\varepsilon = \Omega(1)$.

Remark 4.2 Upon close inspection of its pseudocode, it seems that PEDEL was designed with the implicit assumption that $\varepsilon = \mathcal{O}(H/d^{3/2})$, where d is the dimension of the linear MDP ($d = SAH$ in our tabular setting). This results in cases, e.g. if $\varepsilon = \Omega(1/d)$, where the true sample complexity of PEDEL can be d times larger than $\mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M},\varepsilon)$. We elaborate on this in Appendix 4.8.2. ■

4.3 Proportional Coverage with Implicit PoLicy Elimination

4.3.1 Basic intuition

We now present PRINCIPLE, an algorithm for ε-BPI, which uses CovGame as a subroutine for exploration. Recall that in the PCE algorithm of Chapter 3, we sought to achieve good proportional coverage w.r.t. the set of all policies, i.e., by requiring that $n_h^k(s,a) \geq 2^k \sup_{\pi\in\Pi_{\mathrm{D}}} p_h^\pi(s,a)$ for all h, s, a, k. This is due to the "worst-case" nature of RFE, where any policy can be potentially optimal for some reward function at test time. On the contrary, the mean-reward r is fixed in BPI, a property that we can leverage to perform more adaptive exploration. A natural idea, which led to the tight theoretical guarantees on PEDEL that we saw earlier, is to eliminate policies as soon as we are confident enough that they are sub-optimal. This helps the algorithm adapt its exploration to focus on policies of higher value. The same idea was also used by (Tirinzoni et al., 2022) to perform near-optimal ε-BPI in deterministic MDPs. Unfortunately, while (Tirinzoni et al., 2022) managed to achieve so in a computationally efficient manner for deterministic MDPs, PEDEL needs to enumerate all policies to do the same in stochastic environments, hence yielding an exponential time-memory algorithm. Our strategy, PRINCIPLE, achieves a policy-gaps dependent sample complexity while remaining computationally efficient. Its pseudo-code is reported in Algorithm 12.

Implicit policy eliminations The key idea is to replace the explicit policy eliminations of PEDEL with sequential constraints on the set of state-action distributions corresponding to high-reward policies. While PEDEL computes at each round k a set of policies Π^k that contains an optimal policy with high probability, PRINCIPLE computes a set of state-action distributions Ω^k that w.h.p contains the distribution vector $[\widehat{p}_h^{\pi^\star,k}(s,a)]_{h,s,a}$ of some optimal policy $\pi^\star$ under the empirical transition model $\widehat{p}^k$. In particular, PRINCIPLE maintains, at each phase k, a high-probability lower bound $\underline{V}_1^k$ on the optimal value function $V_1^\star$ computed as

$$\underline{V}_1^k := \sup_{\substack{\rho \in \Omega(\widehat{p}^k), \\ \max_{h,s,a} \rho_h(s,a)/n_h^k(s,a) \le 2^{-k}}} \sum_{h,s,a} \rho_h(s,a)\widehat{r}_h^k(s,a) - \sqrt{2^{2-k} H \beta^{bpi}(t_k, \delta/3)},$$

where $\beta^{bpi}(t,\delta) \propto H^2 \log(1/\delta) + SAH^3 \log\log(t)$ and $\Omega(\widehat{p}^k)$ is the set of valid visitation probabilities in the empirical MDP whose transition kernel is $\widehat{p}^k$. As common, $\underline{V}_1^k$ is computed by subtracting a confidence interval to the maximum expected return estimated on the empirical MDP defined by $(\widehat{p}^k, \widehat{r}^k)$. A notable exception is that we focus only on state-action distributions that are *well-covered* by the current data, i.e., such that $\max_{h,s,a} \rho_h(s,a)/n_h^k(s,a) \le 2^{-k}$. Then, PRINCIPLE defines a set of "active" state-action distributions as

$$\Omega^k := \left\{ \rho \in \Omega(\widehat{p}^k) : \sum_{h,s,a} \rho_h(s,a)\widehat{r}_h^k(s,a) \ge \underline{V}_1^k, \quad \max_{h,s,a} \rho_h(s,a)/n_h^k(s,a) \le 2^{-k} \right\}.$$

Intuitively, ρ is active at phase k if (1) it is a valid state-action distribution in the empirical MDP with transition probabilities $\widehat{p}^k$, (2) it induces an estimated expected return $\sum_{h,s,a} \rho_h(s,a)\widehat{r}_h^k(s,a)$ larger than $\underline{V}_1^k$, and (3) it is well-covered by the current data. Then, as compared to PCE, PRINCIPLE simply replaces the quantity $\sup_{\pi \in \Pi_D} p_h^\pi(s,a)$ in the target function used for CovGame at phase k with $\sup_{\rho \in \Omega^{k-1}} \rho_h(s,a)$, i.e., it restricts the exploration to active state-action distributions. In our analysis, we show that with high probability, state-action distributions corresponding to optimal policies are never eliminated from Ω^k and $\underline{V}_1^k$ gradually approaches $V_1^\star$ from below. That is, Ω^k is dynamically pruned to contain only distributions corresponding to higher returns, hence achieving implicit eliminations of sub-optimal policies.

Computational complexity The computations of $\underline{V}_1^k$ and $\sup_{\rho \in \Omega^{k-1}} \rho_h(s,a)$ amount to solving standard constrained MDPs, which can be done by linear programming (e.g., Efroni et al., 2020). Moreover, PRINCIPLE does not store the set Ω^k but only its associated constraints, whose number is linear in SAH. This implies that PRINCIPLE, unlike PEDEL, requires polynomial (in SAH) time and memory.

4.3.2 Theoretical guarantees

Theorem 4.3 PRINCIPLE is (ε,δ)-PAC for ε-BPI and, with probability $1-\delta$, it has sample complexity

$$\tau \le \widetilde{\mathcal{O}}\left((H^3 \log(1/\delta) + SAH^4) \left[\varphi^\star\left(\left[\sup_{\pi \in \Pi_S} \frac{p_h^\pi(s,a)}{\max(\varepsilon, \Delta(\pi))^2} \right]_{h,s,a} \right) + \frac{\varphi^\star(\mathbb{1})}{\varepsilon} + \varphi^\star(\mathbb{1}) \right] \right),$$

where $\mathbb{1}$ denotes a function equal to 1 for all h,s,a and $\widetilde{\mathcal{O}}$ hides poly-logarithmic factors in $S, A, H, \varepsilon, \log(1/\delta)$ and $\varphi^\star(\mathbb{1})$.

4.3.3 Pseudo-code

Notation: To simplify the presentation of the algorithm and the analysis, we index the counts as well as the empirical estimates of transitions and rewards by their phase number (instead of episode number). Hence, for each triplet (h,s,a), $n_h^k(s,a), \widehat{p}_h^k(.|s,a)$ and $\widehat{r}_h^k(s,a)$ will refer to the number of visits, the empirical transition kernel and the empirical mean reward respectively after t_k episodes, i.e. at the end of the k-th phase. For a transition kernel $\widetilde{p}$, we define the corresponding set of state-action distributions as $\Omega(\widetilde{p}) := \left\{[\widetilde{p}_h^{\pi}(s,a)]_{h,s,a} : \pi \in \Pi_{\mathrm{S}}\right\}$. Finally, for a dataset of episodes $\mathcal{D}$, $n_h(s,a;\mathcal{D})$ denotes the number of visits of (h,s,a) in the episodes stored in $\mathcal{D}$.

Algorithm 12 PRINCIPLE (PRoportIoNal Coverage with Implicit PoLicy Elimination)

Input: Precision ε, Confidence δ, set of reachable states $\mathcal{S}$
Output: A policy $\widehat{\pi}$ that is ε-optimal w.p larger than $1-\delta$
Define target function $c_h^0(s,a) = 1$ for all (h,s,a)
Execute CovGame$\left(c^0,\ \delta/4\right)$ to get dataset $\mathcal{D}_0$ and number of episodes d_0 // Burn-in phase
Initialize episode count $t_0 \leftarrow d_0$ and statistics $n_h^0(s,a), \widehat{r}_h^0(s,a), \widehat{p}_h^0(.|s,a)$ using $\mathcal{D}_0$
Initialize the set of active distributions $\Omega^0 \leftarrow \Omega(\widehat{p}^0)$
for $k = 1, \dots$ **do**
  // Proportional Coverage
  Compute $c_h^k(s,a) := 2^k \min\Big(\sup_{\widehat{\rho}\in\Omega^{k-1}} \widehat{\rho}_h(s,a) + 2\sqrt{H\beta^{bpi}(t_{k-1}+SAH2^k,\delta/2)2^{1-k}},\ 1\Big)$ for all (h,s,a)
  Execute CovGame$\left(c^k,\ \delta/4(k+1)^2\right)$ to get dataset $\widetilde{\mathcal{D}}_k$ and number of episodes T_k
  if $T_k > SAH2^k$ **then**
    Run PruneDataset$(\widetilde{\mathcal{D}}_k, c^k)$ to get *effective* dataset $\mathcal{D}_k$ and *effective phase length* d_k
  else
    Set $d_k \leftarrow T_k$ and $\mathcal{D}_k \leftarrow \widetilde{\mathcal{D}}_k$
  end if
  Update *effective episode count* $t_k \leftarrow t_{k-1} + d_k$ and statistics $n_h^k(s,a), \hat{r}_h^k(s,a), \widehat{p}_h^k(.|s,a)$ using $\mathcal{D}_k$
  // State-action-distribution elimination
  Compute the lower confidence bound

$$\underline{V}_1^k := \sup_{\substack{\widehat{\rho}\in\Omega(\widehat{p}^k),\\ \max_{h,s,a} \widehat{\rho}_h(s,a)/n_h^k(s,a)\le 2^{-k}}} \widehat{\rho}^{\top}\widehat{r}^k - \sqrt{2^{2-k}H\beta^{bpi}(t_k,\delta/2)}$$

  Update the set of active state-action distributions

$$\Omega^k \leftarrow \left\{\widehat{\rho}\in\Omega(\widehat{p}^k) :\ \widehat{\rho}^{\top}\widehat{r}^k \ge \underline{V}_1^k \text{ and } \max_{h,s,a} \widehat{\rho}_h(s,a)/n_h^k(s,a) \le 2^{-k}\right\}$$

  if $\sqrt{2^{2-k}H\beta^{bpi}(t_k,\delta/2)} \le \varepsilon$ **then**
    Compute any $\widehat{\rho}^{\star} \in \arg\max_{\widehat{\rho}\in\Omega^k} \widehat{\rho}^{\top}\widehat{r}^k$ and extract the corresponding policy $\widehat{\pi}$
    return $\widehat{\pi}$
  end if
end for

Algorithm 13 PruneDataset

Input: Target counts c, Dataset $\widetilde{\mathcal{D}}$ such that $n_h(s,a;\widetilde{\mathcal{D}}) \geq c_h(s,a)$ for all (h,s,a)
Output: A dataset $\mathcal{D}$ of $d \leq SAH2^k$ episodes satisfying $n_h(s,a;\mathcal{D}) \geq c_h(s,a)$ for all (h,s,a)
Initialize dataset $\mathcal{D} \leftarrow \emptyset$, episode number $d \leftarrow 0$ and dataset-counts $n_h(s,a;\mathcal{D}) \leftarrow 0$ for all (h,s,a)
for episode $e = (s_\ell^e, a_\ell^e, R_\ell^e)_{1\leq \ell \leq H}$ in $\widetilde{\mathcal{D}}$ **do**
  if $\exists \ell \in [H]$ such that $n_\ell(s_\ell^e, a_\ell^e; \mathcal{D}) < c_\ell(s_\ell^e, a_\ell^e)$ **then**
    Update dataset-counts $n_h(s_h^e, a_h^e; \mathcal{D}) \leftarrow n_h(s_h^e, a_h^e; \mathcal{D}) + 1$ for all $h \in [H]$
    Update dataset $\mathcal{D} \leftarrow \mathcal{D} \cup \{e\}$ and episode number $d \leftarrow d+1$
    if $n_h(s,a;\mathcal{D}) \geq c_h(s,a)$ for all (h,s,a) **then**
      return $(\mathcal{D}, d)$
    end if
  end if
end for

Remark 4.3 — Reachability. While for the PCE algorithm we were able to reduce the sample complexity by ignoring states that are hard to reach (which also allows using PCE when Assumption 3.1 is violated), we did not manage to propose a similar improvement for PRINCIPLE. This is because in reward-free exploration it is sufficient to guarantee that the *true confidence intervals* that depend on the visitation probabilities *under the true MDP* are small, i.e., $\sqrt{\beta^{\mathrm{RF}}(t_k,\delta)\sum_{(h,s,a)} \frac{p_h^\pi(s,a)^2}{n_h^k(s,a)}} \leq 2^k$. This allows us to filter out all (h,s,a) for which $\sup_\pi p_h^\pi(s,a) \leq \mathcal{O}(\varepsilon/SH^2)$, by arguing that their contribution to the true confidence interval is negligible. In contrast, the analysis of PRINCIPLE crucially relies on concentrating the values of policies by minimizing *their empirical confidence intervals*, i.e., $\sqrt{\beta^{bpi}(t_k,\delta)\sum_{(h,s,a)} \frac{\widehat{p}_h^{\pi,k}(s,a)^2}{n_h^k(s,a)}} \leq 2^k$ (see (4.8) and the proof of Lemma4.7). We do not see a straightforward way to ignore the contribution of hard-to-reach states to these empirical confidence intervals. ∎

4.3.4 Comparison with other BPI-algorithms

In this section, we compare PRINCIPLE with other algorithms for Best-Policy Identification algorithms that enjoy problem-dependent guarantees, namely PEDEL (Wagenmaker & Jamieson, 2022) and MOCA (Wagenmaker et al., 2022a). Recalling that $\Delta(\pi) = V_1^\star - V_1^\pi$ denotes the policy gap of π, we first note that by Theorem 4.3, the leading term in the sample complexity of PRINCIPLE in the small (ε,δ) regime is $\mathrm{PRINCIPLE}(\mathcal{M},\varepsilon)\log(1/\delta)$ where

$$\mathrm{PRINCIPLE}(\mathcal{M},\varepsilon) := H^3\varphi^\star\left(\left[\sup_{\pi\in\Pi_{\mathrm{S}}} \frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right).$$

We will now compare this term with the leading terms in the sample complexities of PEDEL and MOCA respectively, in the same asymptotic regime.

4.3.4.1 Comparison with PEDEL

The next lemma shows that (up to H factors) the rate of PEDEL is always better than the complexity measure achieved by PRINCIPLE. Recall that this comes at the cost of an intractable algorithm.

Lemma 4.2 For any MDP $\mathcal{M}$, it holds that $\mathrm{PEDEL}(\mathcal{M}, \varepsilon) \leq H^2 \mathrm{PRINCIPLE}(\mathcal{M}, \varepsilon)$.

Proof. Fix any $h \in [H], \rho \in \Omega, \pi \in \Pi_{\mathrm{D}}$. Then we have

$$\sum_{s,a} \frac{p_h^\pi(s,a)^2}{\rho_h(s,a)} \leq \left(\max_{s,a,h} \frac{p_h^\pi(s,a)}{\rho_h(s,a)} \right) \sum_{s,a} p_h^\pi(s,a) = \max_{s,a,h} \frac{p_h^\pi(s,a)}{\rho_h(s,a)}.$$

Therefore for all $h \in [H]$, using that $\Pi_{\mathrm{D}} \subset \Pi_{\mathrm{S}}$ we have

$$\begin{aligned}
\min_{\rho \in \Omega} \max_{\pi \in \Pi_{\mathrm{D}}} \sum_{s,a} \frac{p_h^\pi(s,a)^2/\rho_h(s,a)}{\max(\varepsilon, \Delta(\pi), \Delta_{\min}(\Pi_{\mathrm{D}}))^2} &\leq \min_{\rho \in \Omega} \max_{\pi \in \Pi_{\mathrm{D}}} \max_{s,a,h} \frac{p_h^\pi(s,a)/\rho_h(s,a)}{\max(\varepsilon, \Delta(\pi), \Delta_{\min}(\Pi_{\mathrm{D}}))^2} \\
&= \min_{\rho \in \Omega} \max_{s,a,h} \max_{\pi \in \Pi_{\mathrm{D}}} \frac{p_h^\pi(s,a)/\rho_h(s,a)}{\max(\varepsilon, \Delta(\pi), \Delta_{\min}(\Pi_{\mathrm{D}}))^2} \\
&\leq \min_{\rho \in \Omega} \max_{s,a,h} \sup_{\pi \in \Pi_{\mathrm{S}}} \frac{p_h^\pi(s,a)}{\rho_h(s,a) \max(\varepsilon, \Delta(\pi))^2} \\
&= \varphi^\star \left(\left[\sup_{\pi \in \Pi_{\mathrm{S}}} \frac{p_h^\pi(s,a)}{\max(\varepsilon, \Delta(\pi))^2} \right]_{h,s,a} \right).
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathrm{PEDEL}(\mathcal{M}, \varepsilon) &:= H^4 \sum_{h=1}^{H} \min_{\rho \in \Omega} \max_{\pi \in \Pi_{\mathrm{D}}} \sum_{s,a} \frac{p_h^\pi(s,a)^2/\rho_h(s,a)}{\max(\varepsilon, \Delta(\pi), \Delta_{\min}(\Pi_{\mathrm{D}}))^2} \\
&\leq H^5 \varphi^\star \left(\left[\sup_{\pi \in \Pi_{\mathrm{S}}} \frac{p_h^\pi(s,a)}{\max(\varepsilon, \Delta(\pi))^2} \right]_{h,s,a} \right) \\
&= H^2 \mathrm{PRINCIPLE}(\mathcal{M}, \varepsilon).
\end{aligned}$$

■

4.3.4.2 Comparison with MOCA

Let us define the complexity functional,

$$\begin{aligned}
\mathrm{MOCA}(\mathcal{M}, \varepsilon) := H^2 \sum_{h=1}^{H} \min_{\rho \in \Omega} \max_{s,a} \frac{1}{\rho_h(s,a)} \min \left(\frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\varepsilon^2} \right) \\
+ \frac{H^4 \left| (h,s,a) : \ \widetilde{\Delta}_h(s,a) \leq 3\varepsilon / W_h(s) \right|}{\varepsilon^2},
\end{aligned}$$

where $W_h(s) := \sup_\pi p_h^\pi(s)$ is the reachability of (h,s) and

$$\widetilde{\Delta}_h(s,a) := \begin{cases} \min_{b \neq a} V_h^\star(s) - Q_h^\star(s,b) & \text{if } a \text{ is the unique optimal action at } (h,s), \\ V_h^\star(s) - Q_h^\star(s,a) & \text{otherwise} \end{cases}$$

is the value gap of (h,s,a). Theorem 1 together with Proposition 2 of (Wagenmaker et al., 2022a) yield that the stopping time of MOCA satisfies

$$\tau \leq \widetilde{\mathcal{O}} \left(\mathrm{MOCA}(\mathcal{M}, \varepsilon) \log(1/\delta) + \frac{\mathrm{poly}\big(SAH, \log(1/\varepsilon), \log(1/\delta)\big)}{\varepsilon} \right).$$

Therefore we see that $\mathrm{MOCA}(\mathcal{M}, \varepsilon) \log(1/\delta)$ is the dominating term in the sample complexity of MOCA in the regime of small ε and small δ. On the other hand, as stated earlier, the leading term in PRINCIPLE's complexity in that regime is $\mathrm{PRINCIPLE}(\mathcal{M}, \varepsilon) \log(1/\delta)$. Therefore we compare $\mathrm{MOCA}(\mathcal{M}, \varepsilon)$ with $\mathrm{PRINCIPLE}(\mathcal{M}, \varepsilon)$ to assess which algorithm is better in this regime.

Lemma 4.3 Fix any $\Delta \in (0,1]$. There exists an MDP $\mathcal{M}$ where

$$\mathrm{MOCA}(\mathcal{M},\varepsilon) = \Omega\left(\frac{H^5SA}{\varepsilon^2}\right) \text{ while } \mathrm{PRINCIPLE}(\mathcal{M},\varepsilon) = \mathcal{O}\left(\frac{H^4SA}{\varepsilon\Delta} + \frac{H^4\log(S)\log(A)}{\varepsilon^2}\right).$$

Proof. Consider the MDP in Figure 4.1 which consists of an initial state s_1 and two sub-MDPs depending on the action taken at step $h = 1$. If the learner takes action a_1 it receives a reward $\Delta > 0$ and makes a transition to a sub-MDP $\mathcal{M}_1$ for which $|\mathcal{S}_1| = \log(S), |\mathcal{A}_1| = \log(A), H_1 = H - 1$ and where the rewards can be anything. On the other hand, if it takes action a_2 the learner will receive zero reward and make a transition to a sub-MDP $\mathcal{M}_2$ for which $|\mathcal{S}_2| = S - \log(S), |\mathcal{A}_2| = A, H_2 = H - 1$, the rewards are equal to zero everywhere and the transitions are deterministic, i.e. $p(s'|s,a) \in \{0,1\}$ for all $(s,a) \in \mathcal{S}_2 \times \mathcal{A}_2$. Note that in this example $\widetilde{\Delta}_h(s,a) = 0$ for all $(h,s,a) \in \mathcal{M}_2$. Therefore

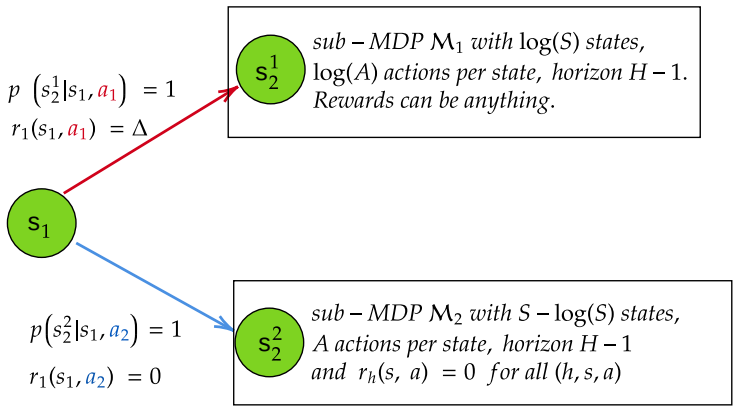


Figure 4.1: MDP instance with large policy gaps and small value gaps.

$$\begin{aligned}\mathrm{MOCA}(\mathcal{M},\varepsilon) &\geq \frac{H^4\left|(h,s,a) : \widetilde{\Delta}_h(s,a) \leq 3\varepsilon/W_h(s)\right|}{\varepsilon^2}, \\ &\geq \frac{H^4(H-1)(S-\log(S))A}{\varepsilon^2}. \end{aligned} \tag{4.2}$$

On the other hand for all triplets (h,s,a) in the sub-MDP $\mathcal{M}_2$ we have

$$\sup_{\pi\in\Pi_S} \frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2} \leq \sup_{\pi\in\Pi_S} \frac{4\pi_1(a_2|s_1)}{(\varepsilon+\Delta(\pi))^2}, \tag{4.3}$$

where we used that $p_h^\pi(s,a) \leq \pi_1(a_2|s_1)$ (since the only path to reach (h,s,a) is by playing action a_2 at s_1) and that $\max(a,b) \geq (a+b)/2$. Now, by the performance-difference lemma ((e.g. Lemma 5.2.1 in Kakade, 2003)) we have

$$\begin{aligned}\Delta(\pi) &= \sum_{h,s,a} p_h^\pi(s,a)[V_h^\star(s) - Q_h^\star(s,a)] \\ &\geq p_1^\pi(s_1,a_2)[V_1^\star(s_1) - Q_1^\star(s_1,a_2)] = \pi_1(a_2|s_1)\Delta.\end{aligned}$$

Plugging this back into (4.3), we get

$$\sup_{\pi\in\Pi_S}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\le\sup_{\pi\in\Pi_S}\frac{4\pi_1(a_2|s_1)}{(\varepsilon+\pi_1(a_2|s_1)\Delta)^2}$$
$$=\sup_{x\in[0,1]}\frac{4x}{(\varepsilon+x\Delta)^2}=\frac{1}{\varepsilon\Delta}$$

For triplets (h,s,a) outside of $\mathcal{M}_2$ (i.e. either at s_1 or in the sub-MDP $\mathcal{M}_1$) we use the crude bound

$$\sup_{\pi\in\Pi_S}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\le\frac{\sup_{\pi\in\Pi_S}p_h^\pi(s,a)}{\varepsilon^2}.$$

Therefore

$$\begin{aligned}
&\text{PRINCIPLE}(\mathcal{M},\varepsilon)=H^3\varphi^\star\left(\left[\sup_{\pi\in\Pi_S}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)\\
&=H^3\varphi^\star\left(\left[\sup_{\pi\in\Pi_S}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}(\mathbb{1}\left((h,s,a)\in\mathcal{M}_2\right)+\mathbb{1}\left((h,s,a)\notin\mathcal{M}_2\right))\right]_{h,s,a}\right)\\
&\overset{(a)}{\le}H^3\varphi^\star\left(\left[\sup_{\pi\in\Pi_S}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\mathbb{1}\left((h,s,a)\in\mathcal{M}_2\right)\right]_{h,s,a}\right)\\
&\quad+H^3\varphi^\star\left(\left[\sup_{\pi\in\Pi_S}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\mathbb{1}\left((h,s,a)\notin\mathcal{M}_2\right)\right]_{h,s,a}\right)\\
&\le H^3\varphi^\star\left(\left[\frac{\mathbb{1}\left((h,s,a)\in\mathcal{M}_2\right)}{\varepsilon\Delta}\right]_{h,s,a}\right)+H^3\varphi^\star\left(\left[\frac{\mathbb{1}\left((h,s,a)\notin\mathcal{M}_2\right)\sup_{\pi\in\Pi_S}p_h^\pi(s,a)}{\varepsilon^2}\right]_{h,s,a}\right)\\
&\overset{(b)}{\le}H^3\sum_{(h,s,a)\in\mathcal{M}_2}\frac{1}{\varepsilon\Delta\sup_{\pi\in\Pi_S}p_h^\pi(s,a)}+H^3\sum_{(h,s,a)\notin\mathcal{M}_2}\frac{1}{\varepsilon^2}\\
&\overset{(c)}{=}\frac{H^3(H-1)(S-\log(S))A}{\varepsilon\Delta}+\frac{H^3(H-1)\log(S)\log(A)}{\varepsilon^2}
\end{aligned}\tag{4.4}$$

where (a) uses the sub-linearity of the flow from Lemma 3.12, (b) uses the bound on $\varphi^\star$ from Lemma 3.2 and (c) uses that the sub-MDP $\mathcal{M}_2$ has deterministic transitions. Combining (4.2) and (4.4) finishes the proof. ■

4.4 Analysis of PRINCIPLE

4.4.1 Good event

We introduce the following events

$$\mathcal{E}_{bpi}:=\Bigg(\forall k\in\mathbb{N}^\star,\forall\pi\in\Pi_S,\ \left|\widehat{V}_1^{\pi,k}-V_1^\pi\right|\le\sqrt{\beta^{bpi}(t_k,\delta/2)\min\left(\sum_{s,a,h}\frac{p_h^\pi(s,a)^2}{n_h^k(s,a)},\sum_{s,a,h}\frac{\widehat{p}_h^{\pi,k}(s,a)^2}{n_h^k(s,a)}\right)}$$
$$\text{and}\ \left|\sum_{s,a,h}\left(\widehat{p}_h^{\pi,k}(s,a)-p_h^\pi(s,a)\right)\widetilde{r}_h(s,a)\right|\le\sqrt{\beta^{bpi}(t_k,\delta/2)\sum_{s,a,h}\frac{p_h^\pi(s,a)^2}{n_h^k(s,a)}}\ \text{for all}\ \widetilde{r}\in[0,1]^{SAH}\Bigg),$$

$$\mathcal{E}_{cov}:=\Big(\forall k\in\mathbb{N},\ \text{CovGame run with inputs}\ (c^k,\delta/4(k+1)^2)\ \text{terminates after at most}$$
$$64m_k\varphi^\star(c^k)+\widetilde{\mathcal{O}}\big(m_k\varphi^\star(\mathbb{1})SAH^2(\log(4(k+1)^2/\delta)+S)\big)\ \text{episodes and returns a dataset}\ \widetilde{\mathcal{D}}_k$$
$$\text{such that for all}\ (h,s,a),n_h(s,a;\widetilde{\mathcal{D}}_k)\ge c_h^k(s,a)\Big),$$

where $m_k = \log_2\left(\frac{\max_{s,a,h} c_h^k(s,a)}{\min_{s,a,h} c_h^k(s,a)\vee 1}\right) \vee 1$ and $\beta^{bpi}(t,\delta) = 16H^2\log(2/\delta) + 96SAH^3\log(1+t)$ is defined in Appendix 3.9. Then our good event is defined as the intersection

$$\mathcal{E}_{good} := \mathcal{E}_{bpi} \cap \mathcal{E}_{cov}.$$

Lemma 4.4 We have that $\mathbb{P}_{\mathcal{M}}(\mathcal{E}_{good}) \geq 1-\delta$.

Proof. Let $\overline{\mathcal{E}}$ denote the complementary event of $\mathcal{E}$. We start by the following decomposition

$$\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{good}}) \leq \mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{cov}}) + \mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{bpi}} \cap \mathcal{E}_{cov}).$$

Now we bound each term separately. First observe that using Corollary 3.1 we have

$$\begin{aligned}\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{cov}}) &\leq \sum_{k=0}^{\infty} \mathbb{P}_{\mathcal{M}}(\text{CovGame with inputs } (c^k, \delta/4(k+1)^2) \text{ fails}) \\ &\leq \sum_{k=0}^{\infty} \frac{\delta}{4(k+1)^2} = \frac{\delta\pi^2}{24} \leq \delta/2.\end{aligned}$$

Next, note that by design of PRINCIPLE $n_h^0(s,a) = n_h(s,a;\widetilde{\mathcal{D}}_0)$ and $c^0 = \mathbb{1}$ so that $\mathcal{E}_{cov} \subset \big(\forall(h,s,a),\ n_h^0(s,a) \geq 1\big)$. Therefore we have

$$\begin{aligned}\mathbb{P}_{\mathcal{M}}(\overline{\mathcal{E}_{bpi}} \cap \mathcal{E}_{cov}) &\leq \mathbb{P}_{\mathcal{M}}\big(\overline{\mathcal{E}_{bpi}} \text{ and } \forall(h,s,a)\ n_h^0(s,a) \geq 1\big) \\ &\leq \delta/2,\end{aligned}$$

where we applied Theorem 3.5 and used the fact that $\beta^p(t,\delta) \leq \beta^{bpi}(t,\delta)$. Combining the two inequalities above yields the desired result. ■

4.4.2 Low Concentrability / Good coverage of optimal policies

Lemma 4.5 Under the good event, for all $k \geq 1$ such that PRINCIPLE did not stop before phase k, it holds that $n_h(s,a,\mathcal{D}_k) \geq c_h^k(s,a)$ for all (h,s,a) and $d_k \leq SAH2^k$.

Proof. Fix $k \geq 1$ such that PRINCIPLE did not stop before phase k. By definition of the good event, we know that at the end of CovGame, $n_h(s,a;\widetilde{\mathcal{D}}_k) \geq c_h^k(s,a)$ for all (h,s,a). Now we distinguish two cases.

If $T_k \leq SAH2^k$: then the result follows immediately since in this case, by design of PRINCIPLE (line 14 in Algorithm 12), $\mathcal{D}_k = \widetilde{\mathcal{D}}_k$ and $d_k = T_k$.

If $T_k > SAH2^k$: the first statement is a direct consequence of the stopping condition of PruneDataset run with parameters $(\widetilde{\mathcal{D}}_k, c^k)$ (lines 7-8 in Algorithm 13). Now for the second statement, observe that each new episode e added by PruneDataset to $\mathcal{D}_k$ increments the dataset-count of at least one triplet (h,s,a) that is not yet covered, i.e. $n_h(s,a;\mathcal{D}_k) < c_h^k(s,a)$. By the pigeon-hole principle it takes at most $\sum_{h,s,a} c_h^k(s,a)$ episodes to ensure that $n_h(s,a,\mathcal{D}_k) \geq c_h^k(s,a)$ for all (h,s,a). Therefore

$$d_k \leq \sum_{h,s,a} c_h^k(s,a) \leq SAH2^k,$$

where we used that $c_h^k(s,a) \leq 2^k$ due to the clipping. ■

The next lemma shows that the set of active state-action distributions always contains the distributions induced by optimal policies.

Lemma 4.6 Under the good event, for all optimal policies $\pi^\star \in \Pi^\star$ and all phases $k \geq 0$, we have that

$$\widehat{p}^{\pi^\star,k} \in \Omega^k \quad \text{and } n_h^k(s,a) \geq 2^k p_h^{\pi^\star}(s,a) \quad \forall (h,s,a).$$

Proof. We fix an optimal policy $\pi^\star$ and prove the statement by induction. For $k=0$, the fact that $\widehat{p}^{\pi^\star,0} \in \Omega^0$ is trivial since $\Omega^0 = \Omega(\widehat{p}^0)$ consists of all possible state-action distributions induced in the MDP whose transition kernel is $\widehat{p}^0$. Furthermore, under the good event we have that, for all (h,s,a), $n_h^0(s,a) \geq c_h^0(s,a) = 1 \geq 2^0 \max\big(p_h^{\pi^\star}(s,a), \widehat{p}_h^{\pi^\star,0}(s,a)\big)$. Now suppose that the property holds for phase k. Then we know that for any (h,s,a)

$$\begin{aligned}
\big|\widehat{p}_h^{\pi^\star,k+1}(s,a) - \widehat{p}_h^{\pi^\star,k}(s,a)\big| &\leq \big|\widehat{p}_h^{\pi^\star,k+1}(s,a) - p_h^{\pi^\star}(s,a)\big| + \big|p_h^{\pi^\star}(s,a) - \widehat{p}_h^{\pi^\star,k}(s,a)\big| \\
&\overset{(a)}{\leq} \sqrt{\beta^{bpi}(t_{k+1},\delta/2) \sum_{s,a,h} \frac{p_h^{\pi^\star}(s,a)^2}{n_h^{k+1}(s,a)}} + \sqrt{\beta^{bpi}(t_k,\delta/2) \sum_{s,a,h} \frac{p_h^{\pi^\star}(s,a)^2}{n_h^{k}(s,a)}} \\
&\overset{(b)}{\leq} 2\sqrt{\beta^{bpi}(t_{k+1},\delta/2) \sum_{s,a,h} \frac{p_h^{\pi^\star}(s,a)^2}{n_h^{k}(s,a)}} \\
&\overset{(c)}{\leq} 2\sqrt{\beta^{bpi}(t_{k+1},\delta/2) H 2^{-k}} \\
&= 2\sqrt{\beta^{bpi}(t_k + d_{k+1},\delta/2) H 2^{-k}} \\
&\overset{(d)}{\leq} 2\sqrt{\beta^{bpi}(t_k + SAH2^{k+1},\delta/2) H 2^{-k}},
\end{aligned} \tag{4.5}$$

where (a) uses the event $\mathcal{E}_{bpi}$ for the reward $\widetilde{r}_\ell(s',a') = \mathbb{1}\big((\ell,s',a') = (h,s,a)\big)$, (b) uses the facts that $t \mapsto \beta(t,\delta)$ is non-decreasing and $n_h^{k+1}(s,a) \geq n_h^k(s,a)$, (c) uses the induction hypothesis which yields that $n_h^k(s,a) \geq 2^k p_h^{\pi^\star}(s,a)$ and (d) uses Lemma 4.5. Similarly, we have that

$$\big|p_h^{\pi^\star}(s,a) - \widehat{p}_h^{\pi^\star,k}(s,a)\big| \leq \sqrt{\beta^{bpi}(t_k + SAH2^{k+1},\delta/2) H 2^{-k}} \tag{4.6}$$

Now thanks to Lemma 4.5, we know that for all (h,s,a), $n_h^{k+1}(s,a) - n_h^k(s,a) = n_h(s,a,\mathcal{D}_{k+1}) \geq c_h^{k+1}(s,a)$. Plugging the definition of c^{k+1} (Line 9 of Algorithm 12) we get that,

$$\begin{aligned}
n_h^{k+1}(s,a) &\geq 2^{k+1} \min\Big(\sup_{\widehat{\rho} \in \Omega^k} \widehat{\rho}_h(s,a) + 2\sqrt{H\beta^{bpi}(t_k + SAH2^{k+1},\delta/2)2^{-k}},\ 1\Big) \\
&\overset{(a)}{\geq} 2^{k+1} \min\Big(\widehat{p}_h^{\pi^\star,k}(s,a) + 2\sqrt{H\beta^{bpi}(t_k + SAH2^{k+1},\delta/2)2^{-k}},\ 1\Big) \\
&\overset{(b)}{\geq} 2^{k+1} \max\big(\widehat{p}_h^{\pi^\star,k+1}(s,a),\ p_h^{\pi^\star}(s,a)\big),
\end{aligned} \tag{4.7}$$

where (a) uses that, by the induction hypothesis, $\widehat{p}^{\pi^\star,k} \in \Omega^k$ and (b) uses (4.5) along with (4.6). In particular we have proved that $\max_{h,s,a} \widehat{p}_h^{\pi^\star,k+1}(s,a)/n_h^{k+1}(s,a) \leq 2^{-(k+1)}$. Now it remains to show that $(\widehat{p}^{\pi^\star,k+1})^\top \widehat{r}^{k+1} \geq \underline{V}_1^{k+1}$. Let us consider $\widetilde{\rho}$ achieving the supremum in the definition of $\underline{V}_1^{k+1}$, i.e.,

$$\widetilde{\rho} \in \operatorname*{arg\,max}_{\substack{\widehat{\rho} \in \Omega(\widehat{p}^{k+1}),\\ \max_{h,s,a} \widehat{\rho}_h(s,a)/n_h^{k+1}(s,a) \leq 2^{-(k+1)}}} \widehat{\rho}^\top \widehat{r}^{k+1},$$

and let $\widetilde{\pi}$ be a policy corresponding to $\widetilde{\rho}$[2]. Then we have that

$$
\begin{aligned}
(\widehat{p}^{\pi^\star,k+1})^\top \widehat{r}^{k+1} &\overset{(a)}{\geq} V_1^\star - \sqrt{\beta^{bpi}(t_{k+1},\delta/2)\sum_{s,a,h}\frac{p_h^{\pi^\star}(s,a)^2}{n_h^{k+1}(s,a)}} \\
&\geq V_1^{\widetilde{\pi}} - \sqrt{\beta^{bpi}(t_{k+1},\delta/2)\sum_{s,a,h}\frac{p_h^{\pi^\star}(s,a)^2}{n_h^{k+1}(s,a)}} \\
&\overset{(b)}{\geq} \widetilde{\rho}^\top \widehat{r}^{k+1} - \sqrt{\beta^{bpi}(t_{k+1},\delta/2)\sum_{s,a,h}\frac{\widetilde{\rho}_h(s,a)^2}{n_h^{k+1}(s,a)}} - \sqrt{\beta^{bpi}(t_{k+1},\delta/2)\sum_{s,a,h}\frac{p_h^{\pi^\star}(s,a)^2}{n_h^{k+1}(s,a)}} \\
&\overset{(c)}{\geq} \widetilde{\rho}^\top \widehat{r}^{k+1} - 2\sqrt{2^{-(k+1)}H\beta^{bpi}(t_{k+1},\delta/2)} \\
&= \underline{V}_1^{k+1}
\end{aligned}
\tag{4.8}
$$

where (a) uses the event $\mathcal{E}_{bpi}$ for policy $\pi^\star$, (b) uses the same event combined with the fact that $\widetilde{\rho} = \widehat{p}^{\widetilde{\pi},k+1}$, and (c) uses (4.7) and the fact that by definition of $\widetilde{\rho}$, $\max_{h,s,a}\widetilde{\rho}_h(s,a)/n_h^{k+1}(s,a) \leq 2^{-(k+1)}$. Now combining (4.7) with (4.8) gives that $\widehat{p}^{\pi^\star,k+1} \in \Omega^{k+1}$. This finishes the proof. ■

4.4.3 Correctness

Lemma 4.7 Under the good event, if PRINCIPLE stops then the recommended policy satisfies $V_1^{\widehat{\pi}} \geq V_1^\star - \varepsilon$.

Proof. Suppose that PRINCIPLE stops at phase $k \geq 1$. Let $\pi^\star$ be any optimal policy and recall the definition $\widehat{\rho}^\star = \arg\max_{\widehat{\rho}\in\Omega^k} \widehat{\rho}^\top \widehat{r}^k$ with ties broken arbitrarily. We have that

$$
\begin{aligned}
V_1^{\widehat{\pi}} &\overset{(a)}{\geq} (\widehat{\rho}^\star)^\top \widehat{r}^k - \sqrt{\beta^{bpi}(t_k,\delta/2)\sum_{s,a,h}\frac{\widehat{\rho}_h^\star(s,a)^2}{n_h^k(s,a)}} \\
&\overset{(b)}{\geq} (\widehat{p}^{\pi^\star,k})^\top \widehat{r}^k - \sqrt{\beta^{bpi}(t_k,\delta/2)\sum_{s,a,h}\frac{\widehat{\rho}_h^\star(s,a)^2}{n_h^k(s,a)}} \\
&\overset{(c)}{\geq} V_1^\star - \sqrt{\beta^{bpi}(t_k,\delta/2)\sum_{s,a,h}\frac{\widehat{p}_h^{\pi^\star,k}(s,a)^2}{n_h^k(s,a)}} - \sqrt{\beta^{bpi}(t_k,\delta/2)\sum_{s,a,h}\frac{\widehat{\rho}_h^\star(s,a)^2}{n_h^k(s,a)}} \\
&\overset{(d)}{\geq} V_1^\star - 2\sqrt{2^{-k}H\beta^{bpi}(t_k,\delta/2)} \overset{(e)}{\geq} V_1^\star - \varepsilon,
\end{aligned}
$$

where (a) uses the event $\mathcal{E}_{bpi}$ for policy $\widehat{\pi}$ and the fact that $\widehat{\rho}^\star = \widehat{p}^{\widehat{\pi},k}$, (b) uses the definition of $\widehat{\rho}^\star$ and the fact that, by Lemma 4.6, $\widehat{p}^{\pi^\star,k} \in \Omega^k$, (c) uses the event $\mathcal{E}_{bpi}$ for the policy $\pi^\star$, and (d) uses that for all $\rho \in \Omega^k$, $\max_{h,s,a}\rho_h(s,a)/n_h^k(s,a) \leq 2^{-k}$ and (e) uses the stopping condition of PRINCIPLE (Line 19 of Algorithm 12). ■

4.4.3.1 Upper bound on the number of phases

Lemma 4.8 Define the index of the final phase of PRINCIPLE, $\kappa_f := \inf\left\{k \in \mathbf{N}_+ : \sqrt{2^{2-k}H\beta^{bpi}(t_k,\delta/2)} \leq \varepsilon\right\}$. Further, let τ denote the number of episodes played by the

[2] i.e. $\widetilde{\pi}$ is the policy obtained by normalization of $\widetilde{\rho}$.

algorithm. Then under the good event, it holds that $\kappa_f < \infty$ and

$$2^{\kappa_f} \leq \frac{8H\beta^{bpi}(\tau, \delta/2)}{\varepsilon^2}.$$

Proof. To prove that κ_f is finite we write

$$\begin{aligned} t_k &= \sum_{j=0}^{k} d_j \\ &\leq d_0 + SAH \sum_{j=1}^{k} 2^j \\ &\leq \widetilde{\mathcal{O}}\Big(\varphi^\star(\mathbb{1})SAH^2\big(\log(4/\delta) + S\big)\Big) + SAH2^{k+1}, \end{aligned} \tag{4.9}$$

where we have used the coverage event $\mathcal{E}_{cov}$ and Lemma 4.5 to upper bound d_0 and $(d_k)_{1\leq j\leq k}$ respectively. This means that $t_k = \mathcal{O}_{k\to\infty}\big(2^k\big)$. Now recall that

$$\beta^{bpi}(t, \delta) = 16H^2 \log(1/\delta) + 96SAH^3 \log(1 + t). \tag{4.10}$$

Combining (4.9) and (4.10) gives that

$$\beta^{bpi}(t_k, \delta/2) = o_{k\to\infty}\big(2^k\big).$$

Therefore $\kappa_f = \inf\big\{k \in \mathbf{N}_+ : \sqrt{2^{2-k}H\beta^{bpi}(t_k, \delta/2)} \leq \varepsilon\big\}$ is indeed finite. The proof of the second statement is straightforward by noting that $\kappa_f - 1$ does not satisfy the stopping condition (Line 19 in Algorithm 12) and using the (crude) upper bound $t_{\kappa_f - 1} \leq \tau$. ■

Lemma 4.9 (Upper bound on phases where a suboptimal policy is active) Consider any suboptimal policy $\pi \in \Pi_S$. Further let k such that PRINCIPLE did not stop at phase k and $\widehat{p}^{\pi,k} \in \Omega^k$. Further, let τ denote the number of episodes played by the algorithm. Then under the good event, we have the inequality

$$2^k \leq \frac{16H\beta^{bpi}(\tau, \delta/2)}{\max(\varepsilon, \Delta(\pi))^2},$$

where $\Delta(\pi) := V_1^\star - V_1^\pi$ denotes the policy gap of π.

Proof. Let $\pi^\star$ be any optimal policy. Then we have

$$\begin{aligned} V_1^\star - \sqrt{\beta^{bpi}(t_k, \delta/2) \sum_{s,a,h} \frac{\widehat{p}_h^{\pi^\star,k}(s,a)^2}{n_h^k(s,a)}} &\overset{(a)}{\leq} (\widehat{p}^{\pi^\star,k})^\top \widehat{r}^k \\ &\overset{(b)}{\leq} \sup_{\substack{\widehat{\rho}\in\Omega(\widehat{p}^k),\\ \max_{h,s,a} \widehat{\rho}_h(s,a)/n_h^k(s,a)\leq 2^{-k}}} \widehat{\rho}^\top \widehat{r}^k \\ &= \underline{V}_1^{\star,k} + \sqrt{2^{2-k}H\beta^{bpi}(t_k, \delta/2)} \\ &\overset{(c)}{\leq} (\widehat{p}^{\pi,k})^\top \widehat{r}^k + \sqrt{2^{2-k}H\beta^{bpi}(t_k, \delta/2)} \\ &\overset{(d)}{\leq} V_1^\pi + \sqrt{\beta^{bpi}(t_k, \delta/2) \sum_{s,a,h} \frac{\widehat{p}_h^{\pi,k}(s,a)^2}{n_h^k(s,a)}} + \sqrt{2^{2-k}H\beta^{bpi}(t_k, \delta/2)}, \end{aligned}$$

where (a) uses the event $\mathcal{E}_{bpi}$ for $\pi^\star$, (b) uses the definition of Ω^k along with Lemma 4.6 which gives that $\widehat{p}^{\pi^\star,k} \in \Omega^k$, (c) uses our assumption that $\widehat{p}^{\pi,k} \in \Omega^k$ and (d) uses the event $\mathcal{E}_{bpi}$ for policy π. Rewriting the inequality above we get that

$$\begin{aligned}\Delta(\pi) &= V_1^\star - V_1^\pi \\ &\leq \sqrt{\beta^{bpi}(t_k,\delta/2)\sum_{s,a,h}\frac{\widehat{p}_h^{\pi^\star,k}(s,a)^2}{n_h^k(s,a)}} + \sqrt{\beta^{bpi}(t_k,\delta/2)\sum_{s,a,h}\frac{\widehat{p}_h^{\pi,k}(s,a)^2}{n_h^k(s,a)}} + \sqrt{2^{2-k}H\beta^{bpi}(t_k,\delta/2)} \\ &\leq 2\sqrt{2^{-k}H\beta^{bpi}(t_k,\delta/2)} + \sqrt{2^{2-k}H\beta^{bpi}(t_k,\delta/2)} = 4\sqrt{2^{-k}H\beta^{bpi}(t_k,\delta/2)}, \qquad (4.11)\end{aligned}$$

where the last inequality uses the fact that $\widehat{p}^{\pi^\star,k} \in \Omega^k$ by Lemma 4.6 and that $\widehat{p}^{\pi,k} \in \Omega^k$ by assumption. Therefore, using a crude bound $t_k \leq \tau$ we get that

$$2^k \leq \frac{16H\beta^{bpi}(\tau,\delta/2)}{\Delta(\pi)^2}.$$

Combining the result above with Lemma 4.8 and the fact that $k \leq \kappa_f$ yields the final result. ■

4.4.3.2 Upper bound on the phase length

Lemma 4.10 Let T_k denote the number of episodes played by PRINCIPLE during phase $k \geq 1$. Then we have

$$\begin{aligned}T_k \leq\ & 256H\beta^{bpi}(\tau,\delta/2)k\varphi^\star\left(\left[\sup_{\pi\in\Pi_S}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right) \\ &+ 48k\sqrt{H\beta^{bpi}(t_{k-1}+SAH2^{k-1},\delta/2)2^k}\varphi^\star(\mathbb{1}) \\ &+ \widetilde{\mathcal{O}}\Big(k\varphi^\star(\mathbb{1})SAH^2\big(\log(4(k+1)^2/\delta)+S\big)\Big).\end{aligned}$$

Proof. Define $m_k = \log_2\left(\frac{\max_{s,a,h} c_h^k(s,a)}{\min_{s,a,h} c_h^k(s,a)\vee 1}\right) \vee 1$. Under the good event, we have

$$\begin{aligned}T_k &\leq 64m_k\varphi^\star(c^k) + \widetilde{\mathcal{O}}\Big(m_k\varphi^\star(\mathbb{1})SAH^2\big(\log(4(k+1)^2/\delta)+S\big)\Big) \\ &\leq 64k\varphi^\star(c^k) + \widetilde{\mathcal{O}}\Big(k\varphi^\star(\mathbb{1})SAH^2\big(\log(4(k+1)^2/\delta)+S\big)\Big), \qquad (4.12)\end{aligned}$$

where the last inequality uses the fact that for all (h,s,a), $c_h^k(s,a) \leq 2^k$. Now we simplify the expression of $\varphi^\star(c^k)$ as follows

$$\begin{aligned}\varphi^\star(c^k) &= \varphi^\star\left(\left[2^k\min\Big(\sup_{\widehat{\rho}\in\Omega^{k-1}}\widehat{\rho}_h(s,a) + 2\sqrt{H\beta^{bpi}(t_{k-1}+SAH2^{k-1},\delta/2)2^{1-k}},\ 1\Big)\right]_{h,s,a}\right) \\ &\leq \varphi^\star\left(\left[\sup_{\substack{\pi\in\Pi_S:\\ \widehat{p}^{\pi,k-1}\in\Omega^{k-1}}} 2^k\widehat{p}_h^{\pi,k-1}(s,a) + 2\sqrt{H\beta^{bpi}(t_{k-1}+SAH2^{k-1},\delta/2)2^{k+1}}\right]_{h,s,a}\right), \qquad (4.13)\end{aligned}$$

where we have used that $\varphi^\star(c) \leq \varphi^\star(c')$ if $\forall(h,s,a)\ c_h(s,a) \leq c'_h(s,a)$. Now fix a policy π in the set $\{\pi \in \Pi_S : \widehat{p}^{\pi,k-1} \in \Omega^{k-1}\}$. Using the event $\mathcal{E}_{bpi}$ for the rewards $\widetilde{r}_\ell(s',a') =$

$\mathbb{1}\big((\ell, s', a') = (h, s, a)\big)$ we have that for all (h, s, a)

$$
\begin{aligned}
2^k \widehat{p}_h^{\pi,k-1}(s,a) &\le 2^k p_h^\pi(s,a) + 2^k \sqrt{\beta^{bpi}(t_{k-1},\delta/2) \sum_{s',a',\ell} \frac{\widehat{p}_\ell^{\pi,k-1}(s',a')^2}{n_\ell^{k-1}(s',a')}} \\
&\overset{(a)}{\le} 2^k p_h^\pi(s,a) + 2^k \sqrt{\beta^{bpi}(t_{k-1},\delta/2) H 2^{1-k}} \\
&\le 2^k p_h^\pi(s,a) + \sqrt{H\beta^{bpi}(t_{k-1} + SAH2^{k-1},\delta/2) 2^{k+1}} \\
&\overset{(b)}{\le} \frac{32H\beta^{bpi}(\tau,\delta/2) p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2} + \sqrt{H\beta^{bpi}(t_{k-1} + SAH2^{k-1},\delta/2) 2^{k+1}},
\end{aligned}
$$

where (a) uses that $\max_{s',a',\ell} \frac{\widehat{p}_\ell^{\pi,k-1}(s',a')}{n_\ell^{k-1}(s',a')} \le 2^{1-k}$ since $\widehat{p}^{\pi,k-1} \in \Omega^{k-1}$ and (b) uses Lemma 4.9. Plugging the inequality above into (4.13) we get that

$$
\begin{aligned}
\varphi^\star(c^k) &\le \varphi^\star\left(\left[\sup_{\pi\in\Pi_S} \frac{32H\beta^{bpi}(\tau,\delta/2) p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2} + 3\sqrt{H\beta^{bpi}(t_{k-1} + SAH2^{k-1},\delta/2)2^{k+1}}\right]_{h,s,a}\right) \\
&\le 32H\beta^{bpi}(\tau,\delta/2)\varphi^\star\left(\left[\sup_{\pi\in\Pi_S} \frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right) \\
&\quad + 3\sqrt{H\beta^{bpi}(t_{k-1} + SAH2^{k-1},\delta/2)2^{k+1}}\varphi^\star(\mathbb{1}), \qquad (4.14)
\end{aligned}
$$

where we used Lemma 3.12 in the last step. Combining (4.12) and (4.14) finishes the proof. ■

4.4.3.3 Total sample complexity

Theorem 4.4 With probability at least $1-\delta$, the total sample complexity of PRINCIPLE satisfies

$$
\tau \le \widetilde{\mathcal{O}}\left((H^3\log(1/\delta) + SAH^4)\left[\varphi^\star\left(\left[\sup_{\pi\in\Pi_S} \frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right) + \frac{\varphi^\star(\mathbb{1})}{\varepsilon} + \varphi^\star(\mathbb{1})\right]\right),
$$

where $\widetilde{\mathcal{O}}$ hides poly-logarithmic factors in $S, A, H, \varepsilon, \log(1/\delta)$ and $\varphi^\star(\mathbb{1})$ and $\Delta(\pi) := V_1^\star - V_1^\pi$ denotes the policy gap of π.

Proof. We write

$$
\begin{aligned}
\tau &= \sum_{k=0}^{\kappa_f} T_k \\
&\le \widetilde{\mathcal{O}}\Big(\varphi^\star(\mathbb{1})^2 SAH^2\big(\log(4/\delta) + S\big)\Big) + \sum_{k=1}^{\kappa_f} T_k \\
&\le \underbrace{\sum_{k=1}^{\kappa_f} 256H\beta^{bpi}(\tau,\delta/2)k\varphi^\star\left(\left[\sup_{\pi\in\Pi_S} \frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)}_{:=A} \\
&\quad + \underbrace{\sum_{k=1}^{\kappa_f} 48k\sqrt{H\beta^{bpi}(t_{k-1} + SAH2^{k-1},\delta/2)2^k}\varphi^\star(\mathbb{1})}_{:=B} + \underbrace{\widetilde{\mathcal{O}}\Big(\textstyle\sum_{k=1}^{\kappa_f} k\varphi^\star(\mathbb{1})SAH^2\big(\log(4(k+1)^2/\delta) + S\big)\Big)}_{:=C},
\end{aligned}
$$

where we have used Lemma 4.10. Now we bound each term separately. First note that

$$\begin{aligned}
A &\leq 256H\beta^{bpi}(\tau,\delta/2)\varphi^\star\left(\left[\sup_{\pi\in\Pi_S}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)\kappa_f^2\\
&\overset{(a)}{\leq} 256H\beta^{bpi}(\tau,\delta/2)\varphi^\star\left(\left[\sup_{\pi\in\Pi_S}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)\log_2^2\left(8H\beta^{bpi}(\tau,\delta/2)/\varepsilon^2\right)\\
&\overset{(b)}{\leq} \mathcal{O}\left([H^3\log(1/\delta)+SAH^4\log(1+\tau)]\varphi^\star\left(\left[\sup_{\pi\in\Pi_S}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)\log_2^2\left(8H\beta^{bpi}(\tau,\delta/2)/\varepsilon^2\right)\right),
\end{aligned}$$

where (a) uses Lemma 4.8 and (b) uses the definition of β^{bpi}. Similarly

$$\begin{aligned}
B &\leq 48\sqrt{H\beta^{bpi}(\tau+SAH2^{\kappa_f-1},\delta/2)2^{\kappa_f}\varphi^\star(\mathbb{1})\kappa_f^2}\\
&\overset{(a)}{\leq} 48\sqrt{\frac{4H^2\beta^{bpi}(\tau+SAH2^{\kappa_f-1},\delta/2)\beta^{bpi}(\tau,\delta/2)}{\varepsilon^2}}\varphi^\star(\mathbb{1})\log_2^2\left(8H\beta^{bpi}(\tau,\delta/2)/\varepsilon^2\right)\\
&\leq \frac{48H}{\varepsilon}\beta^{bpi}(\tau+SAH2^{\kappa_f-1},\delta/2)\varphi^\star(\mathbb{1})\log_2^2\left(8H\beta^{bpi}(\tau,\delta/2)/\varepsilon^2\right)\\
&\overset{(b)}{\leq} \mathcal{O}\left(\frac{\varphi^\star(\mathbb{1})}{\varepsilon}\left[H^3\log(1/\delta)+SAH^4\log\left(1+\tau+\frac{4SAH^2\beta^{bpi}(\tau,\delta/2)}{\varepsilon^2}\right)\right]\log_2^2\left(8H\beta^{bpi}(\tau,\delta/2)/\varepsilon^2\right)\right),
\end{aligned}$$

where (a) and (b) use Lemma 4.8. Finally

$$\begin{aligned}
C &\leq \widetilde{\mathcal{O}}\left(\varphi^\star(\mathbb{1})SAH^2\big(\log(4(\kappa_f+1)^2/\delta)+S\big)\kappa_f^2\right)\\
&\leq \widetilde{\mathcal{O}}\left(\varphi^\star(\mathbb{1})SAH^2\Big[\log\Big(\frac{4\log_2^2\left(8H\beta^{bpi}(\tau,\delta/2)/\varepsilon^2\right)}{\delta}\Big)+S\Big]\log_2^2\left(8H\beta^{bpi}(\tau,\delta/2)/\varepsilon^2\right)\right),
\end{aligned}$$

where we have used Lemma 4.8 again. Combining the three inequalities with the definition of β^{bpi} we get that

$$\begin{aligned}
\tau \leq \mathcal{O}\Bigg((H^3\log(1/\delta)+SAH^4)\left[\varphi^\star\left(\left[\sup_{\pi\in\Pi_S}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)+\frac{\varphi^\star(\mathbb{1})}{\varepsilon}+\varphi^\star(\mathbb{1})\right]\\
\times \operatorname{polylog}(\tau,S,A,H,\varphi^\star(\mathbb{1}),\varepsilon,\log(1/\delta))\Bigg).
\end{aligned}$$

Solving for τ yields

$$\tau \leq \widetilde{\mathcal{O}}\left((H^3\log(1/\delta)+SAH^4)\left[\varphi^\star\left(\left[\sup_{\pi\in\Pi_S}\frac{p_h^\pi(s,a)}{\max(\varepsilon,\Delta(\pi))^2}\right]_{h,s,a}\right)+\frac{\varphi^\star(\mathbb{1})}{\varepsilon}+\varphi^\star(\mathbb{1})\right]\right),$$

where $\widetilde{\mathcal{O}}$ hides poly-logarithmic factors in $S, A, H, \varepsilon, \log(1/\delta)$ and $\varphi^\star(\mathbb{1})$. ■

4.5 Conclusion and open question

We proposed the first general instance-dependent lower bound for online ε-BPI and proved that it is nearly matched by PEDEL. This however comes at the cost of enumerating and storing the set of deterministic policies, which is of size A^{SH}. This is needed by PEDEL in order to both eliminate suboptimal policies and solve an experimental design of the form

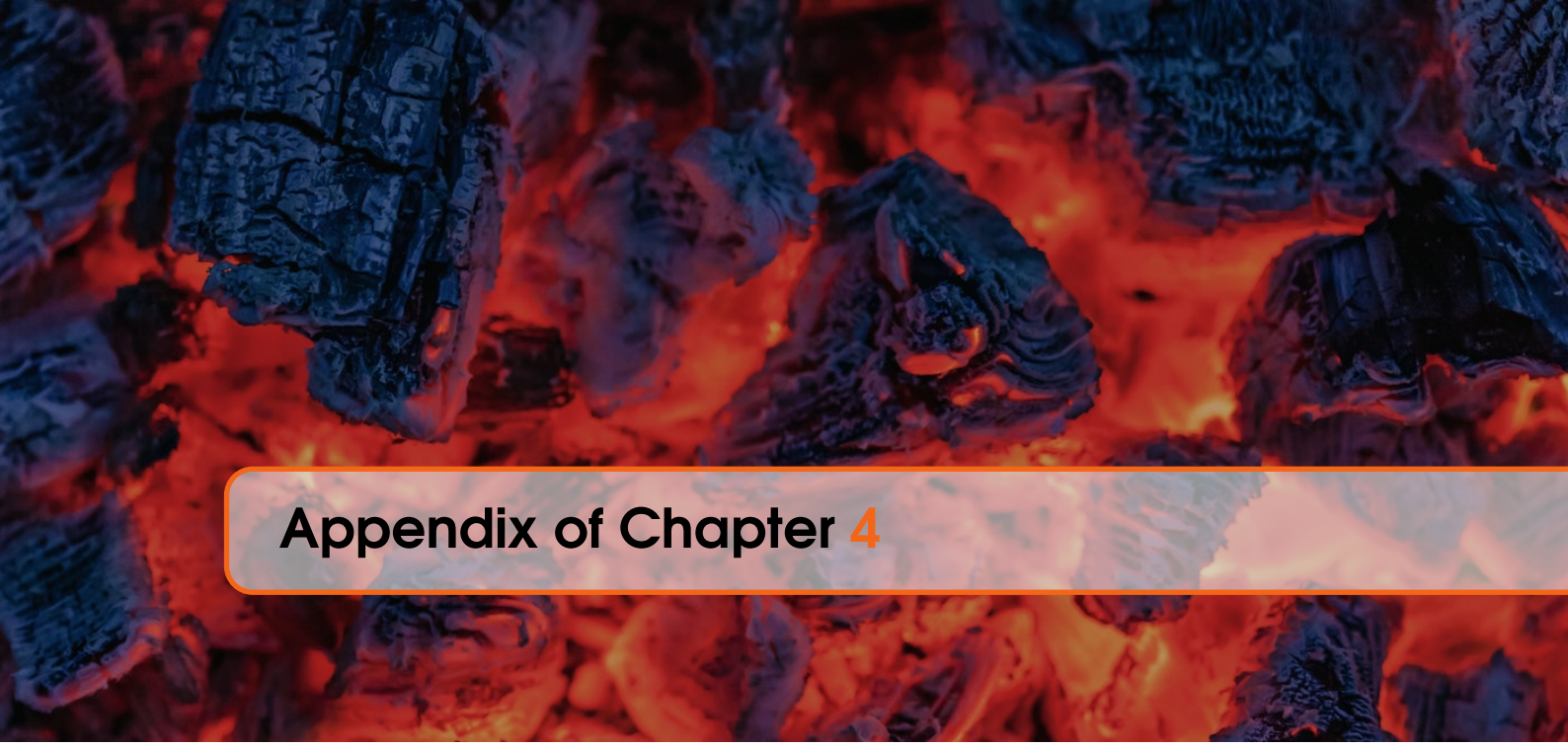
$$\min_{\rho\in\Omega(\mathcal{M})}\max_{\pi\in\Pi_\ell}\sum_{s,a}\frac{\widehat{p}_h^{\pi,\ell}(s,a)^2}{\rho_h(s,a)},$$

where $\Pi_\ell \subset \Pi_D$ is the set of active policies at iteration ℓ (initialized as $\Pi_0 = \Pi_D$) and $\widehat{p}_h^{\pi,\ell}(s,a)$ refers to the visitation probability under the empirical MDP $\widehat{\mathcal{M}}_\ell$. Therefore, we ask the following question

Open question 4.1 Is there an ε-BPI algorithm that can (nearly) match the lower bound of Theorem 4.1 while maintaining a computational and memory complexity that are polynomial in SAH?

We believe that answering this question would shed light on the (still elusive) question of instance-optimality in ε-BPI. Indeed, if the answer is negative then this would indicate a clear separation between MDPs and Bandits where we know that computationally efficient instance-optimality is possible (Garivier & Kaufmann, 2016; Jedra & Proutiere, 2020).

Finally, in an attempt to make policy eliminations tractable, we combined proportional coverage in CovGame with an implicit policy elimination scheme to design an ε-BPI algorithm. Thus we obtained PRINCIPLE, the first computationally efficient algorithm for ε-BPI in stochastic MDPs whose sample complexity scales with policy gaps.



Appendix of Chapter 4

4.6 Proof of Theorem 4.1

As mentioned before, our proof is inspired by the one from (Degenne & Koolen, 2019). The key differences are in Lemma 4.11 which explicits the shape of the characteristic time for the ε-BPI problem and Lemma 4.13 which relies on a slightly different martingale construction to concentrate the likelihood ratio. Indeed, our martingale involves the expected number of visits to state-action pairs instead of the actual number of visits as in (Degenne & Koolen, 2019), which is crucial to obtain the navigation constraints $\rho \in \Omega(\mathcal{M})$ in the optimization program of the lower bound.

Notation: For any $\pi^\varepsilon \in \Pi^\varepsilon$, we define the set of alternative MDPs that have the same transitions as $\mathcal{M}$ but in which π^ε is no longer ε-optimal:

$$\mathrm{Alt}\,(\pi^\varepsilon) := \Big\{ \widetilde{\mathcal{M}} \in \mathfrak{M}_1 :\ \forall (h,s,a),\ p_h(.|s,a;\widetilde{\mathcal{M}}) = p_h(.|s,a;\mathcal{M}) \\ \text{and } \exists \pi \in \Pi^D,\ V_1^{\widetilde{\mathcal{M}},\pi^\varepsilon} < V_1^{\widetilde{\mathcal{M}},\pi} - \varepsilon \Big\}.$$

Finally, we define the characteristic time to learn that π^ε is ε-optimal

$$T(\mathcal{M},\pi^\varepsilon,\varepsilon) := \left(\sup_{\rho\in\Omega(\mathcal{M})} \inf_{\widetilde{\mathcal{M}}\in \mathrm{Alt}(\pi^\varepsilon)} \sum_{h,s,a} \rho_h(s,a) \frac{\big(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\big)^2}{2} \right)^{-1}.$$

Further, for any set of MDPs $E \subset \mathfrak{M}_1$, we let $\overline{E}$ denote the closure of E where the convergence is defined w.r.t the distance $d(\mathcal{M},\mathcal{M}') := \max_{h,s,a} |r_h^{\mathcal{M}}(s,a) - r_h^{\mathcal{M}'}(s,a)|$.

Proof. Let $\xi \in (0,1)$ and define $T := (1-\xi)\min_{\pi^\varepsilon\in\Pi^\varepsilon} T(\mathcal{M},\pi^\varepsilon,\varepsilon)\log(1/\delta)$[3]. Thanks to Markov's inequality we have that

$$\mathbb{E}_{\mathcal{M}}[\tau] \geq T(1 - \mathbb{P}_{\mathcal{M}}(\tau < T)). \tag{4.15}$$

[3]For simplicity, we assume the latter is integer.

We will now upper bound the probability on the right-hand side above. Since the algorithm is (ε, δ)-PAC We have that

$$\begin{aligned}\mathbb{P}_{\mathcal{M}}(\tau < T) &= \mathbb{P}_{\mathcal{M}}\big(\widehat{\pi} \notin \Pi^{\varepsilon}, \tau < T\big) + \sum_{\pi^{\varepsilon} \in \Pi^{\varepsilon}} \mathbb{P}_{\mathcal{M}}\big(\widehat{\pi} = \pi^{\varepsilon}, \tau < T\big) \\ &\leq \delta + \sum_{\pi^{\varepsilon} \in \Pi^{\varepsilon}} \mathbb{P}_{\mathcal{M}}\big(\widehat{\pi} = \pi^{\varepsilon}, \tau < T\big). \end{aligned} \tag{4.16}$$

Now we fix $\pi^{\varepsilon} \in \Pi^{\varepsilon}$ and apply Lemma 4.14 for the event $\mathcal{C} = \big(\widehat{\pi} = \pi^{\varepsilon}, \tau < T\big) \in \mathcal{F}_T$, which yields that there exists $\widetilde{\mathcal{M}}_1, \dots, \widetilde{\mathcal{M}}_{SAH+1} \in \overline{\mathrm{Alt}\,(\pi^{\varepsilon})}$ such that for all $y > 0$

$$\begin{aligned}\mathbb{P}_{\mathcal{M}}\big(\widehat{\pi} = \pi^{\varepsilon}, \tau < T\big) &\leq \exp\Big(y + \frac{T}{T(\mathcal{M}, \pi^{\varepsilon}, \varepsilon)}\Big) \max_{1 \leq i \leq SAH+1} \mathbb{P}_{\widetilde{\mathcal{M}}_i}\big(\widehat{\pi} = \pi^{\varepsilon}, \tau < T\big) \\ &\quad + \sum_{i=1}^{SAH+1} \exp\Big(-\frac{y^2}{2T(\sigma_i \lambda_i^{\star})^2}\Big) \\ &= \delta^{\xi - 1} \exp(y) \max_{1 \leq i \leq SAH+1} \mathbb{P}_{\widetilde{\mathcal{M}}_i}\big(\widehat{\pi} = \pi^{\varepsilon}, \tau < T\big) \\ &\quad + \sum_{i=1}^{SAH+1} \exp\Big(-\frac{y^2}{2T(\sigma_i \lambda_i^{\star})^2}\Big). \end{aligned} \tag{4.17}$$

Now for any $i \in [|1, SAH+1|]$ since $\widetilde{\mathcal{M}}_i \in \overline{\mathrm{Alt}\,(\pi^{\varepsilon})}$ there exists a sequence of MDPs $(\mathcal{M}'_n)_{n \geq 1}$ with values in $\mathrm{Alt}\,(\pi^{\varepsilon})$ such that $\lim_{n \to \infty} \mathcal{M}'_n = \widetilde{\mathcal{M}}_i$[4].

By definition of $\mathrm{Alt}\,(\pi^{\varepsilon})$, we have that $\mathbb{P}_{\mathcal{M}'_n}\big(\widehat{\pi} = \pi^{\varepsilon}, \tau < T\big) \leq \mathbb{P}_{\mathcal{M}'_n}(\widehat{\pi} = \pi^{\varepsilon}) \leq \delta$ for all $n \geq 1$. Therefore

$$\begin{aligned}\mathbb{P}_{\widetilde{\mathcal{M}}_i}\big(\widehat{\pi} = \pi^{\varepsilon}, \tau < T\big) &\leq \mathbb{P}_{\widetilde{\mathcal{M}}_i}(\widehat{\pi} = \pi^{\varepsilon}) \\ &\overset{(a)}{\leq} \liminf_{n \to \infty} \mathbb{P}_{\mathcal{M}'_n}(\widehat{\pi} = \pi^{\varepsilon}) \leq \delta, \end{aligned} \tag{4.18}$$

where (a) uses Fatou's lemma. Combining (4.17) with (4.18) for the value $y = \xi \log(1/\delta)/2$ yields

$$\begin{aligned}\mathbb{P}_{\mathcal{M}}\big(\widehat{\pi} = \pi^{\varepsilon}, \tau < T\big) &\leq \delta^{\xi} \exp(y) + \sum_{i=1}^{SAH+1} \exp\Big(-\frac{y^2}{2T(\sigma_i \lambda_i^{\star})^2}\Big) \\ &\overset{(a)}{=} \delta^{\xi/2} + \sum_{i=1}^{SAH+1} \exp\Big(-\frac{\xi^2 \log(1/\delta)}{4(1-\xi) \min_{\pi^{\varepsilon} \in \Pi^{\varepsilon}} T(\mathcal{M}, \pi^{\varepsilon}, \varepsilon)(\sigma_i \lambda_i^{\star})^2}\Big), \end{aligned} \tag{4.19}$$

where (a) uses the definition of T. Therefore $\lim_{\delta \to 0} \mathbb{P}_{\mathcal{M}}\big(\widehat{\pi} = \pi^{\varepsilon}, \tau < T\big) = 0$. This, combined with (4.16) gives that $\lim_{\delta \to 0} \mathbb{P}_{\mathcal{M}}(\tau < T) = 0$. Plugging this back into (4.15) and using the definition of yields

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_{\mathcal{M}}[\tau]}{\log(1/\delta)} \geq (1 - \xi) \min_{\pi^{\varepsilon} \in \Pi^{\varepsilon}} T(\mathcal{M}, \pi^{\varepsilon}, \varepsilon).$$

To finish the proof of Theorem 4.1, we simply take the limit when ξ goes to zero and use the simplified expression of the characteristic time given by Lemma 4.11. ■

4.6.1 Simplifying the expression of the characteristic time

[4]Recall that the convergence was defined w.r.t the distance $d(\mathcal{M}, \mathcal{M}') := \max_{h,s,a} |r_h^{\mathcal{M}}(s,a) - r_h^{\mathcal{M}'}(s,a)|$

Lemma 4.11 For any $\mathcal{M} \in \mathfrak{M}_1$ and $\pi^\varepsilon \in \Pi^\varepsilon$ we have

$$T(\mathcal{M}, \pi^\varepsilon, \varepsilon) = 2 \inf_{\rho \in \Omega(\mathcal{M})} \max_{\pi \in \Pi^D} \sum_{s,a,h} \frac{\left(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a)\right)^2}{\rho_h(s,a)(\Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon)^2}.$$

Proof. Let us first solve the inner minimization program in the definition of $T(\mathcal{M}, \pi^\varepsilon, \varepsilon)^{-1}$. Using the definition of $\mathrm{Alt}(\pi^\varepsilon)$, we have that

$$\inf_{\widetilde{\mathcal{M}} \in \mathrm{Alt}(\pi^\varepsilon)} \sum_{h,s,a} \rho_h(s,a) \frac{\left(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\right)^2}{2} = \min_{\pi \in \Pi_{\mathrm{D}}} \inf_{\widetilde{\mathcal{M}}: V_1^{\widetilde{\mathcal{M}}, \pi^\varepsilon} < V_1^{\widetilde{\mathcal{M}}, \pi} - \varepsilon} \sum_{h,s,a} \rho_h(s,a) \frac{\left(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\right)^2}{2}. \tag{4.20}$$

Now observe that we can rewrite $V_1^{\widetilde{\mathcal{M}}, \pi^\varepsilon} < V_1^{\widetilde{\mathcal{M}}, \pi} - \varepsilon$ as linear constraint in the rewards of $\widetilde{\mathcal{M}}$:

$$\begin{aligned}
&\sum_{h,s,a} (p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a)) r_h^{\widetilde{\mathcal{M}}}(s,a) > \varepsilon, \\
\iff &\sum_{h,s,a} (p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a)) \left(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\right) > V_1^{\pi^\varepsilon} - V_1^\pi + \varepsilon, \\
\iff &\sum_{h,s,a} (p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a)) \left(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\right) > \Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon
\end{aligned}$$

Therefore, letting $u_h(s,a) = r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)$, the program in (4.20) is equivalent to

$$\min_{\pi \in \Pi_{\mathrm{D}}} \inf_{u \text{ s.t: } \sum_{h,s,a} (p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a)) u_h(s,a) > \Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon} \sum_{h,s,a} \rho_h(s,a) \frac{u_h(s,a)^2}{2}. \tag{4.21}$$

Solving the KKT conditions of the previous program, we get that

$$\inf_{u \text{ s.t: } \sum_{h,s,a} (p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a)) u_h(s,a) > \Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon} \sum_{h,s,a} \rho_h(s,a) \frac{u_h(s,a)^2}{2} = \left(\sum_{h,s,a} \frac{(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon)^2} \right)^{-1}.$$

Summing up all the inequalities, we conclude that

$$T(\mathcal{M}, \pi^\varepsilon, \varepsilon)^{-1} = \frac{1}{2} \sup_{\rho \in \Omega(\mathcal{M})} \min_{\pi \in \Pi_{\mathrm{D}}} \left(\sum_{h,s,a} \frac{(p_h^\pi(s,a) - p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi) - \Delta(\pi^\varepsilon) + \varepsilon)^2} \right)^{-1}.$$

■

4.6.2 A max-min game formulation

We define $\Delta_{SAH+1} := \{\lambda \in \mathbb{R}_+^{SAH+1} : \sum_{i=1}^{SAH+1} \lambda_i = 1\}$ to be the simplex of dimension SAH. Finally $\mathrm{Conv}(E)$ refers to the convex hull of E.

Lemma 4.12 Fix $\pi^\varepsilon \in \Pi^\varepsilon$ and define the set of KL-divergence vectors generated by alternative instances in $\mathrm{Alt}(\pi^\varepsilon)$,

$$\mathcal{D}(\pi^\varepsilon) := \left\{ \left[\frac{\left(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\right)^2}{2} \right]_{h,s,a} \in \mathbb{R}^{SAH} \text{ s.t } \widetilde{\mathcal{M}} \in \mathrm{Alt}(\pi^\varepsilon) \right\}.$$

Then there exists $\rho^\star \in \Omega(\mathcal{M}), \lambda^\star \in \Delta_{SAH+1}$ and $\widetilde{\mathcal{M}}_1, \dots, \widetilde{\mathcal{M}}_{SAH+1} \in \overline{\mathrm{Alt}\,(\pi^\varepsilon)}$ such that

$$T(\mathcal{M}, \pi^\varepsilon, \varepsilon)^{-1} = \sum_{i=1}^{SAH+1} \lambda_i^\star \left[\sum_{h,s,a} \rho_h^\star(s,a) \frac{\left(r_h^{\widetilde{\mathcal{M}}_i}(s,a) - r_h^{\mathcal{M}}(s,a)\right)^2}{2} \right].$$

Furthermore, for any $\rho \in \Omega(\mathcal{M})$ we have that

$$\sum_{i=1}^{SAH+1} \lambda_i^\star \left[\sum_{h,s,a} \rho_h(s,a) \frac{\left(r_h^{\widetilde{\mathcal{M}}_i}(s,a) - r_h^{\mathcal{M}}(s,a)\right)^2}{2} \right] \leq T(\mathcal{M}, \pi^\varepsilon, \varepsilon)^{-1}.$$

Proof. Observe that we can rewrite the expression of the characteristic time $T(\mathcal{M}, \pi^\varepsilon, \varepsilon)$ as follows,

$$\begin{aligned}
T(\mathcal{M}, \pi^\varepsilon, \varepsilon)^{-1} &= \sup_{\rho \in \Omega(\mathcal{M})} \inf_{\widetilde{\mathcal{M}} \in \mathrm{Alt}(\pi^\varepsilon)} \sum_{h,s,a} \rho_h(s,a) \frac{\left(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\right)^2}{2} \\
&= \sup_{\rho \in \Omega(\mathcal{M})} \inf_{\widetilde{d} \in \mathcal{D}(\pi^\varepsilon)} \rho^\top \widetilde{d} \\
&= \sup_{\rho \in \Omega(\mathcal{M})} \inf_{\widetilde{d} \in \overline{\mathcal{D}(\pi^\varepsilon)}} \rho^\top \widetilde{d} \\
&= \sup_{\rho \in \Omega(\mathcal{M})} \inf_{\widetilde{d} \in \mathrm{Conv}(\overline{\mathcal{D}(\pi^\varepsilon)})} \rho^\top \widetilde{d}, \qquad (4.22)
\end{aligned}$$

where $\mathrm{Conv}(\mathcal{D}(\pi^\varepsilon))$ denotes the convex hull of $\mathcal{D}(\pi^\varepsilon)$. Now let $(\rho^\star, d^\star)$ be an optimal solution to (4.22). Since $\mathcal{D}(\pi^\varepsilon) \subset \mathbb{R}^{SAH}$, by Carathéodory's extension theorem we have that there exists $\lambda^\star \in \Delta_{SAH+1}$ and $d_1, \dots, d_{SAH+1} \in \overline{\mathcal{D}(\pi^\varepsilon)}$ such that $d^\star = \sum_{i=1}^{SAH+1} \lambda_i^\star d_i$. This means that there exists $\rho^\star \in \Omega(\mathcal{M})$ and $\widetilde{\mathcal{M}}_1, \dots, \widetilde{\mathcal{M}}_{SAH+1} \in \overline{\mathrm{Alt}\,(\pi^\varepsilon)}$ such that

$$\begin{aligned}
T(\mathcal{M}, \pi^\varepsilon, \varepsilon)^{-1} &= (\rho^\star)^\top d^\star \\
&= \sum_{i=1}^{SAH+1} \lambda_i^\star (\rho^\star)^\top d_i \\
&= \sum_{i=1}^{SAH+1} \lambda_i^\star \left[\sum_{h,s,a} \rho_h^\star(s,a) \frac{\left(r_h^{\widetilde{\mathcal{M}}_i}(s,a) - r_h^{\mathcal{M}}(s,a)\right)^2}{2} \right].
\end{aligned}$$

This proves the first statement. Now for the second statement, using Sion's minimax theorem ((Sion, 1958), Theorem 3.4) we know that

$$(\rho^\star)^\top d^\star = \sup_{\rho \in \Omega(\mathcal{M})} \inf_{\widetilde{d} \in \mathrm{Conv}(\overline{\mathcal{D}(\pi^\varepsilon)})} \rho^\top \widetilde{d} = \inf_{\widetilde{d} \in \mathrm{Conv}(\overline{\mathcal{D}(\pi^\varepsilon)})} \sup_{\rho \in \Omega(\mathcal{M})} \rho^\top \widetilde{d},$$

i.e., $(\rho^\star, d^\star)$ is a saddle point of (4.22). This means that for all $\rho \in \Omega(\mathcal{M})$

$$\rho^\top d^\star \leq (\rho^\star)^\top d^\star = T(\mathcal{M}, \pi^\varepsilon, \varepsilon)^{-1}.$$

Expanding the left-hand side proves the second statement. ■

4.6.3 Log-likelihood ratio of MDPs with same transition kernel

In the following we fix an algorithm $\mathfrak{A}$. For $T \geq 1$ we define the history up to the end of episode T as $\mathcal{H}_T := (s_1^t, a_1^t, R_1^t, \dots, s_H^t, a_H^t, R_H^t, \mathbb{1}(t \leq \tau_\delta))_{1 \leq t \leq T}$. For any MDP $\mathcal{M}$, we write $\mathbb{P}_\mathcal{M}$ to denote the probability distribution over possible histories when $\mathfrak{A}$ interacts

with $\mathcal{M}$[5]. Further $(\mathcal{F}_T)_{T\geq 1}$ will denote the sigma algebra generated by $(\mathcal{H}_T)_{T\geq 1}$. Finally, for a pair of MDPs $\mathcal{M}, \widetilde{\mathcal{M}}$, we define the log-likelihood ratio of observations at the end of any episode T[6]

$$\begin{aligned} L_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) &:= \log \frac{d\mathbb{P}_{\mathcal{M}}}{d\mathbb{P}_{\widetilde{\mathcal{M}}}}(\mathcal{H}_T) \\ &= \log \left(\prod_{t=1}^{T} \prod_{h=1}^{H} \frac{\exp\big(-[R_h^t - r_h^{\mathcal{M}}(s_h^t, a_h^t)]^2/2\big) p_{h-1}^{\mathcal{M}}(s_h^t | s_{h-1}^t, a_{h-1}^t)}{\exp\big(-[R_h^t - r_h^{\widetilde{\mathcal{M}}}(s_h^t, a_h^t)]^2/2\big) p_{h-1}^{\widetilde{\mathcal{M}}}(s_h^t | s_{h-1}^t, a_{h-1}^t)} \right). \end{aligned}$$

Lemma 4.13 For any pair of MDPs $\mathcal{M}, \widetilde{\mathcal{M}} \in \mathfrak{M}_1$, there exists a martingale (under $\mathbb{E}_{\mathcal{M}}$) $\big(M_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}})\big)_{T\geq 1}$ whose increments are $(2d(\mathcal{M}, \mathcal{M}')^2 + d(\mathcal{M}, \mathcal{M}')^4/2)$-subGaussian and such that the likelihood ratio at the end of episode T satisfies

$$L_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) = M_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) + \sum_{h,s,a} \mathbb{E}_{\mathcal{M}}[n_h^T(s,a)] \frac{\big(r_h^{\widetilde{\mathcal{M}}}(s,a) - r_h^{\mathcal{M}}(s,a)\big)^2}{2}.$$

Proof. Using that the MDPs $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ share the same transition kernels and have Gausssian reward distributions with unit variance, we can simplify their log-likelihood ratio as follows,

$$\begin{aligned} L_T(\mathbb{P}_{\mathcal{M}}, \mathbb{P}_{\widetilde{\mathcal{M}}}) &= -\frac{1}{2} \sum_{t=1}^{T} \sum_{h=1}^{H} \left[\big(R_h^t - r_h^{\mathcal{M}}(s_h^t, a_h^t)\big)^2 - \big(R_h^t - r_h^{\widetilde{\mathcal{M}}}(s_h^t, a_h^t)\big)^2 \right] \\ &= \frac{1}{2} \sum_{h,s,a} \sum_{t=1}^{T} \mathbb{1}(s_h^t = s, a_h^t = a) \left[\big(R_h^t - r_h^{\widetilde{\mathcal{M}}}(s,a)\big)^2 - \big(R_h^t - r_h^{\mathcal{M}}(s,a)\big)^2 \right]. \end{aligned} \tag{4.23}$$

Now for any fixed (h,s,a) we can define $\widehat{r}_h^T(s,a) := \frac{\sum_{t=1}^T \mathbb{1}(s_h^t = s, a_h^t = a) R_h^t}{n_h^T(s,a)}$ if $n_h^T(s,a) > 0$ and $\widehat{r}_h^T(s,a) := 0$ otherwise. Then we can write that

$$\begin{aligned} &\sum_{t=1}^{T} \mathbb{1}(s_h^t = s, a_h^t = a) \big(R_h^t - r_h^{\mathcal{M}}(s_h, a_h)\big)^2 \\ &= \sum_{t=1}^{T} \mathbb{1}(s_h^t = s, a_h^t = a) \left[\big(R_h^t - \widehat{r}_h^T(s,a)\big) + \big(\widehat{r}_h^T(s,a) - r_h^{\mathcal{M}}(s,a)\big) \right]^2 \\ &= \sum_{t=1}^{T} \mathbb{1}(s_h^t = s, a_h^t = a) \left[\big(R_h^t - \widehat{r}_h^T(s,a)\big)^2 + \big(\widehat{r}_h^T(s,a) - r_h^{\mathcal{M}}(s,a)\big)^2 \right] \\ &+ 2\big(\widehat{r}_h^T(s,a) - r_h^{\mathcal{M}}(s,a)\big) \underbrace{\sum_{t=1}^{T} \mathbb{1}(s_h^t = s, a_h^t = a) \big(R_h^t - \widehat{r}_h^T(s,a)\big)}_{=0} \\ &= \sum_{t=1}^{T} \mathbb{1}(s_h^t = s, a_h^t = a) \left[\big(R_h^t - \widehat{r}_h^T(s,a)\big)^2 + \big(\widehat{r}_h^T(s,a) - r_h^{\mathcal{M}}(s,a)\big)^2 \right]. \end{aligned} \tag{4.24}$$

[5] Since we will be considering the same algorithm $\mathfrak{A}$ interacting with different MDPs, we do not index the probability distributions by $\mathfrak{A}$.

[6] With the convention that $p_0(.|s_0, a_0) = \mathbb{1}(s_1 = .)$ for all (s_0, a_0). Also note that we have simplified the probabilities of choosing actions $\pi^t(a_h^t | s_h^t, a_{h-1}^t, \ldots, s_1^t, \mathcal{H}_{t-1})$ and of stopping $\pi^t(\tau_\delta = t | \mathcal{H}_t)$ as they only depend on the history, therefore having the same value for $\mathcal{M}$ and $\widetilde{\mathcal{M}}$.

Similarly, one can show that

$$\sum_{h,s,a}\sum_{t=1}^{T}\mathbb{1}(s_h^t=s,a_h^t=a)\big(R_h^t-r_h^{\widetilde{\mathcal{M}}}(s_h^t,a_h^t)\big)^2$$
$$=\sum_{t=1}^{T}\mathbb{1}(s_h^t=s,a_h^t=a)\Big[\big(R_h^t-\widehat{r}_h^T(s,a)\big)^2+\big(\widehat{r}_h^T(s,a)-r_h^{\widetilde{\mathcal{M}}}(s,a)\big)^2\Big]. \tag{4.25}$$

Combining equations (4.23), (4.24) and (4.25) we get that

$$\begin{aligned}L_T(\mathbb{P}_{\mathcal{M}},\mathbb{P}_{\widetilde{\mathcal{M}}})&=\frac{1}{2}\sum_{h,s,a}\sum_{t=1}^{T}\mathbb{1}(s_h^t=s,a_h^t=a)\Big[\big(\widehat{r}_h^T(s,a)-r_h^{\widetilde{\mathcal{M}}}(s,a)\big)^2-\big(\widehat{r}_h^T(s,a)-r_h^{\mathcal{M}}(s,a)\big)^2\Big]\\&=\frac{1}{2}\sum_{h,s,a}n_h^T(s,a)\Big(r_h^{\mathcal{M}}(s,a)-r_h^{\widetilde{\mathcal{M}}}(s,a)\Big)\Big(2\widehat{r}_h^T(s,a)-r_h^{\mathcal{M}}(s,a)-r_h^{\widetilde{\mathcal{M}}}(s,a)\Big).\end{aligned} \tag{4.26}$$

Next we define the sequences

$$\begin{aligned}M_T(h,s,a)&:=\frac{1}{2}\Big[n_h^T(s,a)\big(r_h^{\mathcal{M}}(s,a)-r_h^{\widetilde{\mathcal{M}}}(s,a)\big)\big(2\widehat{r}_h^T(s,a)-r_h^{\mathcal{M}}(s,a)-r_h^{\widetilde{\mathcal{M}}}(s,a)\big)\\&\qquad-\mathbb{E}_{\mathcal{M}}[n_h^T(s,a)]\big(r_h^{\widetilde{\mathcal{M}}}(s,a)-r_h^{\mathcal{M}}(s,a)\big)^2\Big].\\M_T(\mathbb{P}_{\mathcal{M}},\mathbb{P}_{\widetilde{\mathcal{M}}})&:=\sum_{h,s,a}M_T(h,s,a).\end{aligned}$$

Using (4.26) one can check that

$$L_T(\mathbb{P}_{\mathcal{M}},\mathbb{P}_{\widetilde{\mathcal{M}}})=M_T(\mathbb{P}_{\mathcal{M}},\mathbb{P}_{\widetilde{\mathcal{M}}})+\sum_{h,s,a}\mathbb{E}_{\mathcal{M}}[n_h^T(s,a)]\frac{\big(r_h^{\widetilde{\mathcal{M}}}(s,a)-r_h^{\mathcal{M}}(s,a)\big)^2}{2}.$$

This proves the second statement. Now for the first statement we note that for $T\geq 2$

$$\begin{aligned}M_T(h,s,a)-M_{T-1}(h,s,a)&=\frac{1}{2}\Big(r_h^{\mathcal{M}}(s,a)-r_h^{\widetilde{\mathcal{M}}}(s,a)\Big)\mathbb{1}(s_h^T=s,a_h^T=a)\Big(2R_h^T-r_h^{\mathcal{M}}(s,a)-r_h^{\widetilde{\mathcal{M}}}(s,a)\Big)\\&\quad-\frac{1}{2}\mathbb{P}_{\mathcal{M}}(s_h^T=s,a_h^T=a)\Big(r_h^{\widetilde{\mathcal{M}}}(s,a)-r_h^{\mathcal{M}}(s,a)\Big)^2.\end{aligned}$$

Therefore, using that conditionally on the event $(s_h^T=s,a_h^T=a)$ the reward R_h^T is independent of the filtration generated past episodes $\mathcal{F}_{T-1}$, we have that

$$\begin{aligned}&\mathbb{E}_{\mathcal{M}}\Big[M_T(h,s,a)-M_{T-1}(h,s,a)\Big|\mathcal{F}_{T-1}\Big]\\&=\frac{1}{2}\Big(r_h^{\mathcal{M}}(s,a)-r_h^{\widetilde{\mathcal{M}}}(s,a)\Big)\mathbb{P}_{\mathcal{M}}(s_h^T=s,a_h^T=a)\Big(2r_h^{\mathcal{M}}(s,a)-r_h^{\mathcal{M}}(s,a)-r_h^{\widetilde{\mathcal{M}}}(s,a)\Big)\\&-\frac{1}{2}\mathbb{P}_{\mathcal{M}}(s_h^T=s,a_h^T=a)\Big(r_h^{\widetilde{\mathcal{M}}}(s,a)-r_h^{\mathcal{M}}(s,a)\Big)^2\\&=0.\end{aligned}$$

Therefore $\left(M_T(h,s,a)\right)_{T\geq 1}$, and consequently $\left(M_T(\mathbb{P}_{\mathcal{M}},\mathbb{P}_{\widetilde{\mathcal{M}}})\right)_{T\geq 1}$, is a martingale. Furthermore its increments can be decomposed as follows

$$\begin{aligned}
&M_T(\mathbb{P}_{\mathcal{M}},\mathbb{P}_{\widetilde{\mathcal{M}}}) - M_{T-1}(\mathbb{P}_{\mathcal{M}},\mathbb{P}_{\widetilde{\mathcal{M}}}) \\
&= \frac{1}{2}\sum_{h,s,a}\left(r_h^{\mathcal{M}}(s,a) - r_h^{\widetilde{\mathcal{M}}}(s,a)\right)\mathbb{1}(s_h^T = s, a_h^T = a)\left(2R_h^T - r_h^{\mathcal{M}}(s,a) - r_h^{\widetilde{\mathcal{M}}}(s,a)\right) \\
&\quad - \mathbb{P}_{\mathcal{M}}(s_h^T = s, a_h^T = a)\left(r_h^{\mathcal{M}}(s,a) - r_h^{\mathcal{M}}(s,a)\right)^2 \\
&= \underbrace{\left(r_h^{\mathcal{M}}(s,a) - r_h^{\widetilde{\mathcal{M}}}(s,a)\right)\mathbb{1}\left(s_h^T = s, a_h^T = a\right)\left(R_h^T - r_h^{\mathcal{M}}(s,a)\right)}_{:=A_T} \\
&\quad + \underbrace{\frac{1}{2}\left(r_h^{\mathcal{M}}(s,a) - r_h^{\widetilde{\mathcal{M}}}(s,a)\right)^2\left(\mathbb{1}\left(s_h^T = s, a_h^T = a\right) - \mathbb{P}_{\mathcal{M}}\left(s_h^t = s, a_h^t = a\right)\right)}_{:=B_T}.
\end{aligned}$$

Now we prove that each term is sub-Gaussian under $\mathbb{P}_{\mathcal{M}}$. First, we have that

$$|A_T| \leq d(\mathcal{M},\mathcal{M}')\left|\mathbb{1}\left(s_h^T = s, a_h^T = a\right)\left(R_h^T - r_h^{\mathcal{M}}(s,a)\right)\right|.$$

Since the reward ditribution of (h,s,a) under $\mathcal{M}$ is $\mathcal{N}\big(r_h^{\mathcal{M}}(s,a),1\big)$, we conclude that A_T is σ_A^2-sub-Gaussian where $\sigma_A := d(\mathcal{M},\mathcal{M}')$. Similarly, we have that

$$|B_T| \leq \frac{1}{2}d(\mathcal{M},\mathcal{M}')^2\left|\mathbb{1}\left(s_h^T = s, a_h^T = a\right) - \mathbb{P}_{\mathcal{M}}\left(s_h^t = s, a_h^t = a\right)\right| \leq d(\mathcal{M},\mathcal{M}')^2/2.$$

Therefore we conclude that B_T is σ_B^2-sub-Gaussian where $\sigma_B := d(\mathcal{M},\mathcal{M}')^2/2$. Using Lemma 4.15, we conclude that $M_T(\mathbb{P}_{\mathcal{M}},\mathbb{P}_{\widetilde{\mathcal{M}}}) - M_{T-1}(\mathbb{P}_{\mathcal{M}},\mathbb{P}_{\widetilde{\mathcal{M}}})$ is $2(\sigma_A^2+\sigma_B^2)$-subgaussian. ■

4.6.4 The change-of-measure argument

Lemma 4.14 Consider $(\widetilde{\mathcal{M}}_i)_{1\leq i\leq SAH+1} \in \overline{\text{Alt}\,(\pi^\varepsilon)}^{SAH+1}$ given by Lemma 4.12 and let $T \geq 1$. Then for any event $\mathcal{C} \in \mathcal{F}_T$ and any $y \geq 1$ we have

$$\mathbb{P}_{\mathcal{M}}(C) \leq \exp\Big(y + \frac{T}{T(\mathcal{M},\pi^\varepsilon,\varepsilon)}\Big)\max_{1\leq i\leq SAH+1}\mathbb{P}_{\widetilde{\mathcal{M}}_i}(C) + \sum_{i=1}^{SAH+1}\exp\left(-\frac{y^2}{2T(\sigma_i\lambda_i^\star)^2}\right),$$

where $\sigma_i^2 := 2d(\mathcal{M},\widetilde{\mathcal{M}}_i)^2 + d(\mathcal{M},\widetilde{\mathcal{M}}_i)^4/2$.

Proof. Consider the simplex vector $\lambda^\star \in \Delta_{SAH+1}$ given by Lemma 4.12. We define the mixture distribution $\mathbb{Q} = \sum_{i=1}^{SAH+1}\lambda_i^\star\mathbb{P}_{\widetilde{\mathcal{M}}_i}$ and the corresponding log-likelihood ratio

$$L_T(\mathbb{P}_{\mathcal{M}},\mathbb{Q}) := \log\frac{d\mathbb{P}_{\mathcal{M}}}{d\mathbb{Q}}(\mathcal{H}_T).$$

Using Lemma 3.1 from (Garivier & Kaufmann, 2021) we have that for any event $C \in \mathcal{F}_T$ and any $x > 0$,

$$\mathbb{P}_{\mathcal{M}}(C) \leq e^x\mathbb{Q}(C) + \mathbb{P}_{\mathcal{M}}(L_T(\mathbb{P}_{\mathcal{M}},\mathbb{Q}) > x). \tag{4.27}$$

We bound each term in the right-hand side separately. First note that, since $\lambda^\star \in \Delta_{SAH+1}$, for any event C

$$\begin{aligned}\mathbb{Q}(C) &= \sum_{i=1}^{SAH+1} \lambda_i^\star \mathbb{P}_{\widetilde{\mathcal{M}}_i}(C) \\ &\leq \max_{1\leq i\leq SAH+1} \mathbb{P}_{\widetilde{\mathcal{M}}_i}(C) \end{aligned} \tag{4.28}$$

On the other hand, we have that

$$\begin{aligned} L_T(\mathbb{P}_\mathcal{M}, \mathbb{Q}) &\overset{(a)}{\leq} \sum_{i=1}^{SAH+1} \lambda_i^\star \log \frac{d\mathbb{P}_\mathcal{M}}{d\mathbb{P}_{\widetilde{\mathcal{M}}_i}}\Big((s_1^t, a_1^t, R_1^t, \dots, s_H^t, a_H^t, R_H^t)_{1\leq t\leq T}\Big) \\ &= \sum_{i=1}^{SAH+1} \lambda_i^\star L_T(\mathbb{P}_\mathcal{M}, \mathbb{P}_{\widetilde{\mathcal{M}}_i}) \\ &\overset{(b)}{=} \sum_{i=1}^{SAH+1} \lambda_i^\star M_T(\mathbb{P}_\mathcal{M}, \mathbb{P}_{\widetilde{\mathcal{M}}_i}) + \sum_{i=1}^{SAH+1} \lambda_i^\star \sum_{h,s,a} \mathbb{E}_\mathcal{M}[n_h^T(s,a)] \frac{\big(r_h^{\widetilde{\mathcal{M}}_i}(s,a) - r_h^\mathcal{M}(s,a)\big)^2}{2} \\ &= \sum_{i=1}^{SAH+1} \lambda_i^\star M_T(\mathbb{P}_\mathcal{M}, \mathbb{P}_{\widetilde{\mathcal{M}}_i}) + T \sum_{i=1}^{SAH+1} \lambda_i^\star \sum_{h,s,a} \frac{\mathbb{E}_\mathcal{M}[n_h^T(s,a)]}{T} \frac{\big(r_h^{\widetilde{\mathcal{M}}_i}(s,a) - r_h^\mathcal{M}(s,a)\big)^2}{2} \\ &\overset{(c)}{\leq} \sum_{i=1}^{SAH+1} \lambda_i^\star M_T(\mathbb{P}_\mathcal{M}, \mathbb{P}_{\widetilde{\mathcal{M}}_i}) + \frac{T}{T(\mathcal{M}, \pi^\varepsilon, \varepsilon)}, \end{aligned}$$

where (a) uses the convexity of $x \mapsto \log(1/x)$ and Jensen's inequality, (b) uses Lemma 4.13 and (c) uses the second statement of Lemma 4.12 and the fact that the vector $\big[\frac{\mathbb{E}_\mathcal{M}[n_h^T(s,a)]}{T}\big]_{h,s,a}$ belongs to $\Omega(\mathcal{M})$. Therefore for any $y > 0$, we have that

$$\begin{aligned} \mathbb{P}_\mathcal{M}\Big(L_T(\mathbb{P}_\mathcal{M}, \mathbb{Q}) > \frac{T}{T(\mathcal{M}, \pi^\varepsilon, \varepsilon)} + y\Big) &\leq \mathbb{P}_\mathcal{M}\Big(\sum_{i=1}^{SAH+1} \lambda_i^\star M_T(\mathbb{P}_\mathcal{M}, \mathbb{P}_{\widetilde{\mathcal{M}}_i}) > y\Big) \\ &\leq \sum_{i=1}^{SAH+1} \mathbb{P}_\mathcal{M}\Big(M_T(\mathbb{P}_\mathcal{M}, \mathbb{P}_{\widetilde{\mathcal{M}}_i}) > y/\lambda_i^\star\Big) \\ &\leq \sum_{i=1}^{SAH+1} \exp\Big(-\frac{y^2}{2T(\sigma_i\lambda_i^\star)^2}\Big), \end{aligned} \tag{4.29}$$

where in the last line we defined $\sigma_i^2 := 2d(\mathcal{M}, \widetilde{\mathcal{M}}_i)^2 + d(\mathcal{M}, \widetilde{\mathcal{M}}_i)^4/2$ and used Azuma-Hoeffding inequality along with Lemma 4.13. Combining (4.28) and (4.29) with (4.27) for $x = \frac{T}{T(\mathcal{M}, \pi^\varepsilon, \varepsilon)} + y$ gives the result. ■

4.6.5 Sum of subgaussian random variables

Lemma 4.15 Let X an Y be to random variables with values in $\mathbb{R}$ that are σ_X^2 and σ_Y^2 subGaussian respectively. Then $X + Y$ is $2(\sigma_X^2 + \sigma_Y^2)$-subGaussian.

Proof. Using Cauchy-Schwartz's inequality and the definition of sub-Gaussian variables, we

write

$$\begin{aligned}\mathbb{E}[\exp\big(t(X+Y)\big)] &= \mathbb{E}[\exp(tX)\exp(tY)]\\ &\le \mathbb{E}[\exp(2tX)]^{1/2}\mathbb{E}[\exp(2tY)]^{1/2}\\ &\le \exp\big(\frac{4t^2\sigma_X^2}{2}\big)^{1/2}\exp\big(\frac{4t^2\sigma_Y^2}{2}\big)^{1/2}\\ &= \exp\big(\frac{t^2(2\sigma_X^2+2\sigma_Y^2)}{2}\big).\end{aligned}$$

■

4.7 Proof of Lemma 4.1

Proof. Fix any pair of policies π, π'. We write

$$\begin{aligned}(\widehat{V}_1^{\pi,t} - \widehat{V}_1^{\pi',t}) - (V_1^{\pi} - V_1^{\pi'}) &= (p^{\pi} - p^{\pi'})^{\top}(\widehat{r}^t - r)\\ &= \sum_{h,s,a}(p_h^{\pi}(s,a) - p_h^{\pi'}(s,a))(\widehat{r}_h^t(s,a) - r_h(s,a))\\ &= \sum_{h,s,a}\mathbb{1}\big(a \in \{\pi_h(s), \pi'_h(s)\}\big)(p_h^{\pi}(s,a) - p_h^{\pi'}(s,a))(\widehat{r}_h^t(s,a) - r_h(s,a)),\end{aligned}$$

where we used vector notation $p^{\pi} = [p_h^{\pi}(s,a)]_{h,s,a}$. Now applying Lemma 3.15 with $\delta' = \delta/(A^{2SH})$ and $\mathcal{Z} = \big\{(h,s,a) \,:\, (h,s) \in [H]\times\mathcal{S}, a \in \{\pi_h(s), \pi'_h(s)\}\big\}$ we get that with proability at least $1-\delta/(A^{2SH})$, for all $t \ge t_0$,

$$\begin{aligned}\sum_{h,s,a}\mathbb{1}\big(a \in \{\pi_h(s), \pi'_h(s)\}\big)n_h^t(s,a)\big(\widehat{r}_h^t(s,a) - r_h(s,a)\big)^2 &\le 4\log(1/\delta) + 4SH\log(A(1+t))\\ &= \beta^r(t,\delta),\end{aligned}$$

where we used that $|\mathcal{Z}| \le 2SH$. Next we use Lemma 3.17 with $p = p^{\pi} - p^{\pi'}$ which yields that

$$\begin{aligned}\big|(\widehat{V}_1^{\pi,t} - \widehat{V}_1^{\pi',t}) - (V_1^{\pi} - V_1^{\pi'})\big| &\le \sqrt{\beta^r(t,\delta)\sum_{h,s,a}\mathbb{1}\big(a \in \{\pi_h(s), \pi'_h(s)\}\big)\frac{\big(p_h^{\pi}(s,a) - p_h^{\pi'}(s,a)\big)^2}{n_h^t(s,a)}}\\ &\le \sqrt{\beta^r(t,\delta)\sum_{h,s,a}\frac{\big(p_h^{\pi}(s,a) - p_h^{\pi'}(s,a)\big)^2}{n_h^t(s,a)}},\end{aligned}$$

with probability at least $1-\delta/(A^{2SH})$. We conclude the proof with a union bound over pairs of policies $(\pi,\pi') \in \Pi_{\mathrm{D}} \times \Pi_{\mathrm{D}}$. ■

4.8 PEDEL

4.8.1 Proof of Proposition 4.1

First, let us introduce the intermediate complexity measure

$$C(\mathcal{M},\varepsilon) := \min_{\rho\in\Omega}\max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a,h}\frac{p_h^{\pi}(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}.$$

We start by showing that $H^3 C(\mathcal{M},\varepsilon) \le \mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M},\varepsilon) \le H^5 C(\mathcal{M},\varepsilon)$. For $h \in [H]$ consider any $\rho^{\star,h} \in \arg\min_{\rho\in\Omega}\max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a}\frac{p_h^{\pi}(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}$. Now, letting $\widetilde{\rho} :=$

$\frac{1}{H}\sum_{h=1}^{H}\rho^{\star,h}$, we see that since Ω is a convex set, $\widetilde{\rho}\in\Omega$. Furthermore,

$$\begin{aligned}
C(\mathcal{M},\varepsilon) &= \min_{\rho\in\Omega}\max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a,h}\frac{p_h^\pi(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}\\
&\le \max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a,h}\frac{p_h^\pi(s,a)^2}{\widetilde{\rho}_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}\\
&\overset{(a)}{\le} \sum_{h=1}^{H}\max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a}\frac{p_h^\pi(s,a)^2}{\widetilde{\rho}_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}\\
&\overset{(b)}{\le} H\sum_{h=1}^{H}\max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a}\frac{p_h^\pi(s,a)^2}{\rho_h^{\star,h}(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}\\
&= H\sum_{h=1}^{H}\min_{\rho\in\Omega}\max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a}\frac{p_h^\pi(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}\\
&= H^{-3}\mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M},\varepsilon),
\end{aligned}$$

where (a) uses the fact that $max_\pi\sum_h f(\pi,h)\le\sum_h max_\pi f(\pi,h)$ and (b) uses the crude bound $\widetilde{\rho}_h(s,a)\ge\rho_h^{\star,h}(s,a)/H$. On the other hand we have

$$\begin{aligned}
\mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M},\varepsilon) &= H^4\sum_{h=1}^{H}\min_{\rho\in\Omega}\max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a}\frac{p_h^\pi(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}\\
&\le H^4\sum_{h=1}^{H}\min_{\rho\in\Omega}\max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a,h}\frac{p_h^\pi(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}\\
&= H^5 C(\mathcal{M},\varepsilon).
\end{aligned}$$

Therefore, we just proved that

$$H^3 C(\mathcal{M},\varepsilon)\le\mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M},\varepsilon)\le H^5 C(\mathcal{M},\varepsilon). \tag{4.30}$$

Now we compare $C(\mathcal{M},\varepsilon)$ and $LB(\mathcal{M},\varepsilon)$. Using that $a^2\le 2(a-b)^2+2b^2$, we note that for any $\rho\in\Omega$ and any $\pi^\varepsilon\in\Pi^\varepsilon$,

$$\begin{aligned}
\max_{\pi\in\Pi_{\mathrm{D}}}\frac{\sum_{s,a,h}\frac{p_h^\pi(s,a)^2}{\rho_h(s,a)}}{(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2} &\le \max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a,h}\left[\frac{2(p_h^\pi(s,a)-p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}+\frac{2p_h^{\pi^\varepsilon}(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}\right]\\
&\le \max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a,h}\frac{2(p_h^\pi(s,a)-p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}+\max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a,h}\frac{2p_h^{\pi^\varepsilon}(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}\\
&= \max_{\pi\in\Pi_{\mathrm{D}}}\frac{2(p_h^\pi(s,a)-p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}+\sum_{s,a,h}\frac{2p_h^{\pi^\varepsilon}(s,a)^2}{\rho_h(s,a)(\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}.
\end{aligned} \tag{4.31}$$

Now let us define $\rho^0 := \arg\min_{\rho\in\Omega}\max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a,h}\frac{(p_h^\pi(s,a)-p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}$ and $\widetilde{\rho}^1 := \frac{\rho^0+p^{\pi^\varepsilon}}{2}$. Then we have that

$$
\begin{aligned}
C(\mathcal{M},\varepsilon) &= \min_{\rho\in\Omega}\max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a,h}\frac{p_h^\pi(s,a)^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}\\
&\le \max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a,h}\frac{p_h^\pi(s,a)^2}{\widetilde{\rho}_h^1(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}\\
&\overset{(a)}{\le} \max_{\pi\in\Pi_{\mathrm{D}}}\frac{2(p_h^\pi(s,a)-p_h^{\pi^\varepsilon}(s,a))^2}{\widetilde{\rho}_h^1(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}+\sum_{s,a,h}\frac{2p_h^{\pi^\varepsilon}(s,a)^2}{\widetilde{\rho}_h^1(s,a)(\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}\\
&\overset{(b)}{\le} \max_{\pi\in\Pi_{\mathrm{D}}}\frac{4(p_h^\pi(s,a)-p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h^0(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}+\sum_{s,a,h}\frac{4p_h^{\pi^\varepsilon}(s,a)^2}{p_h^{\pi^\varepsilon}(s,a)(\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}\\
&= 4\min_{\rho\in\Omega}\max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a,h}\frac{(p_h^\pi(s,a)-p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}+\frac{4H}{(\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2},
\end{aligned}
$$

where (a) uses (4.31) and (b) uses that for all (h,s,a), $\widetilde{\rho}_h^1(s,a)\ge\max(\rho_h^0(s,a),p_h^{\pi^\varepsilon}(s,a))/2$. Now taking the minimum over $\pi^\varepsilon\in\Pi^\varepsilon$ in both sides of the previous inequality proves that

$$
\begin{aligned}
C(\mathcal{M},\varepsilon) &\le 4\min_{\pi\in\Pi^\varepsilon}\min_{\rho\in\Omega}\max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a,h}\frac{(p_h^\pi(s,a)-p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}+\frac{4H}{(\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}\\
&\le 16\min_{\pi\in\Pi^\varepsilon}\min_{\rho\in\Omega}\max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a,h}\frac{(p_h^\pi(s,a)-p_h^{\pi^\varepsilon}(s,a))^2}{\rho_h(s,a)(\Delta(\pi)+\varepsilon-\Delta(\pi^\varepsilon))^2}+\frac{4H}{(\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}\\
&\le 8LB(\mathcal{M},\varepsilon)+\frac{4H}{(\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2},
\end{aligned}
\tag{4.32}
$$

where in the second inequality we used that $\Delta(\pi)+\varepsilon-\Delta(\pi^\varepsilon)\le 2(\Delta(\pi)\vee\varepsilon\vee\Delta_{\min}(\Pi_{\mathrm{D}}))$. Combining (4.30) and (4.32) proves the first inequality.

4.8.2 On the complexity of PEDEL in the moderate precision regime

PEDEL has a loop structure where at each iteration it seeks to halve the precision of its estimate of the value for all the policies that are still active. Taking a closer look into the design of PEDEL, we notice that it starts the first iteration with the parameter $\ell_0=\lceil\log_2\frac{d^{3/2}}{H}\rceil$ and ends at $\lceil\log\frac{4}{\varepsilon}\rceil$. From Theorem 7 in (Wagenmaker & Jamieson, 2022), we get that the number of episodes played during the initial iteration is

$$
\mathcal{O}\left(H^4\sum_{h=1}^{H}\frac{\inf_{\Lambda_{exp}\in\Omega_h}\max_{\varphi\in\Phi}\|\varphi\|_{\Lambda_{exp}^{-1}}}{\varepsilon_{exp}}\right),\ \text{where}\quad \varepsilon_{exp}:=\frac{\varepsilon_{\ell_0}^2}{\beta_{\ell_0}},
$$

$$
\varepsilon_{\ell_0}:=2^{-\ell_0}=\frac{H}{d^{3/2}},\ \beta_{\ell_0}:=64H^2\log(\frac{4H^2|\Pi|\ell_0^2}{\delta}).
$$

As a consequence, running just the initial iteration of PEDEL requires the number of episodes

$$
\mathcal{C}_0:=\mathcal{O}\left(d^3H^4\log(|\Pi|/\delta)\min_{\rho\in\Omega}\max_{\pi\in\Pi_{\mathrm{D}}}\sum_{s,a,h}\frac{p_h^\pi(s,a)^2}{\rho_h(s,a)}\right).
$$

When $\varepsilon=\Omega(1/d)$, we have that $d^2=\Omega(\frac{1}{(\varepsilon\vee\Delta(\pi)\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2})$ for all policies π so that

$$
\mathcal{C}_0=\Omega\left(dH^2\log(|\Pi|/\delta)\min_{\rho\in\Omega(\mathcal{M})}\max_{\pi\in\Pi_{\mathrm{D}}}\frac{\sum_{s,a,h}\frac{p_h^\pi(s,a)^2}{\rho_h(s,a)}}{(\varepsilon\vee\Delta(\pi)\vee\Delta_{\min}(\Pi_{\mathrm{D}}))^2}\right).
$$

Therefore when $\varepsilon = \Omega(1/SAH)$, we get that the sample complexity of PEDEL for tabular MDPs satisfies

$$\tau = \Omega\left(SAH \times \mathcal{C}_{\mathrm{PEDEL}}(\mathcal{M}, \varepsilon) \log\left(1/\delta\right)\right),$$

with high probabilty.

5. All-epsilon Best Arms Identification

In this Chapter, we investigate the All-ε-BAI problem (see Section 1.4.3). Firt, we derive an instance-dependent lower bound using the KL contraction method. Then, we present an efficient method to solve the the max-min program featured in this lower bound. This leads us to the design of a Track-and-Stop algorithm, whose sample complexity matches the lower bound when δ tends to 0. Finally, we provide an example of a bandit instance where the simulator lower bound of Theorem 1.2 can be tighter than the KL contraction bound. The contents of this chapter are based on the conference paper:

Aymen Al Marjani, Tomáš Kocák, Aurélien Garivier. **On the Complexity of All ε-Best Arms Identification**. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (ECMLPKDD), 2022.

Contents

5.1 Lower bound

Notation: We consider Gaussian bandits of the shape $\nu = \big(\mathcal{N}(\mu_a, 1)\big)_{a\in[K]}$ which we parametrize by their mean-reward vector $\boldsymbol{\mu} \in \mathbb{R}^K$. $G_\varepsilon(\mu) := \{a \in [K] : \ \mu_a \geq \max_b \mu_b - \varepsilon\}$ will denote the set of "good arms". We use $\Sigma_K := \{\omega \in \mathbb{R}_+^K : \ \sum_{a\in[K]} \omega_a = 1\}$ for the simplex of dimension $K-1$. The set of *alternative* bandit instances is defined as $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} \in \mathbb{R}^K : G_\varepsilon(\mu) \neq G_\varepsilon(\boldsymbol{\lambda})\}$. By following the same steps in the proof of (1.27), we derive the lower bound below.

Proposition 5.1 For any δ-correct algorithm $\mathbb{A}$ and any bandit instance $\boldsymbol{\mu}$, the expected stopping time τ_δ can be lower-bounded as

$$\mathbb{E}_{\mu,\mathbb{A}}[\tau_\delta] \geq T^\star_\varepsilon(\boldsymbol{\mu}) \log(1/2.4\delta)$$

where

$$T^\star_\varepsilon(\boldsymbol{\mu})^{-1} := \sup_{\omega\in\Sigma_K} T_\varepsilon(\boldsymbol{\mu},\omega)^{-1} \qquad \text{and} \tag{5.1}$$

$$T_\varepsilon(\boldsymbol{\mu},\omega)^{-1} := \inf_{\lambda\in\mathrm{Alt}(\boldsymbol{\mu})} \sum_{a\in[K]} \omega_a \frac{(\mu_a-\lambda_a)^2}{2}\,. \tag{5.2}$$

The characteristic time $T^\star_\varepsilon(\boldsymbol{\mu})$ above is an instance-specific quantity that determines the difficulty of our problem. The optimization program in the definition of $T^\star_\varepsilon(\boldsymbol{\mu})$ can be seen as a two-player game between an algorithm which samples each arm a proportionally to ω_a and an adversary who chooses an alternative instance λ that is difficult to distinguish from μ under the algorithm's sampling scheme. This suggests that an optimal strategy should play the optimal allocation ω^* that maximizes the optimization problem (5.1) and, as a consequence, rules out all alternative instances as fast as possible. This motivates our algorithm, presented in Section 5.2.

Proof. Let $\mathrm{kl}(p,q)$ be the KL-divergence between two Bernoulli distributions with parameters p and q. We start by applying Lemma 1 from (Kaufmann et al., 2016) which states that for any $\mathcal{F}_\tau$-measurable event $\mathcal{E}$, and any pair of bandit problems μ and λ,

$$\sum_{a\in[K]} \frac{(\mu_a-\lambda_a)^2}{2}\mathbb{E}_{\mu,\mathbb{A}}[N_a(\tau_\delta)] \geq \mathrm{kl}(\mathbb{P}_{\mu,\mathbb{A}}(\mathcal{E}),\ \mathbb{P}_{\lambda,\mathbb{A}}(\mathcal{E}))$$

We let $\mathcal{E} := (\widehat{G} \neq G_\varepsilon(\mu))$, where $\widehat{G}$ is the set answered by $\mathbb{A}$ at the end of exploration. For this choice of event and since $\mathbb{A}$ is δ-correct, we have $\mathbb{P}_{\mu,\mathbb{A}}(\mathcal{E}) \leq \delta$. On the other hand, by choosing $\lambda \in \mathrm{Alt}(\boldsymbol{\mu})$, we get $\mathbb{P}_{\lambda,\mathbb{A}}(\mathcal{E}) \geq 1-\delta$. Therefore, using the monotonicity properties of $(p,q) \mapsto \mathrm{kl}(p,q)$ we have $\mathrm{kl}(\mathbb{P}_{\mu,\mathbb{A}}(\mathcal{E}),\ \mathbb{P}_{\lambda,\mathbb{A}}(\mathcal{E})) \geq \mathrm{kl}(\delta,\,1-\delta)$. Since this holds for any alternative problem $\lambda \in \mathrm{Alt}(\boldsymbol{\mu})$, we get

$$\begin{aligned}
\mathrm{kl}(\delta,\,1-\delta) &\leq \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu})} \mathrm{kl}(\mathbb{P}_{\mu,\mathbb{A}}(\mathcal{E}),\ \mathbb{P}_{\lambda,\mathbb{A}}(\mathcal{E})) \\
&\leq \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu})} \sum_{a\in[K]} \mathbb{E}_{\mu,\mathbb{A}}[\tau_\delta]\frac{(\mu_a-\lambda_a)^2}{2}\frac{\mathbb{E}_{\mu,\mathbb{A}}[N_a(\tau_\delta)]}{\mathbb{E}_{\mu,\mathbb{A}}[\tau_\delta]} \\
&\leq \mathbb{E}_{\mu,\mathbb{A}}[\tau_\delta] \sup_{\boldsymbol{\omega}\in\Sigma_K} \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu})} \sum_{a\in[K]} \omega_a \frac{(\mu_a-\lambda_a)^2}{2},
\end{aligned}$$

where we used that the vector Fraction $(\mathbb{E}_{\mu,\mathbb{A}}[N_a]/\mathbb{E}_{\mu,\mathbb{A}}[\tau_\delta])_{a\in[K]}$ is in the simplex. We conclude by noting that $\mathrm{kl}(\delta,\,1-\delta) \geq \log(1/2.4\delta)$. ∎

5.2 Track-and-Stop for All-ε-BAI

We propose an adaptation of the Track-and-Stop strategy similar to the one proposed by (Garivier & Kaufmann, 2016) for the problem of Best-Arm Identification. It starts by sampling once from every arm $a \in [K]$ and constructs an initial estimate $\widehat{\boldsymbol{\mu}}_K$ of the vector of mean rewards $\boldsymbol{\mu}$. After this burn-in phase, the algorithm enters a loop where at every iteration it plays arms according to the estimated optimal sampling rule (5.3) and updates its estimate $\widehat{\boldsymbol{\mu}}_t$ of the arms' expectations. Finally, the algorithm checks if the stopping rule (5.4) is satisfied, in which case it stops and returns the set of empirically ε-good arms. The full pseudo-code is provided in Algorithm 14.

5.2.1 Sampling rule

For our sampling rule we rely on C-tracking: first, we compute $\widetilde{\boldsymbol{\omega}}(\widehat{\boldsymbol{\mu}}_t)$, an allocation vector which is $\frac{1}{\sqrt{t}}$-optimal in the lower-bounf problem (5.1) for the instance $\widehat{\boldsymbol{\mu}}_t$. Then we project $\widetilde{\boldsymbol{\omega}}(\widehat{\boldsymbol{\mu}}_t)$ on the set $\Delta_K^{\eta_t} = \Delta_K \cap [\eta_t, 1]^K$. Given the projected vector $\widetilde{\boldsymbol{\omega}}^{\eta_t}(\widehat{\boldsymbol{\mu}}_t)$, the next arm to sample from is defined by:

$$a_{t+1} = \arg\min_a N_a(t) - \sum_{s=1}^{t} \widetilde{\boldsymbol{\omega}}_a^{\eta_t}(\widehat{\boldsymbol{\mu}}_s), \tag{5.3}$$

where $N_a(t)$ is the number of times arm a has been pulled up to time t. In other words, we sample the arm whose number of visits is farther behind its corresponding sum of empirical optimal allocations. In the long run, as our estimate $\widehat{\boldsymbol{\mu}}_t$ tends to the true value $\boldsymbol{\mu}$, the sampling frequency $N_a(t)/t$ of every arm a will converge to the oracle optimal allocation $\boldsymbol{\omega}_a^*(\boldsymbol{\mu})$. The projection on $\Delta_K^{\eta_t}$ ensures exploration at a minimal rate of $\eta_t = \frac{1}{2\sqrt{(K^2+t)}}$ so that no arm is left behind because of bad initial estimates.

5.2.2 Stopping rule

To be sample-efficient, the algorithm must stop as soon as the collected samples are sufficiently informative to declare that $G_\varepsilon(\widehat{\boldsymbol{\mu}}_t) = G_\varepsilon(\boldsymbol{\mu})$ with probability larger than $1-\delta$. For this purpose we use the Generalized Likelihood Ratio (GLR) test (Chernoff, 1959). We define the Z-statistic

$$Z(t) := t \times T_\varepsilon\big(\widehat{\boldsymbol{\mu}}_t, \frac{\mathbf{N}(t)}{t}\big)^{-1}$$

where $\mathbf{N}(t) = \big(N_a(t)\big)_{a\in[K]}$. As shown in (Degenne et al., 2019a; Garivier & Kaufmann, 2021), the Z-statistic is equal to the ratio of the likelihood of observations under the most likely model where $G_\varepsilon(\widehat{\boldsymbol{\mu}}_t)$ is the correct answer, i.e. $\widehat{\boldsymbol{\mu}}_t$, to the likelihood of observations under the most likely model where $G_\varepsilon(\widehat{\boldsymbol{\mu}}_t)$ is not the set of ε-good arms. The algorithm rejects the hypothesis $G_\varepsilon(\widehat{\boldsymbol{\mu}}_t) \neq G_\varepsilon(\boldsymbol{\mu})$ and stops as soon as this ratio of likelihoods becomes larger than a certain threshold $\beta(\delta, t)$, properly tuned to ensure that the algorithm is δ-PAC. Following this intuition, we define the stopping rule as

$$\tau_\delta := \inf\big\{t \in \mathbb{N} \;:\; Z(t) > \beta(t,\delta)\big\} \tag{5.4}$$

One can find many suitable thresholds from the bandit literature (Garivier, 2013; Magureanu et al., 2014; Kaufmann & Koolen, 2021), all of which are of the order $\beta(\delta,t) \approx \log(1/\delta) + \frac{K}{2}\log(\log(t/\delta))$. Such $\beta(t,\delta)$ is enough to ensure that $\mathbb{P}_{\mu,\mathbb{A}}\big(G_\varepsilon(\widehat{\boldsymbol{\mu}}_{\tau_\delta}) \neq G_\varepsilon(\boldsymbol{\mu})\big) \leq \delta$, i.e. that the algorithm is δ-correct.

Now we state our sample complexity result which we adapted from Theorem 14 in (Garivier & Kaufmann, 2016). Notably, while their Track-and-Stop strategy relies on tracking the exact optimal weights to prove that the expected stopping time matches the

Algorithm 14 Track-and-Stop

Input: risk δ, accuracy parameter ε.
Pull each arm once and observe rewards $(r_a)_{a\in[K]}$.
Set initial estimate $\widehat{\boldsymbol{\mu}}_K = (r_1, \dots, r_K)^T$.
Set $t \leftarrow K$ and $N_a(t) \leftarrow 1$ for all arms a.
for $t = 1, 2, \dots$ **do:**
  Compute $\tilde{\boldsymbol{\omega}}(\widehat{\boldsymbol{\mu}}_t)$, a $\frac{1}{\sqrt{t}}$-optimal vector for (5.1) using mirror-ascent.
  Pull next arm a_{t+1} given by (5.3) and observe reward r_t.
  Update $\widehat{\boldsymbol{\mu}}_t$ according to r_t.
  Set $t \leftarrow t+1$ and update $\big(N_a(t)\big)_{a\in[K]}$.
  if $t \times T_\varepsilon\big(\widehat{\boldsymbol{\mu}}_t, \frac{\mathbf{N}(t)}{t}\big)^{-1} > \beta(t,\delta)$ **then:**
    Stop and return $G_\varepsilon(\widehat{\boldsymbol{\mu}}_{\tau_\delta})$
  end if
end for

lower bound when δ tends to zero, our proof shows that it is enough to track some slightly sub-optimal weights with a decreasing optimality gap in the order of $\frac{1}{\sqrt{t}}$ to enjoy the same sample complexity guarantees. The proof is deferred to Section 5.6.

Theorem 5.1 For all $\delta \in (0,1)$, Track-and-Stop terminates almost-surely. Moreover, its stopping time τ_δ satisfies

$$\limsup_{\delta\to 0} \frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq T^\star_\varepsilon(\boldsymbol{\mu}).$$

Remark 5.1 Suppose that the arms are ordered decreasingly $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$. (Mason et al., 2020) define the upper margin $\alpha_\varepsilon = \min_{k\in G_\varepsilon(\boldsymbol{\mu})} \mu_k - (\mu_1 - \varepsilon)$ and provide a lower bound of the form $f(\nu)\log(1/\delta)$ where

$$f(\nu) := 2\sum_{a=1}^{K} \max\left(\frac{1}{(\mu_1 - \varepsilon - \mu_i)^2}, \frac{1}{(\mu_1 + \alpha_\varepsilon - \mu_a)^2}\right).$$

It can be seen directly (or deduced from Theorem 5.1) that $f(\nu) \leq T^\star_\varepsilon(\boldsymbol{\mu})$. In a second step, they proposed FAREAST, an algorithm whose sample complexity in the asymptotic regime $\delta \to 0$ matches their bound up to some universal constant c that does not depend on the instance ν. From Proposition 5.1, we deduce that $T^\star_\varepsilon(\boldsymbol{\mu}) \leq c f(\nu)$, which can be seen directly from the particular changes of measure considered in that paper. The sample complexity of our algorithm improves upon previous work by multiplicative constants. ■

Note that Algorithm 14 requires to solve the best response problem, i.e. the minimization problem in (5.2), in order to be able to compute the Z-statistic of the stopping rule, and also to solve the entire lower bound problem in (5.1) to compute the optimal weights for the sampling rule. The rest of this chapter is dedicated to presenting the tools necessary to solve these two problems.

5.3 Solving the Min Problem: Best Response Oracle

For a given vector $\boldsymbol{\omega}$, we want to compute the best response

$$\boldsymbol{\lambda}^*_{\varepsilon,\boldsymbol{\mu}}(\boldsymbol{\omega}) := \operatorname*{arg\,min}_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a\in[K]} \omega_a \frac{(\mu_a - \lambda_a)^2}{2}. \tag{5.5}$$

To simplify the presentation, we assume that the arms are ordered decreasingly $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$. We also denote by $B_\varepsilon(\boldsymbol{\mu}) := [K] \setminus G_\varepsilon(\boldsymbol{\mu})$ the set of bad arms.

Since an alternative problem $\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})$ must have a different set of ε-optimal arms than the original problem $\boldsymbol{\mu}$, we can obtain it from $\boldsymbol{\mu}$ by changing the expected reward of some arms. We have two options to create an alternative problem $\boldsymbol{\lambda}$:

- **Making one of the ε-optimal arms bad**. We can achieve it by decreasing the expectation of some ε-optimal arm k while increasing the expectation of some other arm ℓ to the point where k is no more ε-optimal. This is illustrated in Figure 5.1.
- **Making one of the ε-sub-optimal arms good.** We can achieve it by increasing the expectation of some sub-optimal arm k while decreasing the expectations of the arms with the largest means -as many as it takes- to the point where k becomes ε-optimal. This is illustrated in Figure 5.1.

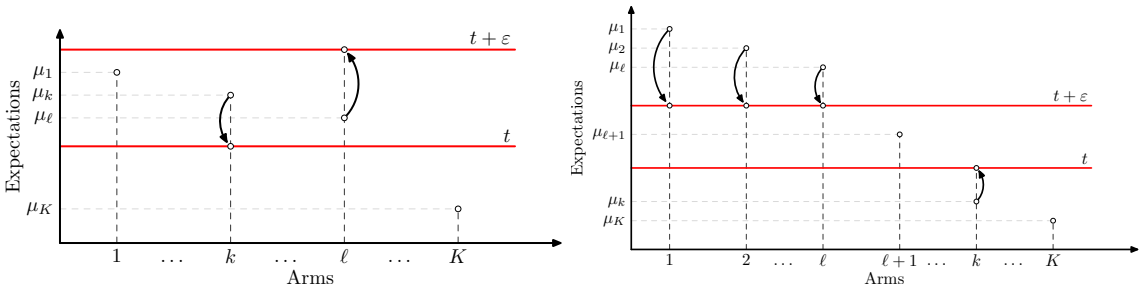


Figure 5.1: Left: Making One of the ε-Optimal Arms Bad. Right: Making One of the ε-Sub-Optimal Arms Good.

In the following, we solve both cases separately.

Case 1: Making one of the ε-optimal arms bad. Let $k \in G_\varepsilon(\boldsymbol{\mu})$ be one of the ε-optimal arms. In order to make arm k sub-optimal, we need to set the expectation of arm k to some value $\lambda_k = t$ and the maximum expectation over all arms to $\max_a \lambda_a = t + \varepsilon$. Note that the index of the arm ℓ with maximum expectation can be chosen in $G_\varepsilon(\boldsymbol{\mu})$. Indeed, if we choose some arm from $B_\varepsilon(\mu)$ to become the arm with maximum expectation in λ then we would make an ε-suboptimal arm good which is covered in the other case below. The expectations of all the other arms should stay the same as in the instance $\boldsymbol{\mu}$, since changing their values would only increase the value of the objective. Now given indices k and ℓ, computing the optimal value of t is rather straightforward since the objective function simplifies to

$$\omega_k \frac{(\mu_k - t)^2}{2} + \omega_\ell \frac{(\mu_\ell - t - \varepsilon)^2}{2}$$

for which the optimal value of t is:

$$t = \overline{\boldsymbol{\mu}}^{k,\ell}_\varepsilon(\boldsymbol{\omega}) := \frac{\omega_k \mu_k + \omega_\ell(\mu_\ell - \varepsilon)}{\omega_k + \omega_\ell}.$$

and the corresponding alternative bandit is:

$$\boldsymbol{\lambda}^{k,\ell}_\varepsilon(\boldsymbol{\omega}) := (\mu_1, \ldots, \underbrace{\overline{\boldsymbol{\mu}}^{k,\ell}_\varepsilon(\boldsymbol{\omega})}_{\text{index } k}, \ldots, \underbrace{\overline{\boldsymbol{\mu}}^{k,\ell}_\varepsilon(\boldsymbol{\omega}) + \varepsilon}_{\text{index } \ell}, \ldots, \mu_K)^\mathsf{T}.$$

The last step is taking the pair of indices $(k, \ell) \in G_\varepsilon(\boldsymbol{\mu}) \times (G_\varepsilon(\boldsymbol{\mu}) \setminus \{k\})$ with the minimal value in the objective (5.2).

Case 2: Making one of the sub-optimal arms good. Let $k \in B_\varepsilon(\boldsymbol{\mu})$ be a sub-optimal arm, if such arm exists, and denote by t the value of its expectation in $\boldsymbol{\lambda}$. In order to make this arm ε-optimal, we need to decrease the expectations of all the arms that are above the threshold $t + \varepsilon$. We pay a cost of $\frac{1}{2}\omega_k(t - \mu_k)^2$ for moving arm k and of $\frac{1}{2}\omega_i(t + \varepsilon - \mu_i)^2$ for every arm i such that $\mu_i > t + \varepsilon$. Consider the functions:

$$f_k(t) := \frac{1}{2}\omega_k(t - \mu_k)^2 \text{ and } f_i(t) := \begin{cases} \frac{1}{2}\omega_i(t + \varepsilon - \mu_i)^2 & \text{for } t < \mu_i - \varepsilon, \\ 0 & \text{for } t \geq \mu_i - \varepsilon. \end{cases} \qquad \forall i \in [K] \setminus \{k\}$$

Each of these functions is convex. Therefore the function $f(t) := \sum_{i=1}^{K} f_i(t)$ is convex and has a unique minimizer t^*. One can easily check that $f'(\mu_k) \leq 0$ and $f'(\mu_1 - \varepsilon) \geq 0$, implying that $\mu_k - \varepsilon < \mu_k \leq t^* \leq \mu_1 - \varepsilon$. Therefore

$$\ell := \min\{i \geq 1 \ : \ t^* > \mu_i - \varepsilon\} - 1$$

is well defined and satisfies $\ell \in [|1, k-1|]$. Note that by definition $\mu_{\ell+1} - \varepsilon < t^*$ and $t^* \leq \mu_a - \varepsilon$ for all $a \leq \ell$, hence

$$0 = f'(t^*) = \omega_k(t^* - \mu_k) + \sum_{a=1}^{\ell} \omega_a(t^* + \varepsilon - \mu_a).$$

This implies that[1]

$$t^* = \overline{\boldsymbol{\mu}}_\varepsilon^{k,\ell}(\boldsymbol{\omega}) := \frac{\omega_k\mu_k + \sum_{a=1}^{\ell} \omega_a(\mu_a - \varepsilon)}{\omega_k + \sum_{a=1}^{\ell} \omega_a}$$

and the alternative bandit in this case writes as:

$$\boldsymbol{\lambda}_\varepsilon^{k,\ell}(\boldsymbol{\omega}) := (\underbrace{\overline{\boldsymbol{\mu}}_\varepsilon^{k,\ell}(\boldsymbol{\omega}) + \varepsilon}_{\text{indices 1to } \ell}, \mu_{\ell+1}, \ldots, \underbrace{\overline{\boldsymbol{\mu}}_\varepsilon^{k,\ell}(\boldsymbol{\omega})}_{\text{index } k}, \ldots, \mu_K)^\mathsf{T}.$$

Observe that since ℓ depends on t^*, we can't directly compute t^* from the expression above. Instead, we use the fact that ℓ is unique by definition. Therefore, to determine t^* one can compute $\overline{\boldsymbol{\mu}}_\varepsilon^{k,\ell}(\boldsymbol{\omega})$ for all values of $\ell \in [|1, k-1|]$ and search for the index ℓ satisfying $\mu_{\ell+1} - \varepsilon < \overline{\boldsymbol{\mu}}_\varepsilon^{k,\ell}(\boldsymbol{\omega}) \leq \mu_\ell - \varepsilon$ and with minimum value in the objective (5.2).

As a summary, we have reduced the minimization problem over the infinite set $\mathrm{Alt}(\boldsymbol{\mu})$ to a combinatorial search over a finite number of alternative bandit instances whose analytical expression is given in the next definition.

Definition 5.1 Let $\boldsymbol{\lambda}_\varepsilon^{k,\ell}(\boldsymbol{\omega})$ be a vector created form $\boldsymbol{\mu}$ by replacing elements on positions k and ℓ (resp. 1 to ℓ), defined as

$$\boldsymbol{\lambda}_\varepsilon^{k,\ell}(\boldsymbol{\omega}) := (\mu_1, \ldots, \underbrace{\overline{\boldsymbol{\mu}}_\varepsilon^{k,\ell}(\boldsymbol{\omega})}_{\text{index } k}, \ldots, \underbrace{\overline{\boldsymbol{\mu}}_\varepsilon^{k,\ell}(\boldsymbol{\omega}) + \varepsilon}_{\text{index } \ell}, \ldots, \mu_K)^\mathsf{T}$$

for $k \in G_\varepsilon(\boldsymbol{\mu})$ and

$$\boldsymbol{\lambda}_\varepsilon^{k,\ell}(\boldsymbol{\omega}) := (\underbrace{\overline{\boldsymbol{\mu}}_\varepsilon^{k,\ell}(\boldsymbol{\omega}) + \varepsilon}_{\text{indices 1to } \ell}, \mu_{\ell+1}, \ldots, \underbrace{\overline{\boldsymbol{\mu}}_\varepsilon^{k,\ell}(\boldsymbol{\omega})}_{\text{index } k}, \ldots, \mu_K)^\mathsf{T}$$

for $k \in B_\varepsilon(\boldsymbol{\mu})$ where $\overline{\boldsymbol{\mu}}_\varepsilon^{k,\ell}(\boldsymbol{\omega})$ is a weighted average of elements on positions k and ℓ (resp.

[1] $\overline{\boldsymbol{\mu}}_\varepsilon^{k,\ell}(\boldsymbol{\omega})$ has a different definition depending on k being a good or a bad arm.

1 to ℓ) defined as:

$$\overline{\boldsymbol{\mu}}_{\varepsilon}^{k,\ell}(\boldsymbol{\omega}) := \frac{\omega_k \mu_k + \omega_\ell(\mu_\ell - \varepsilon)}{\omega_k + \omega_\ell}$$

for $k \in G_\varepsilon(\boldsymbol{\mu})$ and

$$\overline{\boldsymbol{\mu}}_{\varepsilon}^{k,\ell}(\boldsymbol{\omega}) := \frac{\omega_k \mu_k + \sum_{a=1}^{\ell} \omega_a(\mu_a - \varepsilon)}{\omega_k + \sum_{a=1}^{\ell} \omega_a}$$

for $k \in B_\varepsilon(\boldsymbol{\mu})$.

The next lemma is a direct conclusion of the reasoning above.

Lemma 5.1 Using the previous definition, $\boldsymbol{\lambda}^*_{\varepsilon,\boldsymbol{\mu}}(\boldsymbol{\omega})$ can be computed as

$$\boldsymbol{\lambda}^*_{\varepsilon,\boldsymbol{\mu}}(\boldsymbol{\omega}) = \operatorname*{arg\,min}_{\boldsymbol{\lambda} \in \Lambda_G \cup \Lambda_B} \sum_{a \in [K]} \omega_a \frac{(\mu_a - \lambda_a)^2}{2}$$

with ties broken arbitrarily and where

$$\Lambda_G := \{\boldsymbol{\lambda}_{\varepsilon}^{k,\ell}(\boldsymbol{\omega}) : k \in G_\varepsilon(\boldsymbol{\mu}), \ell \in G_\varepsilon(\boldsymbol{\mu})/\{k\}\}$$

and

$$\Lambda_B := \{\boldsymbol{\lambda}_{\varepsilon}^{k,\ell}(\boldsymbol{\omega}) : k \in B_\varepsilon(\boldsymbol{\mu}), \ell \in [|1, k-1|] \text{ s.t. } \mu_\ell \geq \overline{\boldsymbol{\mu}}_{\varepsilon}^{k,\ell}(\boldsymbol{\omega}) + \varepsilon > \mu_{\ell+1}\}.$$

5.4 Solving the Max-Min Problem: Optimal Weights

First observe that we can rewrite $T_\varepsilon(\boldsymbol{\mu}, .)^{-1}$ as a minimum of linear functions:

$$T_\varepsilon(\boldsymbol{\mu}, \boldsymbol{\omega})^{-1} = \inf_{\boldsymbol{d} \in \mathcal{D}_{\varepsilon,\boldsymbol{\mu}}} \boldsymbol{\omega}^\intercal \boldsymbol{d} \text{ where } \mathcal{D}_{\varepsilon,\boldsymbol{\mu}} := \left\{ \left(\frac{(\lambda_a - \mu_a)^2}{2} \right)_{a \in [K]}^\intercal \;\middle|\; \boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\mu}) \right\}. \quad (5.6)$$

Note that by using $\mathcal{D}_{\varepsilon,\boldsymbol{\mu}}$ instead of $\text{Alt}(\boldsymbol{\mu})$, the optimization function becomes simpler for the price of more complex domain (see Figure 5.2 for an example). As a result, $T_\varepsilon(\boldsymbol{\mu}, .)^{-1}$ is concave and we can compute its subgradients thanks to Danskin's Theorem (Danskin, 1966) which we recall in the lemma below.

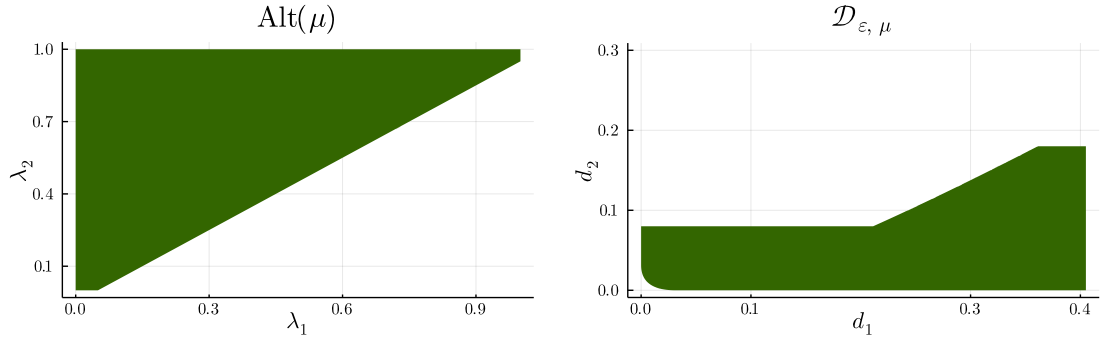


Figure 5.2: Comparison of $\text{Alt}(\boldsymbol{\mu})$ with Simple Linear Boundaries (First Figure) and $\mathcal{D}_{\varepsilon,\boldsymbol{\mu}}$ with Non-Linear Boundaries (Second Figure) for $\boldsymbol{\mu} = [0.9, 0.6]$ and $\varepsilon = 0.05$.

Lemma 5.2 (Danskin's Theorem) Let $\boldsymbol{\lambda}^*(\boldsymbol{\omega})$ be a best response to $\boldsymbol{\omega}$ and define $\boldsymbol{d}^*(\boldsymbol{\omega}) := \left(\frac{(\boldsymbol{\lambda}^*(\boldsymbol{\omega})_a - \mu_a)^2}{2}\right)_{a\in[K]}^\mathsf{T}$. Then $\boldsymbol{d}^*(\boldsymbol{\omega})$ is a subgradient of $T_\varepsilon(\boldsymbol{\mu}, .)^{-1}$ at $\boldsymbol{\omega}$.

Next we prove that $T_\varepsilon(\boldsymbol{\mu}, .)^{-1}$ is Liptschiz.

Lemma 5.3 The function $\boldsymbol{\omega} \mapsto T_\varepsilon(\boldsymbol{\mu}, \boldsymbol{\omega})^{-1}$ is L-Lipschitz with respect to $\|\cdot\|_1$ for any

$$L \geq \max_{a,b\in[K]} \frac{(\mu_a - \mu_b + \varepsilon)^2}{2} .$$

Proof. As we showed in Lemma 5.1, the best response $\boldsymbol{\lambda}^*_{\varepsilon,\boldsymbol{\mu}}(\boldsymbol{\omega})$ to $\boldsymbol{\omega}$ is created from $\boldsymbol{\mu}$ by replacing some of the elements by $\overline{\boldsymbol{\mu}}^{k,\ell}_\varepsilon(\boldsymbol{\omega})$ or $\overline{\boldsymbol{\mu}}^{k,\ell}_\varepsilon(\boldsymbol{\omega}) + \varepsilon$. We also know that $\overline{\boldsymbol{\mu}}^{k,\ell}_\varepsilon(\boldsymbol{\omega})$ is a weighted average of an element of $\boldsymbol{\mu}$ with one or more elements of $\boldsymbol{\mu}$ decreased by ε. This means that

$$\max_{a\in[K]} \mu_a \geq \overline{\boldsymbol{\mu}}^{k,\ell}_\varepsilon(\boldsymbol{\omega}) \geq \min_{a\in[K]} \mu_a - \varepsilon$$

and, as a consequence, we have

$$|\mu_i - \boldsymbol{\lambda}^*_{\varepsilon,\boldsymbol{\mu}}(\boldsymbol{\omega})_i| \leq \max_{a,b\in[K]} (\mu_a - \mu_b + \varepsilon)$$

for any $i \in [K]$. Let $f(\boldsymbol{\omega}) := T_\varepsilon(\boldsymbol{\mu}, \boldsymbol{\omega})^{-1}$. Using the last inequality and the definition of $\boldsymbol{d}^*(\boldsymbol{\omega})$, we can obtain for any $\boldsymbol{\omega}$, $\boldsymbol{\omega}' \in \Sigma_K$,

$$\begin{aligned} f(\boldsymbol{\omega}) - f(\boldsymbol{\omega}') &\leq (\boldsymbol{\omega} - \boldsymbol{\omega}')^\mathsf{T} \boldsymbol{d}^*(\boldsymbol{\omega}') \\ &\leq \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_1 \|\boldsymbol{d}^*(\boldsymbol{\omega}')\|_\infty \\ &\leq \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_1 \max_{a,b\in[K]} \frac{(\mu_a - \mu_b + \varepsilon)^2}{2} \end{aligned}$$

■

As a summary $T_\varepsilon(\boldsymbol{\mu}, .)^{-1}$ is concave, Lipschitz and we have a simple expression to compute its subgradients through the best response oracle. Therefore we have all the necessary ingredients to apply a gradient-based algorithm in order to find the optimal weights and therefore, the value of $T^\star_\varepsilon(\boldsymbol{\mu})$. The algorithm of our choice is the mirror ascent algorithm which enjoys the following guarantees.

Proposition 5.2 — (Bubeck, 2015). Let $\boldsymbol{\omega}_1 = (\frac{1}{K}, \dots, \frac{1}{K})^\mathsf{T}$ and define the learning rate $\alpha_n = \frac{1}{L}\sqrt{\frac{2\log K}{n}}$. Then using mirror ascent algorithm to maximize a L-Lipschitz function f, with respect to $\|\cdot\|_1$, defined on Δ_K with generalized negative entropy $\Phi(\boldsymbol{\omega}) = \sum_{a\in[K]} \omega_a \log(\omega_a)$ as the mirror map enjoys the following guarantees:

$$f(\boldsymbol{\omega}^*) - f\left(\frac{1}{N}\sum_{n=1}^N \boldsymbol{\omega}_n\right) \leq L\sqrt{\frac{2\log K}{N}} .$$

Remark 5.2 — Computational complexity of our algorithm.. To simplify the presentation and analysis, we chose to focus on the vanilla version of Track-and-Stop. However, in practice this requires solving the optimization program that appears in the lower bound at every time step, which can result in large run times. Nonetheless, we note that there are many possible adaptations of Track-and-Stop that reduce the computational

complexity, while retaining the guarantees of asymptotic optimality in terms of the sample complexity (and with a demonstrated small performance loss experimentally). A first solution is to use Franke-Wolfe style algorithms (Ménard, 2019; Wang et al., 2021), which only perform a gradient step of the optimization program at every step. Once can also apply the Gaming approach initiated by (Degenne et al., 2019a) which only needs to solve the best response problem, and runs a no-regret learner such as AdaHedge to determine the weights to be tracked at each step. This approach was used for example by (Jourdan et al., 2021) in a similar setting of pure exploration with semi-bandit feedback. Another adaptation is the Lazy Track-and-Stop (Jedra & Proutiere, 2020), which updates the weights that are tracked by the algorithms every once in a while. ■

5.5 Comparing the Simulator Lower Bound to the Characteristic Time

In this section, we show that the simulator bound of Theorem 1.2 can be arbitrarily large compared to $T^\star_\varepsilon(\boldsymbol{\mu})\log(1/\delta)$. Fix $\delta = 0.1$ and let $\varepsilon, \beta > 0$ with $\beta \ll \varepsilon$ and consider the instance such that $\mu_1 = \beta, \mu_K = -\varepsilon$ and $\mu_a = -\beta$ for $a \in [|2, K-1|]$. Note that in this case $\beta_\varepsilon = \beta$, where β_ε was defined in (1.51). By symmetry, $\omega_a = \omega_2$ for all $a \in [|2, K-1|]$. In this case, using Lemma 5.1 we have

$$\begin{aligned} T^\star_\varepsilon(\boldsymbol{\mu})^{-1} &\overset{(a)}{=} \sup_{\omega\in\Sigma_K} \min\left(\frac{\omega_1\omega_K\beta^2}{2(\omega_1+\omega_K)}, \frac{\omega_1\omega_2(\varepsilon-2\beta)^2}{2(\omega_1+\omega_2)}, \frac{\omega_2\omega_3\varepsilon^2}{2(\omega_2+\omega_3)}\right) \\ &= \sup_{\omega\in\Sigma_K} \min\left(\frac{\omega_1\omega_K\beta^2}{2(\omega_1+\omega_K)}, \frac{\omega_1\omega_2(\varepsilon-2\beta)^2}{2(\omega_1+\omega_2)}, \frac{\omega_2\varepsilon^2}{4}\right) \\ &\geq \sup_{\omega\in\Sigma_K} \min\left(\frac{\omega_1\omega_K\beta^2}{2(\omega_1+\omega_K)}, \frac{\omega_1\omega_2(\varepsilon-2\beta)^2}{2(\omega_1+\omega_2)}, \frac{\omega_2(\varepsilon-2\beta)^2}{4}\right). \end{aligned}$$

The first term of the min in (a) corresponds to the cost of making arm K a good arm by simultaneously increasing its mean reward and decreasing the mean reward of the first arm, the second term to that of making arm 2 a bad arm by simultaneously decreasing its mean reward and increasing the mean reward of the first arm. The third term, corresponds to the cost of making arm 2 a bad arm by simultaneously decreasing its mean reward and increasing the mean reward of the arm 3 (which is the same cost if we replace arms 2 and 3 by any other pair of arms in $[|2, K-1|]$). Now we look for $\boldsymbol{\omega}$ such that $\omega_1 = \omega_K > \omega_2$. This means that the third term of the min in the last line is always smaller than the second term. If we note S the set of such omegas then one can write

$$T^\star_\varepsilon(\boldsymbol{\mu})^{-1} \geq \sup_{\omega\in\Sigma_K\cap S} \min\left(\frac{\omega_1\beta^2}{4}, \frac{\omega_2(\varepsilon-2\beta)^2}{4}\right) \tag{5.7}$$

Note that the right hand side is maximized when both terms of the min are equal. Let $\widetilde{\omega}$ be the maximizer. Then

$$2\widetilde{\omega}_1 + (K-2)\widetilde{\omega}_2 = 1, \quad \text{and} \quad \widetilde{\omega}_1\beta^2 = \widetilde{\omega}_2(\varepsilon-2\beta)^2.$$

Solving for $\widetilde{\omega}$ and injecting in (5.7) we get

$$T^\star_\varepsilon(\boldsymbol{\mu})^{-1} \geq \frac{(\varepsilon-2\beta)^2\beta^2}{8(\varepsilon-2\beta)^2 + 4(K-2)\beta^2}$$

or equivalently

$$T^\star_\varepsilon(\boldsymbol{\mu}) \leq \frac{8}{\beta^2} + \frac{4(K-2)}{(\varepsilon-2\beta)^2}.$$

When $\beta \ll \varepsilon$ and δ is fixed, this yields $T^\star_\varepsilon(\boldsymbol{\mu})\log(1/\delta) = \mathcal{O}(1/\beta^2 + K/\varepsilon^2)$. In contrast note that for this particular instance $|G_\beta(\boldsymbol{\mu})| = 1$ so that the lower bound of Theorem 1.2 is at least of order $\Omega(K/\beta^2)$. Therefore, we see that the simulator bound exhibits an improved scaling w.r.t the number of arms K.

5.6 Proof of Theorem 5.1

We start with a few technical lemmas. The first two are adapted from (Garivier & Kaufmann, 2016):

Lemma 5.4 (Lemma 7, (Garivier & Kaufmann, 2016)) For all $t \geq 1$, the C-Tracking with weights $(\widetilde{\boldsymbol{\omega}}(\widehat{\boldsymbol{\mu}}_s))_{s\in\mathbb{N}^*}$ ensures that $N_a(t) \geq \sqrt{t+K^2} - 2K$ and that

$$\max_{1\leq 1\leq K} \left| N_a(t) - \sum_{s=1}^{t} \widetilde{\boldsymbol{\omega}}_a(\widehat{\boldsymbol{\mu}}_s) \right| \leq K(1+\sqrt{t})$$

Lemma 5.5 (Lemma 19, (Garivier & Kaufmann, 2016)) For $\xi > 0$, define $I_\xi \triangleq [\mu_1 - \xi, \mu_1 + \xi] \times \dots [\mu_K - \xi, \mu_K + \xi]$. And for $T \geq 1$, consider the event: $\mathcal{E}_T = \bigcap_{t=\lfloor T^{1/4}\rfloor}^{T} (\widehat{\boldsymbol{\mu}}_t \in I_\xi)$. Then there exists two constants B, C that only depend on $\boldsymbol{\mu}$ and ξ such that

$$\mathbb{P}_{\mu,\mathbb{A}}(\mathcal{E}_T^c) \leq BT\exp(-CT^{1/8})$$

where $\mathcal{E}_T^c$ is the complementary event of $\mathcal{E}_T$.

The last lemma states that $\boldsymbol{\mu} \mapsto T_\varepsilon(\boldsymbol{\mu}, \boldsymbol{\omega})^{-1}$ is Lipschitz. Its proof is deferred to the end.

Lemma 5.6 For all vectors $\boldsymbol{\omega}$ in the simplex, for all instances $\boldsymbol{\mu}$, $\boldsymbol{\mu}'$ in $[\mu_{min}, \mu_{max}]^K$ we have

$$|T_\varepsilon(\boldsymbol{\mu}', \boldsymbol{\omega})^{-1} - T_\varepsilon(\boldsymbol{\mu}, \boldsymbol{\omega})^{-1}| \leq 4(\mu_{max} - \mu_{min} + \varepsilon) \left\|\boldsymbol{\mu}' - \boldsymbol{\mu}\right\|_\infty .$$

Now we are ready to prove the Theorem. We denote by $L_1([\mu_{min}, \mu_{max}]^K) \triangleq 4(\mu_{max} - \mu_{min} + \varepsilon)$ the Lipschitz constant of the mapping $\boldsymbol{\mu} \mapsto T_\varepsilon(\boldsymbol{\mu}, \boldsymbol{\omega})^{-1}$ in the domain $[\mu_{min}, \mu_{max}]^K$ and by $L_2(\boldsymbol{\mu}) \triangleq \max_{a,b\in[K]} \frac{(\mu_a - \mu_b + \varepsilon)^2}{2}$ the Lipschitz constant of the mapping $\boldsymbol{\omega} \mapsto T_\varepsilon(\boldsymbol{\mu}, \boldsymbol{\omega})^{-1}$.

We will prove a lower bound on $T_\varepsilon\big(\widehat{\boldsymbol{\mu}}_t, \frac{N(t)}{t}\big)^{-1}$ under $\mathcal{E}_T$ which will result into an upper bound on the stopping time $\tau_\delta = \inf\left\{t \in \mathbb{N} \,:\, tT_\varepsilon\big(\widehat{\boldsymbol{\mu}}_t, \frac{N(t)}{t}\big)^{-1} \geq \beta(\delta, t)\right\}$. First observe that under $\mathcal{E}_T$, the L_1 constant is upper bounded: $L_1(I_\xi) \leq L_{1,max} \triangleq 4(\max_a \mu_a - \min_b \mu_b + \varepsilon + 2\xi)$. Similarly, we have for all $\lfloor T^{1/4}\rfloor \leq t \leq T$, $L_2(\widehat{\boldsymbol{\mu}}_t) \leq L_{2,max} \triangleq \frac{(\max_a \mu_a - \min_b \mu_b + 2\xi + \varepsilon)^2}{2}$. Now applying Lemma 5.4 and the Lipschitz property w.r.t the weights, we have for all

$\lfloor T^{1/4} \rfloor \leq t \leq T$

$$
\begin{aligned}
T_\varepsilon\left(\widehat{\boldsymbol{\mu}}_t, \frac{N(t)}{t}\right)^{-1} &\geq T_\varepsilon\left(\widehat{\boldsymbol{\mu}}_t, \frac{\sum_{s=1}^{t} \widetilde{\boldsymbol{\omega}}(\widehat{\boldsymbol{\mu}}_s)}{t}\right)^{-1} - L_{2,max}\frac{K(1+\sqrt{t})}{t} \\
&\geq \frac{\sum_{s=1}^{t} T_\varepsilon\big(\widehat{\boldsymbol{\mu}}_t, \widetilde{\boldsymbol{\omega}}(\widehat{\boldsymbol{\mu}}_s)\big)^{-1}}{t} - L_{2,max}\frac{K(1+\sqrt{t})}{t} \\
&\geq \frac{\sum_{s=\lfloor T^{1/4} \rfloor}^{t} T_\varepsilon\big(\widehat{\boldsymbol{\mu}}_t, \widetilde{\boldsymbol{\omega}}(\widehat{\boldsymbol{\mu}}_s)\big)^{-1}}{t} - L_{2,max}\frac{K(1+\sqrt{t})}{t},
\end{aligned}
$$

where we used the fact that the mapping $\boldsymbol{\omega} \mapsto T_\varepsilon(\boldsymbol{\mu}, \boldsymbol{\omega})^{-1}$ is concave (resp. non-negative) in the second (resp. last) inequality. Now observe that for all $s, t \geq \lfloor T^{1/4} \rfloor$, $\|\widehat{\boldsymbol{\mu}}_t - \widehat{\boldsymbol{\mu}}_s\|_\infty \leq 2\xi$. Therefore the Lipschitz property w.r.t $\boldsymbol{\mu}$ implies that

$$
\begin{aligned}
T_\varepsilon\big(\widehat{\boldsymbol{\mu}}_t, \frac{N(t)}{t}\big)^{-1} &\geq \frac{\sum_{s=\lfloor T^{1/4} \rfloor}^{t} T_\varepsilon\big(\widehat{\boldsymbol{\mu}}_s, \widetilde{\boldsymbol{\omega}}(\widehat{\boldsymbol{\mu}}_s)\big)^{-1}}{t} - \frac{2\xi L_{1,max}(t - \lfloor T^{1/4} \rfloor)}{t} - L_{2,max}\frac{K(1+\sqrt{t})}{t} \\
&\geq \frac{\sum_{s=\lfloor T^{1/4} \rfloor}^{t} T_\varepsilon^*\big(\widehat{\boldsymbol{\mu}}_s\big)^{-1}}{t} - \frac{\sum_{s=\lfloor T^{1/4} \rfloor}^{t} \frac{1}{\sqrt{s}}}{t} - 2\xi L_{1,max} - L_{2,max}\frac{K(1+\sqrt{t})}{t} \qquad (5.8)
\end{aligned}
$$

where in the second inequality we used the fact that by definition $\widetilde{\boldsymbol{\omega}}(\widehat{\boldsymbol{\mu}}_s)$ is at most $\frac{1}{\sqrt{s}}$ sub-optimal. Now observe that

$$
\begin{aligned}
&T_\varepsilon^*\big(\widehat{\boldsymbol{\mu}}_s\big)^{-1} \underset{s\to\infty}{\longrightarrow} T_\varepsilon^*\big(\boldsymbol{\mu}\big)^{-1} \quad \text{almost surely (since } N_a(t) \geq \sqrt{t + K^2} - 2K\text{).} \\
&\frac{\sum_{s=\lfloor T^{1/4} \rfloor}^{t} \frac{1}{\sqrt{s}}}{t} \underset{t\to\infty}{\sim} \frac{\int_1^t \frac{dx}{\sqrt{x}}}{t} \longrightarrow 0. \\
&\frac{K(1+\sqrt{t})}{t} \longrightarrow 0.
\end{aligned}
$$

Therefore for $\eta > 0$, there exists t_η such that for all $t \geq t_\eta$,

$$
\frac{\sum_{s=\lfloor T^{1/4} \rfloor}^{t} T_\varepsilon^*\big(\widehat{\boldsymbol{\mu}}_s\big)^{-1}}{t} - \frac{\sum_{s=\lfloor T^{1/4} \rfloor}^{t} \frac{1}{\sqrt{s}}}{t} - L_{2,max}\frac{K(1+\sqrt{t})}{t} \geq T_\varepsilon^*\big(\boldsymbol{\mu}\big)^{-1} - \eta. \qquad (5.9)
$$

Summing up (5.8) and (5.9), we get for all $t \geq t_\eta$,

$$
T_\varepsilon\big(\widehat{\boldsymbol{\mu}}_t, \frac{N(t)}{t}\big)^{-1} \geq T_\varepsilon^*\big(\boldsymbol{\mu}\big)^{-1} - 2\xi L_{1,max} - \eta.
$$

Therefore for every T such that $T \geq \max\big(t_\eta, \frac{\beta(\delta,T)}{T_\varepsilon^*\big(\boldsymbol{\mu}\big)^{-1} - 2\xi L_{1,max} - \eta}\big)$, we have $\mathcal{E}_T \subset (\tau_\delta \leq T)$ thus $\mathbb{P}(\tau_\delta > T) \leq \mathbb{P}\big(\mathcal{E}_T^c\big) \leq BT\exp(-CT^{1/8})$. Hence for $T_0(\delta) := \inf\Big\{T \geq 1 : \ T \geq$

$\max\left(t_\eta, \frac{\beta(\delta,T)}{T_\varepsilon^*(\boldsymbol{\mu})^{-1}-2\xi L_{1,max}-\eta}\right)\Big\}$ it holds that

$$\begin{aligned}\mathbb{E}[\tau_\delta] &= \sum_{T=1}^{\infty}\mathbb{P}(\tau_\delta > T)\\ &\leq T_0(\delta) + \sum_{T=1}^{\infty} BT\exp(-CT^{1/8}).\end{aligned}$$

Note that

$$\max\left(t_\eta, \frac{\beta(\delta,T_0(\delta))}{T_\varepsilon^*(\boldsymbol{\mu})^{-1}-2\xi L_{1,max}-\eta}\right) \leq T_0(\delta) \leq \max\left(t_\eta, \frac{\beta(\delta,T_0(\delta))}{T_\varepsilon^*(\boldsymbol{\mu})^{-1}-2\xi L_{1,max}-\eta}\right)+1.$$

Since $\lim_{\delta\to 0}\frac{\beta(\delta,t)}{\log(1/\delta)}=1$, the last inequality implies that $\limsup_{\delta\to 0}\frac{T_0(\delta)}{\log(1/\delta)} \leq \frac{1}{T_\varepsilon^*(\boldsymbol{\mu})^{-1}-2\xi L_{1,max}-\eta}$ and consequently $\limsup_{\delta\to 0}\frac{\mathbb{E}[\tau_\delta]}{\log(1/\delta)} \leq \frac{1}{T_\varepsilon^*(\boldsymbol{\mu})^{-1}-2\xi L_{1,max}-\eta}$. We conclude by letting η and ξ go to zero. ■

5.6.1 Proof of Lemma 5.6

First case: arms in $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ have the same order

Without loss of generality, suppose that $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_K$ and $\mu'_1 \geq \mu'_2 \geq \ldots \geq \mu'_K$. Then we see that for all $k \neq l \in [K]$, $\overline{\boldsymbol{\mu}}_\varepsilon^{k,\ell}(\boldsymbol{\omega})$ and $\overline{\boldsymbol{\mu}'}_\varepsilon^{k,\ell}(\boldsymbol{\omega})$ have the same formula and : $|\overline{\boldsymbol{\mu}'}_\varepsilon^{k,\ell}(\boldsymbol{\omega}) - \overline{\boldsymbol{\mu}}_\varepsilon^{k,\ell}(\boldsymbol{\omega})| \leq \|\boldsymbol{\mu}'-\boldsymbol{\mu}\|_\infty$, which implies that $|\boldsymbol{\lambda}'^{k,\ell}_\varepsilon(\boldsymbol{\omega}) - \boldsymbol{\lambda}^{k,\ell}_\varepsilon(\boldsymbol{\omega})| \leq \|\boldsymbol{\mu}'-\boldsymbol{\mu}\|_\infty$. Therefore, letting f denote the function $f(\boldsymbol{\mu},\boldsymbol{\lambda}) \triangleq \sum_{a\in[K]}\omega_a\frac{(\mu_a-\lambda_a)^2}{2}$, we have

$$\begin{aligned}|f(\boldsymbol{\mu}', \boldsymbol{\lambda}'^{k,\ell}_\varepsilon(\boldsymbol{\omega})) - f(\boldsymbol{\mu}, \boldsymbol{\lambda}^{k,\ell}_\varepsilon(\boldsymbol{\omega}))| &\leq \tfrac{1}{2}\sum_{a\in[K]}\omega_a(\mu'_a-\mu_a+\boldsymbol{\lambda}^{k,\ell}_\varepsilon(\boldsymbol{\omega})_a-\boldsymbol{\lambda}'^{k,\ell}_\varepsilon(\boldsymbol{\omega})_a)(\mu'_a+\mu_a-\boldsymbol{\lambda}^{k,\ell}_\varepsilon(\boldsymbol{\omega})_a-\boldsymbol{\lambda}'^{k,\ell}_\varepsilon(\boldsymbol{\omega})_a)\\ &\leq \frac{\omega_a\times 2\|\boldsymbol{\mu}'-\boldsymbol{\mu}\|_\infty \times 2(\mu_{max}-\mu_{min}+\varepsilon)}{2}\\ &= 2(\mu_{max}-\mu_{min}+\varepsilon)\|\boldsymbol{\mu}'-\boldsymbol{\mu}\|_\infty.\end{aligned}$$

where in the second inequality we used the fact that $\boldsymbol{\lambda}^{k,\ell}_\varepsilon(\boldsymbol{\omega})$ (resp. $\boldsymbol{\lambda}'^{k,\ell}_\varepsilon(\boldsymbol{\omega})$) is a weighted average of some arm in $\boldsymbol{\mu}$(resp. $\boldsymbol{\mu}'$) with one or more arms of $\boldsymbol{\mu}$(resp. $\boldsymbol{\mu}'$) decreased by ε and therefore lies in $[\mu_{min}-\varepsilon,\mu_{max}]^K$. Let (k_0,l_0) be such that $\boldsymbol{\lambda}^*_{\varepsilon,\boldsymbol{\mu}}(\boldsymbol{\omega}) = \boldsymbol{\lambda}^{k_0,l_0}_\varepsilon(\boldsymbol{\omega})$ then

$$\begin{aligned}T_\varepsilon(\boldsymbol{\mu}',\boldsymbol{\omega})^{-1} - T_\varepsilon(\boldsymbol{\mu},\boldsymbol{\omega})^{-1} &= T_\varepsilon(\boldsymbol{\mu}',\boldsymbol{\omega})^{-1} - f\big(\omega,\boldsymbol{\lambda}^{k_0,l_0}_\varepsilon(\boldsymbol{\omega})\big)\\ &\leq f\big(\omega,\boldsymbol{\lambda}'^{k_0,l_0}_\varepsilon(\boldsymbol{\omega})\big) - f\big(\omega,\boldsymbol{\lambda}^{k_0,l_0}_\varepsilon(\boldsymbol{\omega})\big)\\ &\leq 2(\mu_{max}-\mu_{min}+\varepsilon)\|\boldsymbol{\mu}'-\boldsymbol{\mu}\|_\infty.\end{aligned}$$

By symmetry we get for all instances $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ with the same arm ordering:

$$|T_\varepsilon(\boldsymbol{\mu}',\boldsymbol{\omega})^{-1} - T_\varepsilon(\boldsymbol{\mu},\boldsymbol{\omega})^{-1}| \leq 2(\mu_{max}-\mu_{min}+\varepsilon)\|\boldsymbol{\mu}'-\boldsymbol{\mu}\|_\infty.$$

Second case: arms in $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ have a different order

Then for n large enough we can find a sequence $(\mu^i)_{0\leq i\leq 2^n}$ of instances in the segment $[\boldsymbol{\mu},\boldsymbol{\mu}']$ such that $\mu^0=\boldsymbol{\mu}$, $\mu^{2^n}=\boldsymbol{\mu}'$ and:

$\forall i \in [|0,2^n-1|]$, μ^i and μ^{i+1} have the same arm ordering and $\|\mu^{i+1}-\mu^i\|_\infty \leq \frac{\|\boldsymbol{\mu}'-\boldsymbol{\mu}\|_\infty}{2^{n-1}}$.

We can construct such a sequence in the following way: Split $[\mu_{min}, \mu_{max}]^K$ into $K!$ regions such that any two instances in the same region share the same arm ordering. The boundaries between these regions correspond to instances where two or more arms are equal. Starting from $\mu^0 \triangleq \boldsymbol{\mu}$, span the segment $[\boldsymbol{\mu}, \boldsymbol{\mu}']$ and define μ^{i+1} to be the first instance where: either the L^∞ distance from μ^i is equal to $\frac{\|\boldsymbol{\mu}'-\boldsymbol{\mu}\|_\infty}{2^{n-1}}$, or we cross a boundary between two regions. Since there can be at most $K! - 1$ changes in the arm ordering, for n large enough such sequence always exists. Now we have:

$$\begin{aligned}
|T_\varepsilon(\boldsymbol{\mu}', \boldsymbol{\omega})^{-1} - T_\varepsilon(\boldsymbol{\mu}, \boldsymbol{\omega})^{-1}| &\leq \sum_{i=0}^{2^n-1} |T_\varepsilon(\boldsymbol{\omega}, \boldsymbol{\mu}^{i+1})^{-1} - T_\varepsilon(\boldsymbol{\omega}, \boldsymbol{\mu}^i)^{-1}| \\
&\leq \sum_{i=0}^{2^n-1} 2(\mu_{max} - \mu_{min} + \varepsilon) \frac{\|\boldsymbol{\mu}' - \boldsymbol{\mu}\|_\infty}{2^{n-1}} \\
&\leq 4(\mu_{max} - \mu_{min} + \varepsilon) \left\|\boldsymbol{\mu}' - \boldsymbol{\mu}\right\|_\infty .
\end{aligned}$$

where in the second inequality we use the first case and the fact that μ^i and μ^{i+1} have the same arm ordering. As a summary, we always have:

$$|T_\varepsilon(\boldsymbol{\mu}', \boldsymbol{\omega})^{-1} - T_\varepsilon(\boldsymbol{\mu}, \boldsymbol{\omega})^{-1}| \leq 4(\mu_{max} - \mu_{min} + \varepsilon) \left\|\boldsymbol{\mu}' - \boldsymbol{\mu}\right\|_\infty .$$

5.7 Conclusion

We shed a new light on the sample complexity of finding all the ε-good arms in a multi-armed bandit with Gaussian rewards. We derived an instance-dependent lower bound, identifying the characteristic time that reflects the true hardness of the problem in the asymptotic regime. Then, capitalizing on an method to solve the optimization program that defines the characteristic time, we proposed an efficient Track-and-Stop strategy whose sample complexity matches the lower bound for small values of the risk level. Finally, we proved that the simulator bound from Chapter 1 can have a better scaling in the number of arms and can be arbitrarily larger than the first bound for moderate values of the risk.

Bibliography

Articles

Abbasi-Yadkori, Y., Pál, D., & Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, *24* (cited on pages 101–103).

Agarwal, A., Kakade, S., & Yang, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal (J. Abernethy & S. Agarwal, Editors). *Proceedings of Thirty Third Conference on Learning Theory*, *125*, 67–83 (cited on page 19).

Al Marjani, A., Kocak, T., & Garivier, A. (2022). On the complexity of all \epsilon-best arms identification. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 317–332 (cited on page 35).

Al-Marjani, A., Garivier, A., & Proutiere, A. (2021). Navigating to the best policy in markov decision processes. *Advances in Neural Information Processing Systems (NeurIPS)*, *34* (cited on pages 24, 61, 65, 67).

Al-Marjani, A., & Proutiere, A. (2021). Adaptive sampling for best policy identification in markov decision processes. *International Conference on Machine Learning*, 7459–7468 (cited on pages 28, 52, 54, 69).

Al-Marjani, A., Tirinzoni, A., & Kaufmann, E. (2023). Active coverage for PAC reinforcement learning. *Proceedings of the 36th Conference On Learning Theory (COLT)* (cited on pages 47, 84, 109).

Azar, M. G., Munos, R., & Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, *91*(3), 325–349 (cited on pages 18–20, 115).

Azar, M. G., Osband, I., & Munos, R. (2017). Minimax regret bounds for reinforcement learning. *Proceedings of the 34th International Conference on Machine Learning, (ICML)* (cited on pages 47, 83, 84).

Barrier, A., Garivier, A., & Kocák, T. (2022). A non-asymptotic approach to best-arm identification for gaussian bandits (G. Camps-Valls, F. J. R. Ruiz, & I. Valera, Editors). *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, *151*, 10078–10109 (cited on page 68).

Brandizi, M., Kurbatova, N., Sarkans, U., & Rocca-Serra, P. (2012). Graph2tab, a library to convert experimental workflow graphs into tabular formats. *Bioinformatics*, *28*(12), 1665–1667 (cited on page 41).

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning* (cited on page 150).

Burda, Y., Edwards, H., Storkey, A., & Klimov, O. (2019). Exploration by random network distillation. *International Conference on Learning Representations* (cited on page 76).

Burnetas, A. N., & Katehakis, M. N. (1997). Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, *22*(1), 222–255 (cited on pages 54, 72).

Cesa-Bianchi, N., Mansour, Y., & Stoltz, G. (2005). Improved second-order bounds for prediction with expert advice. *Machine Learning*, *66*, 321–352 (cited on pages 84, 89).

Chen, J., & Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. *International Conference on Machine Learning*, 1042–1051 (cited on pages 76, 82).

Chen, J., Modi, A., Krishnamurthy, A., Jiang, N., & Agarwal, A. (2022). On the statistical efficiency of reward-free exploration in non-linear rl. *Advances in Neural Information Processing Systems* (cited on page 76).

Chernoff, H. (1959). Sequential design of Experiments. *The Annals of Mathematical Statistics*, *30*(3), 755–770 (cited on page 145).

Cheung, W. C. (2019). Exploration-exploitation trade-off in reinforcement learning on online markov decision processes with global concave rewards. *arXiv preprint arXiv:1905.06466* (cited on page 76).

Dann, C., & Brunskill, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)* (cited on page 19).

Dann, C., Lattimore, T., & Brunskill, E. (2017). Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 5713–5723 (cited on page 106).

Dann, C., Li, L., Wei, W., & Brunskill, E. (2019). Policy certificates: Towards accountable reinforcement learning. *International Conference on Machine Learning (ICML)* (cited on page 19).

Dann, C., Marinov, T. V., Mohri, M., & Zimmert, J. (2021). Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)* (cited on page 112).

Danskin, J. M. (1966). The theory of max-min, with applications. *Siam Journal on Applied Mathematics*, *14*, 641–664 (cited on page 149).

Degenne, R., & Koolen, W. M. (2019). Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems (NeurIPS)* (cited on pages 29, 112, 131).

Degenne, R., Koolen, W. M., & Ménard, P. (2019a). Non-asymptotic pure exploration by solving games. *Neural Information Processing Systems* (cited on pages 24, 29, 145, 151).

Degenne, R., Koolen, W. M., & Ménard, P. (2019b). Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems (NeurIPS)*, 14492–14501 (cited on page 82).

Degenne, R., Menard, P., Shang, X., & Valko, M. (2020). Gamification of pure exploration for linear bandits (H. D. III & A. Singh, Editors). *Proceedings of the 37th International Conference on Machine Learning*, *119*, 2432–2442 (cited on pages 28, 74).

Domingues, O. D., Ménard, P., Kaufmann, E., & Valko, M. (2021). Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. *Algorithmic Learning Theory*, 578–598 (cited on page 19).

Efroni, Y., Mannor, S., & Pirotta, M. (2020). Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189* (cited on page 117).

Even-Dar, E., Mannor, S., & Mansour, Y. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, *7*(Jun), 1079–1105 (cited on page 19).

Eysenbach, B., Gupta, A., Ibarz, J., & Levine, S. (2019). Diversity is all you need: Learning skills without a reward function. *International Conference on Learning Representations* (cited on page 76).

Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C., & Mannor, S. (2009). Regularized fitted q-iteration for planning in continuous-space markovian decision problems. *2009 American Control Conference*, 725–730 (cited on page 76).

Farahmand, A.-m., Szepesvári, C., & Munos, R. (2010). Error propagation for approximate policy and value iteration. *Advances in Neural Information Processing Systems (NeurIPS)*, *23* (cited on page 76).

Fiechter, C.-N. (1994). Efficient reinforcement learning. *Proceedings of the Seventh Conference on Computational Learning Theory (COLT)* (cited on page 18).

Fort, G., Moulines, É., & Priouret, P. (2011). Convergence of adaptive and interacting markov chain monte carlo algorithms. *Annals of Statistics*, *39*, 3262–3289 (cited on pages 61, 62, 67).

Foster, D. J., Krishnamurthy, A., Simchi-Levi, D., & Xu, Y. (2022). Offline reinforcement learning: Fundamental barriers for value function approximation. *Conference on Learning Theory*, 3489–3489 (cited on page 76).

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139 (cited on page 47).

Garivier, A. (2013). Informational confidence bounds for self-normalized averages and applications. *2013 IEEE Information Theory Workshop (ITW)* (cited on page 145).

Garivier, A., & Kaufmann, E. (2016). Optimal best arm identification with fixed confidence (V. Feldman, A. Rakhlin, & O. Shamir, Editors). *29th Annual Conference on Learning Theory*, *49* (cited on pages 22, 28, 29, 54, 113, 130, 145, 152).

Garivier, A., & Kaufmann, E. (2021). Nonasymptotic sequential tests for overlapping hypotheses applied to near-optimal arm identification in bandit models. *Sequential Analysis*, *40*(1), 61–96 (cited on pages 29, 137, 145).

Garivier, A., Ménard, P., & Stoltz, G. (2019). Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, *44*(2), 377–399 (cited on page 25).

Hazan, E., Kakade, S. M., Singh, K., & Soest, A. V. (2019). Provably efficient maximum entropy exploration. *International Conference on Machine Learning (ICML)* (cited on page 76).

Jaksch, T., Ortner, R., & Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, *11*, 1563–1600 (cited on page 77).

Jedra, Y., & Proutiere, A. (2020). Optimal best-arm identification in linear bandits (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin, Editors). *Advances in*

Neural Information Processing Systems, *33*, 10007–10017 (cited on pages 24, 29, 130, 151).

Jin, C., Krishnamurthy, A., Simchowitz, M., & Yu, T. (2020). Reward-free exploration for reinforcement learning. *International Conference on Machine Learning (ICML)* (cited on pages 21, 76, 84, 85, 89–91).

Jin, C., Yang, Z., Wang, Z., & Jordan, M. I. (2019). Provably efficient reinforcement learning with linear function approximation. *Annual Conference Computational Learning Theory* (cited on pages 115, 116).

Jin, Y., Yang, Z., & Wang, Z. (2021). Is pessimism provably efficient for offline rl? *International Conference on Machine Learning*, 5084–5096 (cited on page 76).

Johanson, M. B., Hughes, E., Timbers, F., & Leibo, J. Z. (2022). Emergent bartering behaviour in multi-agent reinforcement learning. *ArXiv*, *abs/2205.06760* (cited on page 9).

Johari, R., Koomen, P., Pekelis, L., & Walsh, D. (2022). Always valid inference: Continuous monitoring of a/b tests. *Operations Research*, *70*(3), 1806–1821 (cited on page 9).

Jourdan, M., Mutný, M., Kirschner, J., & Krause, A. (2021). Efficient pure exploration for combinatorial bandits with semi-bandit feedback (V. Feldman, K. Ligett, & S. Sabato, Editors). *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, *132*, 805–849 (cited on page 151).

Kalyanakrishnan, S., & Stone, P. (2010). Efficient selection of multiple bandit arms: Theory and practice. *International Conference on Machine Learning* (cited on page 16).

Kaufmann, E., Cappé, O., & Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, *17*(1), 1–42 (cited on pages 23, 33, 34, 144).

Kaufmann, E., & Koolen, W. M. (2021). Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, *22*(246), 1–44 (cited on page 145).

Kaufmann, E., Ménard, P., Domingues, O. D., Jonsson, A., Leurent, E., & Valko, M. (2021). Adaptive reward-free exploration. *Algorithmic Learning Theory (ALT)* (cited on pages 19, 21, 76, 91).

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online controlled experiments at large scale. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1168–1176 (cited on page 9).

Kozuno, T., Yang, W., Vieillard, N., Kitamura, T., Tang, Y., Mei, J., M'enard, P., Azar, M. G., Vaĺko, M., Munos, R., Pietquin, O., Geist, M., & Szepesvari, C. (2022). Kl-entropy-regularized rl with a generative model is minimax optimal. *ArXiv*, *abs/2205.14211* (cited on page 19).

Krause, A., & Golovin, D. (2014). Submodular function maximization. *Tractability*, *3*, 71–104 (cited on page 44).

Lai, T., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, *6*(1), 4–22 (cited on page 8).

Laskin, M., Yarats, D., Liu, H., Lee, K., Zhan, A., Lu, K., Cang, C., Pinto, L., & Abbeel, P. (2021). Urlb: Unsupervised reinforcement learning benchmark. *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (cited on page 76).

Li, G., Wei, Y., Chi, Y., Gu, Y., & Chen, Y. (2020). Breaking the sample size barrier in model-based reinforcement learning with a generative model (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin, Editors). *Advances in Neural Information Processing Systems*, *33*, 12861–12872 (cited on page 19).

Littlestone, N., & Warmuth, M. K. (1994). The weighted majority algorithm. *Information and computation*, *108*(2), 212–261 (cited on page 84).

Locatelli, A., Gutzeit, M., & Carpentier, A. (2016). An optimal algorithm for the thresholding bandit problem (M. F. Balcan & K. Q. Weinberger, Editors). *Proceedings of The 33rd International Conference on Machine Learning*, *48*, 1690–1698 (cited on page 16).

Magureanu, S., Combes, R., & Proutiere, A. (2014). Lipschitz bandits: Regret lower bound and optimal algorithms (M. F. Balcan, V. Feldman, & C. Szepesvári, Editors). *Proceedings of The 27th Conference on Learning Theory*, *35*, 975–999 (cited on pages 38, 145).

Mason, B., Jain, L., Tripathy, A., & Nowak, R. (2020). Finding all \epsilon-good arms in stochastic bandits (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin, Editors). *Advances in Neural Information Processing Systems*, *33*, 20707–20718 (cited on pages 16, 32, 35, 146).

Ménard, P. (2019). Gradient ascent for active exploration in bandit problems. *arXiv e-prints*arXiv 1905.08165 (cited on page 151).

Ménard, P., Domingues, O. D., Jonsson, A., Kaufmann, E., Leurent, E., & Valko, M. (2021). Fast active learning for pure exploration in reinforcement learning. *International Conference on Machine Learning (ICML)* (cited on pages 19, 21, 76, 85).

Modi, A., Chen, J., Krishnamurthy, A., Jiang, N., & Agarwal, A. (2021). Model-free representation learning and exploration in low-rank mdps. *arXiv preprint arXiv:2102.07035* (cited on page 77).

Munos, R. (2003). Error bounds for approximate policy iteration. *International Conference on Machine Learning (ICML)* (cited on page 76).

Munos, R., & Szepesvári, C. (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, *9*(5) (cited on page 76).

Mutti, M., De Santi, R., & Restelli, M. (2022). The importance of non-markovianity in maximum state entropy exploration. *International Conference on Machine Learning*, 16223–16239 (cited on page 76).

Ortner, R. (2010). Online regret bounds for markov decision processes with deterministic transitions [Algorithmic Learning Theory (ALT 2008)]. *Theoretical Computer Science*, *411*(29), 2684–2695 (cited on page 39).

Papini, M., Tirinzoni, A., Restelli, M., Lazaric, A., & Pirotta, M. (2021). Leveraging good representations in linear contextual bandits. *International Conference on Machine Learning*, 8371–8380 (cited on page 86).

Pesquerel, F., & Maillard, O.-A. (2022). Imed-rl: Regret optimal learning of ergodic markov decision processes (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh, Editors). *Advances in Neural Information Processing Systems*, *35*, 26363–26374 (cited on page 54).

Réda, C., Tirinzoni, A., & Degenne, R. (2021). Dealing with misspecification in fixed-confidence linear top-m identification. *Advances in Neural Information Processing Systems (NeurIPS)*, *34* (cited on page 102).

Rosenberg, A., & Mansour, Y. (2019). Online convex optimization in adversarial markov decision processes. *International Conference on Machine Learning* (cited on page 85).

Russac, Y., Katsimerou, C., Bohle, D., Cappé, O., Garivier, A., & Koolen, W. M. (2021). A/b/n testing with control in the presence of subpopulations (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan, Editors). *Advances in Neural Information Processing Systems*, *34*, 25100–25110 (cited on page 9).

Sidford, A., Wang, M., Wu, X., Yang, L., & Ye, Y. (2018a). Near-optimal time and sample complexities for solving markov decision processes with a generative model (S.

Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett, Editors). *Advances in Neural Information Processing Systems*, *31* (cited on page 19).

Sidford, A., Wang, M., Wu, X., & Ye, Y. (2018b). Variance reduced value iteration and faster algorithms for solving markov decision processes. *Proceedings of the 2018 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 770–787 (cited on page 19).

Simchowitz, M., Jamieson, K., & Recht, B. (2017). The simulator: Understanding adaptive sampling in the moderate-confidence regime (S. Kale & O. Shamir, Editors). *Proceedings of the 2017 Conference on Learning Theory*, *65*, 1794–1834 (cited on pages 29, 30).

Sion, M. (1958). On general minimax theorems. *Pacific Journal of Mathematics*, *8*, 171–176 (cited on page 134).

Soare, M., Lazaric, A., & Munos, R. (2014). Best-arm identification in linear bandits (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger, Editors). *Advances in Neural Information Processing Systems*, *27* (cited on page 38).

Strehl, A. L., & Littman, M. L. (2008). An analysis of model-based interval estimation for markov decision processes [Learning Theory 2005]. *Journal of Computer and System Sciences*, *74*(8), 1309–1331 (cited on page 60).

Tarbouriech, J., & Lazaric, A. (2019). Active exploration in markov decision processes (K. Chaudhuri & M. Sugiyama, Editors). *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, *89*, 974–982 (cited on page 54).

Tarbouriech, J., Pirotta, M., Valko, M., & Lazaric, A. (2021). A provably efficient sample collection strategy for reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, *34*, 7611–7624 (cited on pages 76, 81).

Tarbouriech, J., Shekhar, S., Pirotta, M., Ghavamzadeh, M., & Lazaric, A. (2020). Active model estimation in markov decision processes. *Conference on Uncertainty in Artificial Intelligence*, 1019–1028 (cited on page 78).

Taupin, J., Jedra, Y., & Proutière, A. (2022). Best policy identification in linear mdps. *ArXiv*, *abs/2208.05633* (cited on page 115).

Tirinzoni, A., Al-Marjani, A., & Kaufmann, E. (2022). Near instance-optimal PAC reinforcement learning for deterministic MDPs. *Advances in Neural Information Processing Systems (NeurIPS)* (cited on pages 44, 81, 112, 115, 116).

Tirinzoni, A., Pirotta, M., & Lazaric, A. (2021). A fully problem-dependent regret lower bound for finite-horizon mdps. *ArXiv*, *abs/2106.13013* (cited on page 113).

Tirinzoni, A., Pirotta, M., Restelli, M., & Lazaric, A. (2020). An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin, Editors). *Advances in Neural Information Processing Systems*, *33*, 1417–1427 (cited on page 28).

Wagenmaker, A., & Jamieson, K. (2022). Instance-dependent near-optimal policy identification in linear mdps via online experiment design. *Advances in Neural Information Processing Systems (NeurIPS)* (cited on pages 48, 49, 115, 116, 119, 141).

Wagenmaker, A., Simchowitz, M., & Jamieson, K. G. (2022a). Beyond no regret: Instance-dependent PAC reinforcement learning. *Conference On Learning Theory (COLT)* (cited on pages 20, 81, 89, 90, 107, 115, 119, 120).

Wagenmaker, A. J., Chen, Y., Simchowitz, M., Du, S., & Jamieson, K. (2022b). Reward-free RL is no harder than reward-aware RL in linear Markov decision processes. *Proceedings of the 39th International Conference on Machine Learning (ICML)* (cited on page 115).

Wang, P.-A., Tzeng, R.-C., & Proutiere, A. (2021). Fast pure exploration via frank-wolfe (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan, Editors). *Ad-*

vances in Neural Information Processing Systems, *34*, 5810–5821 (cited on pages 24, 29, 151).

Wu, J., Braverman, V., & Yang, L. (2022). Gap-dependent unsupervised exploration for reinforcement learning (G. Camps-Valls, F. J. R. Ruiz, & I. Valera, Editors). *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, *151*, 4109–4131 (cited on page 21).

Xie, T., Foster, D. J., Bai, Y., Jiang, N., & Kakade, S. M. (2022). The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157* (cited on pages 76, 82).

Xie, T., & Jiang, N. (2020). Q* approximation schemes for batch reinforcement learning: A theoretical comparison. *Conference on Uncertainty in Artificial Intelligence*, 550–559 (cited on page 76).

Xie, T., & Jiang, N. (2021). Batch value-function approximation with only realizability. *International Conference on Machine Learning*, 11404–11413 (cited on page 76).

Yarats, D., Fergus, R., Lazaric, A., & Pinto, L. (2021). Reinforcement learning with prototypical representations. *International Conference on Machine Learning*, 11920–11931 (cited on page 76).

Zahavy, T., O'Donoghue, B., Desjardins, G., & Singh, S. (2021). Reward is enough for convex mdps. *Neural Information Processing Systems (NeurIPS)* (cited on page 76).

Zanette, A., & Brunskill, E. (2019a). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *Proceedings of the 36th International Conference on Machine Learning, (ICML)* (cited on page 89).

Zanette, A., & Brunskill, E. (2019b). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *ICML*, *97*, 7304–7312 (cited on page 84).

Zanette, A., Kochenderfer, M. J., & Brunskill, E. (2019). Almost horizon-free structure-aware best policy identification with a generative model. *Advances in Neural Information Processing Systems (NeurIPS)*, 5626–5635 (cited on pages 20, 115).

Zanette, A., Lazaric, A., Kochenderfer, M. J., & Brunskill, E. (2020). Provably efficient reward-agnostic navigation with linear value iteration. *Advances in Neural Information Processing Systems (NeurIPS)*, *33*, 11756–11766 (cited on page 77).

Zhang, Z., Du, S., & Ji, X. (2021a). Near optimal reward-free reinforcement learning. *International Conference on Machine Learning, (ICML)* (cited on page 89).

Zhang, Z., Du, S., & Ji, X. (2021b). Near optimal reward-free reinforcement learning (M. Meila & T. Zhang, Editors). *Proceedings of the 38th International Conference on Machine Learning*, *139*, 12402–12412 (cited on page 21).

Zhang, Z., Du, S., & Ji, X. (2021c). Near optimal reward-free reinforcement learning. *International Conference on Machine Learning*, 12402–12412 (cited on page 76).

Zhang, Z., Ji, X., & Du, S. (2021d). Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *Conference on Learning Theory*, 4528–4531 (cited on page 84).

Zheng, S., Trott, A. R., Srinivasa, S., Naik, N. V., Gruesbeck, M., Parkes, D. C., & Socher, R. (2020). The ai economist: Improving equality and productivity with ai-driven tax policies. *ArXiv*, *abs/2004.13332* (cited on page 9).

Books

Aliprantis, C. D., & Border, K. C. (2006). *Infinite dimensional analysis: A hitchhiker's guide*. Berlin; London, Springer. (Cited on page 62).

Lattimore, T., & Szepesvari, C. (2019). *Bandit Algorithms*. Cambridge University Press. (Cited on pages 13, 30, 102, 103).

Levin, D. A., Peres, Y., & Wilmer, E. L. (2006). *Markov chains and mixing times*. American Mathematical Society. (Cited on pages 29, 54, 72).

Puterman, M. (1994). *Markov Decision Processes. Discrete Stochastic. Dynamic Programming*. Wiley. (Cited on pages 8, 10, 11, 13, 28, 49, 77, 82).

Sterelny, K. (2007). *21814 Cognitive Load and Human Decision, or, Three Ways of Rolling the Rock Uphil*. Oxford University Press. (Cited on page 8).

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. Cambridge, MA, USA, A Bradford Book. (Cited on pages 8, 76).

Tsybakov, A. (2008). *Introduction to nonparametric estimation*. Springer Series in Statistics. (Cited on page 18).

List of Figures

List of Algorithms