Sample-Efficient Reinforcement Learning : Exploration, Imitation, and Online Learning

Daniil Tiapkin

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2025IPPAX069

Thèse de doctorat

Sample-Efficient Reinforcement Learning: Exploration, Imitation, and Online Learning

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques et Informatique

Thèse présentée et soutenue à Palaiseau, le 16 Septembre 2025, par

DANIIL TIAPKIN

Composition du Jury :

Erwan Le Pennec Professeur, École polytechnique	Président du jury
Émilie Kaufmann Chargée de recherche (HdR), CNRS, Université de Lille (CRIStAL)	Rapporteure
Shie Mannor Professeur, Technion et Nvidia	Rapporteur
Claire Vernade Professeur assistant, Université de Tübingen	Examinatrice
Aurélien Garivier Professeur, Ecole Normale Supérieure de Lyon	Examinateur
Éric Moulines Professeur, École polytechnique et MBZUAI	Directeur de thèse
Gilles Stoltz Directeur de recherche, CNRS, Université Paris-Saclay	Co-directeur de thèse

Sample-Efficient Reinforcement Learning: Exploration, Imitation, and Online Learning

Apprentissage par renforcement efficace en taille d'échantillon : exploration, imitation et apprentissage séquentiel

Daniil Tiapkin

Centre de Mathématiques Appliquées – CNRS – École polytechnique – Institut Polytechnique de Paris
Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay

2025

This work has been supported by the Paris Île-de-France Région in the framework of the DIM AI4IDF. Additionally, this work was funded by the European Union (ERC-2022-SYG-OCEAN-101071601). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

To my mum for her infinite support

Acknowledgments

First of all, I would like to express my gratitude to my advisors, Éric and Gilles, for guiding me through this fantastic journey. My path first crossed with Éric during my undergraduate studies in Moscow, and later he introduced me to Gilles. Gilles likes to recall that the story of my PhD began with a small misunderstanding at the doctoral school, when my application was briefly mistaken for that of a postdoc candidate. In any case, I am truly grateful to consider both of them not only as mentors but also as friends.

I would like to thank Émilie and Shie for accepting to review this thesis and for their genuine interest in my work. I am also very grateful to Claire and Aurélien for serving as jury members, and to Erwan for chairing the jury and making the defense possible.

During my (relatively short) PhD, I was lucky to meet many collaborators and, more importantly, friends. First, I want to thank Michal and Pierre for our long-standing collaboration since 2021. I first learned reinforcement learning from them, and together we produced a lot of research that laid the foundation for this thesis. I am very happy that this collaboration continues to grow and has introduced me to many great scientists and friends such as Denis, Rémi, Kashif, Daniele, Yunhao, Mark, and Michal.

I am also grateful to Evgenii and Sholom for our joint work (we really should write something all together at some point), part of which contributed directly to this thesis, and for helping me adjust to life in Paris when I first arrived.

I would also like to thank our amortized sampling / GFlowNet group, which I was fortunate to help create during the summer before I moved to Paris. It has been a pleasure to work with Sergey, Alexey, Nikita, Ian, Askar, Artem, Timofei, Dmitry, Nikolay, and Kirill, and I hope the group will keep growing and continue producing great research.

I am grateful to Maxim, Martin, and Fakhri for inviting me for a research visit in Abu Dhabi, and to Michael and Mathieu for our brief but enjoyable collaborations.

This is also the only section of my thesis where I can officially thank Mathieu, Alexandre, Sarah, Nino, Johan, and Daniele (for the second time) for an amazing internship at Google DeepMind, and Lucas, Mathieu, and Joëlle for the opportunity to return.

I would like to thank everyone I met at CMAP and LMO (despite my rare presence on site) and within the OCEAN project. I am grateful for the discussions, for both successful and less successful collaborations, and for the overall research atmosphere. I also want to thank everyone involved in the memorable OCEAN retreat in Venice, as well as for many great conferences and meetings all around the world. I especially want to thank the PhD and former PhD students (Yazid, Badr, Navid, Safwan, Lorenzo, Liza, Antoine, Aymeric, Pierre, Adrien, Mahmoud (for working with me during some intense weekends), Sacha, Etienne, Eugène), the (former) postdocs (Yazid, Paul, Antonio, Andrea) and the professors (Aymeric, Alain, Rémi, Vianney, El Mahdi, Etienne, and Mike) ith whom I had the pleasure of discussing, collaborating, and exchanging ideas.

I am deeply grateful to all my friends for their support throughout these years. There are too many to name here, and many have already been mentioned. With many of you, I could only meet during conferences, short visits, or through calls, but each of you has made this journey much more memorable. I am also grateful to my family, especially my mum, for their support and encouragement over all my life. And, last but not least, to Kseniia, for being with me all along, regardless of the actual distance.

Abstract

Keywords: reinforcement learning, exploration, imitation learning, adversarial MDPs, reinforcement learning from human feedback

Reinforcement learning (RL) provides a formal framework for sequential decision-making in an unknown environment. An agent learns a policy to maximize a cumulative reward signal by interacting with a Markov Decision Process (MDP). This thesis investigates three extensions to the standard RL setting, providing new algorithms and associated theoretical analyses. First, we study the problem of exploration in the absence of a reward signal. We consider two distinct formulations of maximum entropy exploration and propose computationally efficient algorithms. For these algorithms, we derive sample complexity bounds that improve upon previously known results. Our analysis further establishes that learning the maximum visitation entropy policy is statistically less demanding than both a common variant of reward-free exploration and the task of learning an optimal policy in a standard rewarded MDP. Second, we address the problem of accelerating reinforcement learning using expert demonstrations, a setting motivated by modern applications such as Reinforcement Learning from Human Feedback (RLHF). We analyze the framework of demonstration-regularized RL, where the agent's learning objective is augmented with a Kullback-Leibler penalty for deviating from a reference policy learned from the demonstrations. We prove that this approach leads to a linear reduction in sample complexity with respect to the number of demonstrations. This result establishes how expert data can improve the efficiency of the subsequent RL task. Third, we address the problem of learning in MDPs with rewards selected by an (oblivious) adversary. We propose a new algorithm that uses a black-box online linear optimization oracle, making it easy to implement. We provide a regret analysis showing that its performance bound matches the minimax lower bound for the stochastic (non-adversarial) case in its dependence on the number of states, actions, and the number of episodes. Collectively, the contributions of this thesis consist of novel algorithms for various problems in reinforcement learning, accompanied by proofs of their performance in terms of sample complexity and regret bounds.

Résumé

Mots-clés : apprentissage par renforcement, exploration, apprentissage par imitation, MDPs adversaires, apprentissage par renforcement à partir du retour humain

L'apprentissage par renforcement (RL) fournit un cadre formel pour la prise de décision séquentielle en environnement inconnu. Un agent apprend une politique afin de maximiser un signal de récompense cumulée en interagissant avec un processus de décision markovien (MDP). Cette thèse étudie trois extensions du cadre standard de l'apprentissage par renforcement, en proposant de nouveaux algorithmes ainsi que leurs analyses théoriques associées. Premièrement, nous étudions le problème de l'exploration en l'absence de signal de récompense. Nous considérons deux formulations distinctes de l'exploration par maximisation d'entropie et proposons des algorithmes computationnellement efficaces. Pour ces algorithmes, nous dérivons des bornes de complexité en tailles d'échantillons, qui améliorent les résultats précédemment connus. Notre analyse établit en outre que l'apprentissage de la politique de maximisation de l'entropie de visite est statistiquement moins exigeant que l'exploration sans récompense dans l'une de ses variantes usuelles, ainsi que l'apprentissage d'une politique optimale dans un PDM standard avec récompenses. Deuxièmement, nous abordons le problème de l'accélération de l'apprentissage par renforcement à l'aide de démonstrations expertes, un cadre motivé par des applications modernes telles que l'apprentissage par renforcement à partir de retours (*feedback*) humains (RLHF). Nous analysons le cadre de l'apprentissage par renforcement régularisé par des démonstrations, où la fonction objectif de l'agent est augmentée d'une pénalité de Kullback-Leibler pour toute déviation par rapport à une politique de référence apprise à partir de démonstrations. Nous prouvons que cette approche entraîne une réduction linéaire de la complexité en tailles d'échantillon par rapport au nombre de démonstrations. Ce résultat établit comment les données expertes peuvent améliorer l'efficacité de la tâche d'apprentissage par renforcement. Troisièmement, nous étudions le problème de l'apprentissage dans des MDP où les récompenses sont choisies par un adversaire oublieux (*oblivious*). Nous proposons un nouvel algorithme qui utilise un oracle boîte noire d'optimisation linéaire séquentielle, le rendant facile à mettre en œuvre. Nous fournissons une analyse de regret montrant que sa borne de performance correspond à la borne inférieure minimax du cas stochastique (non adverse) en termes de dépendance aux nombres d'états, d'actions et d'épisodes. Globalement, les contributions de cette thèse consistent en de nouveaux algorithmes pour divers problèmes en apprentissage par renforcement, accompagnés de preuves de leur performance en termes de complexité en tailles d'échantillons et de bornes de regret.

Notations

General notations.

- $\mathbb{R}_+$ is the set of non-negative real numbers;
- $\mathbb{N}_+$ is the set of positive natural numbers;
- $[n]$ is the set $\{1, 2, \dots, n\}$ for $n \in \mathbb{N}_+$;
- e is Euler's number;
- Δ_d is the $(d-1)$-dimensional probability simplex: $\Delta_d \triangleq \{x \in \mathbb{R}_+^d : \sum_{j=1}^d x_j = 1\}$;
- $\Delta_{\mathcal{X}}$ or $\mathcal{P}(\mathsf{X})$ is the set of distributions over a finite set $\mathcal{X}$: $\Delta_{\mathcal{X}} = \Delta_{|\mathcal{X}|}$;
- $\mathrm{clip}(x, m, M)$ is the clipping procedure: $\mathrm{clip}(x, m, M) \triangleq \max(\min(x, M), m)$ for $m < M$;
- $\mathcal{H}(p)$ is the Shannon entropy for $p \in \Delta_{\mathcal{X}}$: $\mathcal{H}(p) \triangleq -\sum_{i \in \mathcal{X}} p_i \log p_i$.

Probability theory notations. Let $(\mathsf{X}, \mathcal{X})$ be a measurable space and $\mathcal{P}(\mathsf{X})$ be the set of all probability measures on this space. For $p \in \mathcal{P}(\mathsf{X})$ we denote by $\mathbb{E}_p$ the expectation with respect to p. For a random variable $\xi : \mathsf{X} \to \mathbb{R}$, the notation $\xi \sim p$ means $\mathrm{Law}(\xi) = p$. For any measures $p, q \in \mathcal{P}(\mathsf{X})$ we denote their product measure by $p \otimes q$. We also write $\mathbb{E}_{\xi \sim p}$ instead of $\mathbb{E}_p$.

For any $p, q \in \mathcal{P}(\mathsf{X})$, the Kullback-Leibler divergence $\mathrm{KL}(p, q)$ is given by

$$\mathrm{KL}(p, q) = \begin{cases} \mathbb{E}_p\left[\log \frac{\mathrm{d}p}{\mathrm{d}q}\right], & p \ll q \\ +\infty, & \text{otherwise} \end{cases}$$

For any $p \in \mathcal{P}(\mathsf{X})$ and $f : \mathsf{X} \to \mathbb{R}$, we define $pf = \mathbb{E}_p[f]$. In particular, for any $p \in \Delta_d$ and $f : \{0, \dots, d\} \to \mathbb{R}$, we have $pf = \sum_{\ell=0}^d f(\ell) p(\ell)$.

We define $\mathrm{Var}_p(f) = \mathbb{E}_{s' \sim p}\left[(f(s') - pf)^2\right] = p[f^2] - (pf)^2$.

Matrix norms and inner products. For any symmetric positive definite matrix A, we define the corresponding A-scalar product and A-norm as follows:

$$\langle x, y \rangle_A = \langle x, Ay \rangle, \qquad \|x\|_A = \sqrt{\langle x, x \rangle_A}$$

Note that if $\|A\|_2 \leqslant c$, then $\|x\|_A \leqslant \sqrt{c}\|x\|_2$.

Résumé substantiel en français

Cette thèse porte sur l'apprentissage par renforcement, un cadre théorique pour la prise de décision séquentielle en environnement inconnu. Dans ce contexte, un agent interagit avec un processus de décision markovien afin d'apprendre une politique qui maximise une fonction de récompense cumulée. Le travail s'intéresse à des variantes du cadre standard pour lesquelles les garanties classiques d'efficacité statistique et computationnelle doivent être repensées. Trois directions principales sont explorées : (i) l'exploration en l'absence de signal de récompense, (ii) l'apprentissage à partir de démonstrations expertes, et (iii) l'apprentissage dans des environnements où les récompenses sont adverses. Pour chacune de ces extensions, la thèse propose de nouveaux algorithmes accompagnés d'analyses théoriques de leur performance en termes de complexité en taille d'échantillon ou de regret.

Première partie — Exploration en l'absence de récompense. La première partie étudie le problème de l'exploration pure, où un agent doit découvrir un environnement inconnu sans disposer d'un signal de récompense extrinsèque. L'objectif n'est plus de maximiser une fonction de récompense donnée, mais de couvrir efficacement l'espace des états et des actions. Deux formulations fondées sur la maximisation d'entropie sont analysées.

La première repose sur la *maximisation de l'entropie trajectorielle*. Ce critère est montré équivalent à un problème d'apprentissage par renforcement régularisé par l'entropie, ce qui permet l'utilisation de méthodes de type politique régularisée. Un algorithme dédié, noté `RFExploreEnt`, est introduit. Son analyse établit une borne de complexité en taille d'échantillon de l'ordre de $O(\mathrm{poly}(H, S, A)/\varepsilon)$ pour obtenir une politique ε-optimale, où H désigne l'horizon temporel du processus, S le nombre d'états et A le nombre d'actions possibles. Cette borne améliore la dépendance standard en $O(1/\varepsilon^2)$ issue des approches directes.

La seconde formulation concerne la *maximisation de l'entropie de visite*, qui favorise la couverture équilibrée des états et actions visités. Le problème correspondant est non-concave dans l'espace des politiques, mais il peut être reformulé comme une maximisation concave sur le polytope convexe des mesures d'occupation états-actions. Un nouvel algorithme inspiré de la théorie des jeux, `EntGame`, est proposé. Il interprète la maximisation de l'entropie de visite comme un jeu à somme nulle entre un prédicteur et un échantillonneur récompensé pour son imprévisibilité. Cette reformulation rend le calcul plus efficace, car elle évite la résolution complète d'un sous-problème d'apprentissage par renforcement à chaque itération, étape coûteuse des approches précédentes. Une version régularisée, `RegEntGame`, est également analysée. Les résultats obtenus montrent une complexité en taille d'échantillon en $O(H^2SA/\varepsilon^2)$, améliorant la dépendance antérieure en $O(1/\varepsilon^3)$. L'analyse établit en outre que l'apprentissage d'une politique de maximisation de l'entropie de visite est, du point de vue statistique, moins exigeant que l'exploration sans récompense classique ou que la recherche d'une politique optimale dans un processus de décision markovien standard avec récompenses.

Deuxième partie — Apprentissage régularisé par des démonstrations. La deuxième partie s'intéresse à la manière dont l'apprentissage par renforcement peut être accéléré grâce à des démonstrations expertes obtenues préalablement. Ce cadre est motivé par des applications récentes, notamment l'apprentissage par renforcement à partir de retours humains, mais il est ici étudié sous un angle théorique.

Le travail introduit un *cadre d'apprentissage par renforcement régularisé par les démonstrations*,

dans lequel la fonction objectif de l'agent comporte un terme de pénalisation de Kullback-Leibler pour toute déviation par rapport à une politique de référence apprise à partir de données expertes. La méthodologie comprend deux étapes : (i) apprentissage d'une politique initiale par *clonage comportemental* à partir du jeu de démonstrations; (ii) optimisation d'une politique séquentielle, régularisée par rapport à cette politique de référence.

L'analyse théorique montre que ce schéma réduit la complexité en taille d'échantillon proportionnellement au nombre de démonstrations N^{E}, atteignant un taux $O(H^6S^3A^2/(\varepsilon^2 N^{\mathrm{E}}))$. Ce résultat formalise comment les démonstrations peuvent contribuer à une accélération linéaire de l'apprentissage. La preuve repose sur deux éléments principaux : une analyse optimale au sens minimax du clonage comportemental en divergence de Kullback–Leibler sur les trajectoires, et un nouvel algorithme d'exploration régularisée (`UCBVIEnt+`) pour l'estimation de la meilleure politique régularisée. Le cadre est ensuite étendu au cas où la récompense n'est pas directement disponible mais apprise à partir de préférences, comme dans les approches d'apprentissage par renforcement à partir de retours humains. L'analyse montre que, sous des conditions de taille suffisante des ensembles de démonstrations et de préférences, le schéma standard en trois étapes – apprentissage supervisé initial, modélisation de la récompense, puis apprentissage par renforcement régularisé – possède des garanties de convergence et permet de limiter les dérives du signal de récompense.

Troisième partie — Apprentissage dans des environnements adverses. La troisième partie traite du problème de l'apprentissage dans des processus de décision markoviens où les fonctions de récompense sont choisies de manière adverse, mais indépendamment des actions présentes et passées (adversaire dit oublieux, "*oblivious*" en anglais). Le but est de concevoir un algorithme dont la borne de regret corresponde, en ordre de grandeur, à celle obtenue dans le cadre stochastique classique.

L'algorithme proposé, appelé `APO-MVP`, utilise un *oracle d'optimisation linéaire séquentielle* pour chaque état, chargé de gérer la nature adverse des récompenses à l'aide d'estimations d'avantage optimistes. Afin de traiter la non-stationnarité du modèle de transition inconnu, une technique de doublement des époques est employée pour geler périodiquement les estimations du modèle.

L'analyse repose sur une nouvelle décomposition du regret qui sépare les composantes adverses, contrôlées par les garanties de l'oracle d'optimisation linéaire séquentielle, et les composantes liées à l'apprentissage du modèle. Cette séparation permet d'éviter les dépendances récursives qui avaient conduit à des bornes sous-optimales dans les travaux antérieurs. Il en résulte une borne de regret de $O(\sqrt{H^7SAT})$, correspondant en dépendance principale à la borne inférieure minimax $O(\sqrt{\mathrm{poly}(H)SAT})$ du cas stochastique, où H désigne l'horizon temporel, S le nombre d'états, A le nombre d'actions et T le nombre total d'épisodes d'apprentissage.

Conclusion. Dans l'ensemble, cette thèse propose une étude unifiée de plusieurs extensions du cadre standard de l'apprentissage par renforcement, couvrant des problématiques d'exploration, d'utilisation de données expertes et de robustesse face à l'adversité. Les contributions consistent en la conception d'algorithmes efficaces en calcul et en l'établissement de garanties précises en taille d'échantillon et en regret. Ces résultats apportent une meilleure compréhension théorique des limites statistiques et computationnelles de l'apprentissage par renforcement dans des environnements complexes, au-delà de la simple maximisation d'un signal de récompense connu.

Contents

Chapter 1

Introduction

Contents

Chapter roadmap. This chapter has two roles: (i) it gathers the background on classical reinforcement learning needed for the rest of the thesis and (ii) it previews the three research directions developed in the subsequent chapters.

- **Section 1.1** (Reinforcement Learning Problem) introduces finite-horizon Markov decision processes, standard learning objectives, their minimax lower bounds, and the UCBVI algorithm that nearly matches those bounds.
- **Section 1.2** summarises Chapter 2 and explains how to find a policy that maximizes certain notions of entropy.
- **Section 1.3** previews Chapter 3 where expert demonstrations are incorporated into reinforcement learning via KL-regularised learning.

- **Section 1.4** previews Chapter 4 on reinforcement learning with adversarial rewards.
- **Section 1.5** briefly lists related works not covered in this thesis.

Readers already comfortable with the material in Section 1.1 may skip ahead to the summary sections.

1.1 Reinforcement Learning Problem

1.1.1 Overview

Reinforcement learning (RL) is a paradigm of machine learning centered on the *reward hypothesis*: intelligent behaviour can be learned by trial-and-error interaction that maximises a scalar reward (Silver et al. 2021). Deep RL has confirmed this hypothesis in practice. It first reached super-human play on Atari games and Go (Mnih et al. 2015; Silver et al. 2016), mastered Chess and StarCraft (Silver et al. 2018; Vinyals et al. 2019). All these domains share a clear, fixed reward signal.

However, many real-world applications present additional challenges. In many settings the reward is sparse, supplemented by demonstrations, or even non-stationary. This thesis focuses on three such challenges:

1. **Pure exploration.** Many tasks provide little or no reward — examples include automatic website testing (Zheng et al. 2021) and exploratory pretraining for downstream control tasks (Seo et al. 2021; Zhang et al. 2021a; Mutti et al. 2022). In these settings the agent must explore efficiently rather than maximise immediate returns. We investigate reward-free exploration in Chapter 2 and preview it in Section 1.2.

2. **Learning from demonstrations.** Robotics (Tang et al. 2025) and reinforcement learning from human feedback (RLHF) for language model fine-tuning (Ziegler et al. 2020; Ouyang et al. 2022) additionally provide large datasets of expert behaviour. The challenge is to blend these demonstrations with reward signals so the agent learns faster than with either RL or imitation alone. We tackle this in Chapter 3 and outline the idea in Section 1.3.

3. **Non-stationary rewards.** In some applications, rewards may change over time. For example, this occurs when one trains the destruction process in molecule generation (Bengio et al. 2021; Malkin et al. 2022; Tiapkin et al. 2024b) or when human preferences shift in RLHF (Pettigrew 2019). Algorithms must adapt online, even in the face of adversarial changes. We address this challenge in Chapter 4 and provide an overview in Section 1.4.

These three problems highlight the need to extend classical RL theory to new settings. This thesis develops theoretical and algorithmic frameworks for each, aiming to advance current reinforcement learning methods. The subsequent chapters introduce these challenges in detail.

1.1.2 Markov Decision Process

To formalize the reinforcement learning problem, we define a so-called Markov decision process (MDP). In this thesis, we focus on *finite-horizon heterogeneous MDPs* that are defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \boldsymbol{P}, H, \boldsymbol{r}, s_1)$ that consists of:

- a finite set of states $\mathcal{S}$ of size S;
- a finite set of actions $\mathcal{A}$ of size A;
- a planning horizon $H \in \mathbb{N}$;
- a collection of transition kernels $\boldsymbol{P} \triangleq (P_h)_{h=1}^{H}$, where each $P_h \colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ specifies the probability of transitioning to the next state at step h;

- a collection of reward functions $\boldsymbol{r} \triangleq (r_h)_{h=1}^H$, in which each $r_h \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$;
- an initial state $s_1 \in \mathcal{S}$.

For simplicity, we assume that rewards are deterministic and bounded within $[0, 1]$. This assumption does not significantly limit generality. In most reinforcement learning problems, the primary difficulty is learning the transition kernel when it is unknown.[1] We also assume a fixed initial state s_1. If the initial state instead comes from a distribution $\rho \in \Delta(\mathcal{S})$, this scenario can be reduced to the fixed-state case. As described by Fiechter (1994), one can introduce an auxiliary state s°, extend the horizon to $H + 1$, and define the initial transition as $P_0(s|s^\circ, a) = \rho(s)$ for any $a \in \mathcal{A}$, assigning zero reward to this auxiliary transition.

Policy. A general (or history-dependent) policy $\boldsymbol{\pi} = (\psi_h)_{h\in[H]}$ is a collection of functions $\psi_h \colon (\mathcal{S} \times \mathcal{A} \times [0, 1])^{h-1} \times \mathcal{S} \times [0, 1] \to \mathcal{A}$ that each maps a history $I_h = (S_1, A_1, U_1, \ldots, S_{h-1}, A_{h-1}, U_{h-1})$ where U_j for $j \leqslant h - 1$ are i.i.d. uniformly distributed on the unit interval, a state S_h, and an auxiliary independent uniformly distributed random variable U_h to an action $A_h = \psi_h(I_h, S_h, U_h)$. The set of all general (or history-dependent) policies is denoted as Π_{H}.

A policy is called *Markovian* if the action at each step depends only on the current state and an independent randomization variable, i.e., $A_h = \psi_h(S_h, U_h)$. In this case, the policy can be equivalently represented as $\boldsymbol{\pi} = (\pi_h)_{h\in[H]}$, where each $\pi_h \colon \mathcal{S} \to \Delta(\mathcal{A})$ maps states to probability distributions over actions, and $A_h \sim \pi_h(S_h)$. Importantly, for any distribution $\pi_h(S_h) \in \Delta(\mathcal{A})$, there exists a measurable function ψ_h such that $A_h = \psi_h(S_h, U_h)$ generates A_h according to $\pi_h(S_h)$, e.g., via the inverse transform sampling method (see also Douc et al. (2018, Theorem 1.3.6) for a more general statement). The set of all Markovian policies is denoted as Π.

A policy $\boldsymbol{\pi} = (\pi_h)_{h\in[H]}$ is called *Markovian deterministic* if, for each step $h \in [H]$, the policy $\pi_h \colon \mathcal{S} \to \mathcal{A}$ deterministically maps each state to a single action. That is, for every $s \in \mathcal{S}$, $\pi_h(s)$ returns a specific action in $\mathcal{A}$ without any randomization. Note that in the finite-horizon setting, each π_h may differ across steps h, in contrast to the stationary policies often considered in infinite-horizon MDPs (see, e.g., Puterman 2014, Section 5).

Interaction protocol. At the beginning of each episode, the agent selects a (possibly history-dependent) policy $\boldsymbol{\pi} = (\psi_h)_{h\in[H]}$. The interaction protocol is as follows:

- The agent starts at the initial state $S_1 = s_1$;
- For each step $h = 1, \ldots, H$:
  - The agent samples an action $A_h = \psi_h(I_h, S_h, U_h)$, where I_h denotes the history at the beginning of step h, and U_h is an independent randomization variable;
  - The environment transitions to the next state $S_{h+1} \sim P_h(\cdot \mid S_h, A_h)$;
  - The agent observes the reward $r_h(S_h, A_h)$.

Trajectory distribution. The interaction protocol induces a probability distribution over the space of trajectories $\mathcal{T} \triangleq (\mathcal{S} \times \mathcal{A})^H$. For a Markovian policy $\boldsymbol{\pi} = (\pi_h)_{h=1}^H$, the probability of observing a trajectory $\tau = (s_1, a_1, s_2, a_2, \ldots, s_H, a_H)$ is

$$q^{\boldsymbol{\pi}}(\tau) \triangleq \pi_1(a_1 \mid s_1) \prod_{h=1}^{H-1} P_h(s_{h+1} \mid s_h, a_h)\, \pi_{h+1}(a_{h+1} \mid s_{h+1})\,. \tag{1.1}$$

Occupancy measure (visitation distribution). For a Markovian policy $\boldsymbol{\pi} = (\pi_h)_{h=1}^H$, the *occupancy measure* or *visitation distribution* at step h is the distribution $d_h^{\boldsymbol{\pi}}$ over state-action pairs,

[1] However, stochastic heavy-tailed rewards introduce additional analytical challenges; see Zhuang and Sui (2021) and Huang et al. (2023).

defined as

$$d_h^{\boldsymbol{\pi}}(s,a) \triangleq \mathbb{P}_{\boldsymbol{\pi},\boldsymbol{P}}[S_h = s,\, A_h = a]\,, \tag{1.2}$$

where the probability is taken over trajectories generated by following $\boldsymbol{\pi}$ in the MDP with transition kernels P, i.e., when trajectories τ follow the distributions q^{π}. A set of occupancy measures for all steps is denoted by $\boldsymbol{d}^{\boldsymbol{\pi}} = (d_h^{\boldsymbol{\pi}})_{h\in[H]}$.

Value and Q-value. Given a general policy $\boldsymbol{\pi}$, the *value function* $V_h^{\boldsymbol{\pi}}(s)$ at step h and state s is defined as the expected cumulative reward obtained by starting from state s at step h and following policy $\boldsymbol{\pi}$ thereafter:

$$V_h^{\boldsymbol{\pi}}(s) = \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{P}}\left[\sum_{h'=h}^{H} r_{h'}(S_{h'}, A_{h'}) \,\middle|\, S_h = s\right]. \tag{1.3}$$

The *Q-value function* $Q_h^{\boldsymbol{\pi}}(s,a)$ is the expected cumulative reward when taking action a in state s at step h, and then following policy $\boldsymbol{\pi}$:

$$Q_h^{\boldsymbol{\pi}}(s,a) = \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{P}}\left[\sum_{h'=h}^{H} r_{h'}(S_{h'}, A_{h'}) \,\middle|\, S_h = s,\, A_h = a\right], \tag{1.4}$$

where the expectation is taken over trajectories $\tau = (s_1, a_1, \ldots, s_H, a_H)$ generated according to the interaction protocol described above.

Optimal values and optimal policy. The *optimal value function* $V_h^\star \colon \mathcal{S} \to \mathbb{R}$ and the *optimal Q-value function* $Q_h^\star \colon \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ at step h are defined as

$$V_h^\star(s) \triangleq \sup_{\boldsymbol{\pi}\in\Pi_{\mathsf{H}}} V_h^{\boldsymbol{\pi}}(s), \qquad Q_h^\star(s,a) \triangleq \sup_{\boldsymbol{\pi}\in\Pi_{\mathsf{H}}} Q_h^{\boldsymbol{\pi}}(s,a). \tag{1.5}$$

A policy $\boldsymbol{\pi}^\star \in \Pi_{\mathsf{H}}$ is called *optimal* if it achieves the supremum, i.e., $V_h^{\boldsymbol{\pi}^\star}(s) = V_h^\star(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$. A classical result (see Bellman 1957, Chapter IV, also see Puterman 2014, Theorem 4.4.2 for a more modern exposition) guarantees the existence of an optimal policy that is both Markovian and deterministic, and, therefore, the class of policies in the definition (1.5) can be replaced with Π. Consequently, unless specified explicitly, we restrict our attention to Markovian policies in the remainder of this thesis.

Bellman equations. The Bellman equations (Howard 1960, Chapter 2) describe the recursive relationship between value and Q-value functions for a Markovian policy $\boldsymbol{\pi} = (\pi_h)_{h=1}^H$. For all $h \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$:

$$Q_h^{\boldsymbol{\pi}}(s,a) = r_h(s,a) + P_h V_{h+1}^{\boldsymbol{\pi}}(s,a), \qquad V_h^{\boldsymbol{\pi}}(s) = \pi_h Q_h^{\boldsymbol{\pi}}(s), \tag{1.6}$$

where $V_{H+1}^{\boldsymbol{\pi}} \triangleq 0$ by convention. Here, $P_h V_{h+1}^{\boldsymbol{\pi}}(s,a) \triangleq \mathbb{E}_{S'\sim P_h(\cdot|s,a)}[V_{h+1}^{\boldsymbol{\pi}}(S')]$ is the expected value at the next state, and $\pi_h Q_h^{\boldsymbol{\pi}}(s) \triangleq \mathbb{E}_{A\sim\pi_h(s)}[Q_h^{\boldsymbol{\pi}}(s,A)]$ is the expected Q-value under the policy at step h. We refer to Puterman (2014, Section 4.2) and Howard (1960) for details.

Bellman optimality equations. The Bellman optimality equations (Bellman 1957, Chapter III) characterize the optimal value and Q-value functions recursively, for all $h \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$:

$$Q_h^\star(s,a) = r_h(s,a) + P_h V_{h+1}^\star(s,a), \qquad V_h^\star(s) = \max_{a\in\mathcal{A}} Q_h^\star(s,a), \tag{1.7}$$

with $V_{H+1}^\star \triangleq 0$. These equations imply that the optimal value at each state is achieved by choosing the action that maximizes the expected sum of immediate reward and the optimal value at the

next step. The existence of a Markovian deterministic optimal policy follows from these equations (see, e.g., Puterman 2014, Theorem 4.4.2, for a proof), since such a policy can be described as $\pi_h^\star(s) = \arg\max_{a\in\mathcal{A}} Q_h^\star(s,a)$, where ties in the arg max definition are resolved arbitrary, i.e., $\pi_h^\star(s)$ is equal to a some action from the set $\arg\max_{a\in\mathcal{A}} Q_h^\star(s,a)$.

1.1.3 Learning Objectives

When both the transition kernel $\boldsymbol{P}$ and reward function $\boldsymbol{r}$ are known, the RL problem reduces to solving the Bellman optimality equations (1.7) and extracting a Markovian deterministic optimal policy. In practice, however, these quantities are unknown and must be learned through interaction with the environment. In this subsection, we summarize the three most common learning objectives in RL.

Learning algorithms and objectives. This leads to the standard reinforcement learning objectives, which can be broadly grouped as follows.

- *Regret minimization (Lai and Robbins 1985; Burnetas and Katehakis 1997).* The learner seeks to minimize the cumulative regret over T episodes, defined as the difference between the total expected reward of the best policy in hindsight and that of the learner's sequence of policies:

$$\mathfrak{R}(T) \triangleq \sum_{t=1}^{T} (V_1^\star(s_1) - V_1^{\boldsymbol{\pi}_t}(s_1)),$$

  where $\boldsymbol{\pi}_t$ denotes the policy executed in episode t. Under this criterion, an RL algorithm is formalized as a sequence of random policies $(\boldsymbol{\pi}_t)_{t\in\mathbb{N}}$, where each $\boldsymbol{\pi}_t$ depends on the history up to episode $t-1$. The learner updates its policy based on observed trajectories and rewards, aiming to minimize regret over time. An algorithm is called a *regret minimizer* if, with high probability, $\mathfrak{R}(T) = o(T)$ as $T \to \infty$.

- *Best policy identification (BPI, Fiechter 1994).* The learner's goal is to output a policy $\widehat{\boldsymbol{\pi}}$ such that, with probability at least $1-\delta$, its value is within ε of optimal:

$$V_1^\star(s_1) - V_1^{\widehat{\boldsymbol{\pi}}}(s_1) \leqslant \varepsilon.$$

  The focus here is on minimizing the sample complexity, i.e., the number of episodes required before the learner can confidently output such a policy. Formally, an RL algorithm for this objective is specified by a sequence of exploration policies $(\boldsymbol{\pi}_t)_{t\in\mathbb{N}}$, a stopping time $\mathfrak{t}$ (possibly random and adapted to the observed data), and an output policy $\widehat{\boldsymbol{\pi}}$, which may depend on the entire interaction history up to $\mathfrak{t}$. Importantly, $\widehat{\boldsymbol{\pi}}$ need not coincide with the last policy executed, and can be constructed using all collected data.

- *Provably Approximately Correct (PAC) for exploration (Kakade 2003).* The learner aims to explore the MDP so as to minimize the number of episodes in which the executed policy is more than ε-suboptimal:

$$N_\varepsilon \triangleq \sum_{t=1}^{\infty} \mathbb{1}\{V_1^\star(s_1) - V_1^{\boldsymbol{\pi}_t}(s_1) > \varepsilon\},$$

  where $\boldsymbol{\pi}_t$ denotes the policy executed in episode t. Under this objective, an RL algorithm is described by a sequence of random policies $(\boldsymbol{\pi}_t)_{t\in\mathbb{N}}$, with each $\boldsymbol{\pi}_t$ depending on the history up to episode $t-1$, as in regret minimization. An algorithm is called (ε,δ)-PAC-MDP if, with probability at least $1-\delta$, the number of ε-suboptimal episodes satisfies $N_\varepsilon \leqslant \mathrm{poly}(1/\varepsilon, \log(1/\delta))$.

The regret minimization framework is well suited for interactive learning scenarios where an agent is continuously deployed, such as in recommendation systems (Li et al. 2010). The objective is to minimize the cumulative regret over its entire operational period, which requires balancing exploration and exploitation. The PAC-MDP framework also suits this scenario; however, it requires specifying the accuracy ε beforehand, that is less likely to be known in advance, in contrast to a deployment period T.

The best policy identification (BPI) framework formalizes a more traditional "learning" paradigm. During training, the learner interacts with the environment, collects data, and improves its policy. After training (that happens at a random stopping time $\mathbf{t}$), the objective is to output a single policy that is ε-optimal for the given MDP. This setting is also called a *pure exploration* problem (Kaufmann et al. 2021; Ménard et al. 2021). In contrast to the BPI setting, neither regret minimisation nor PAC-MDP specifies which policy the practitioner should deploy. The last policy executed may still be suboptimal, even after N_ε episodes in the PAC-MDP setting.

On the connection between learning objectives. The three frameworks—regret minimization, BPI, and PAC-MDP—are distinct but related. Their connections highlight the trade-offs between online performance and final policy quality.

Regret Minimization and BPI. A regret minimization algorithm, which minimizes cumulative regret online, can be converted into a BPI algorithm that outputs a single high-quality policy. A common approach is to output a uniform mixture over all policies executed during training, that we define as $\widehat{\boldsymbol{\pi}} = \mathcal{U}\mathrm{nif}\left\{(\boldsymbol{\pi}_t)_{t\in[T]}\right\}$ and it is defined by the following trajectory sampling procedure:

1. Sample a timestep t from a uniform distribution over $[T]$.
2. Follow a policy $\boldsymbol{\pi}_t$ till the end of the episode.

The performance of this mixture policy is tied to the average regret, since $V_1^\star(s_1) - V_1^{\widehat{\boldsymbol{\pi}}}(s_1) = \mathfrak{R}(T)/T$. However, this requires storing all past policies and results in a history-dependent final policy. Another method involves using a data-dependent early stopping rule (Ménard et al. 2021), though this is not a black-box conversion. In the opposite direction, the BPI criterion does not imply low regret; deploying an ε-optimal policy from the start would lead to a linear regret of εT.

BPI and PAC-MDP. The BPI framework's goal of identifying an ε-optimal policy is stronger than the PAC-MDP guarantee. A BPI algorithm can be trivially converted into a PAC-MDP one: after the stopping time, one can repeatedly deploy the ε-optimal policy, ensuring all subsequent episodes are not ε-suboptimal. However, a PAC-MDP algorithm does not necessarily solve the BPI problem, as it only guarantees that few episodes are highly suboptimal but does not specify which policy should be used to guarantee ε-optimality.

Regret Minimization and PAC-MDP. Both frameworks are suited for interactive learning, but they are not equivalent. As shown by Dann et al. (2017), a regret-minimizing algorithm can be suboptimal in the PAC sense, and a PAC algorithm may not achieve low regret. The key difference lies in their objectives: regret minimization focuses on the cumulative loss over a known period T, while the PAC objective requires achieving ε-optimality over a bounded number of episodes.

Lower bounds. After defining our objectives, we now address a fundamental question: *What is the worst-case regret or sample complexity that any algorithm must suffer?* This question is answered by minimax lower bounds, which establish the performance limits for any algorithm interacting with a worst-case MDP.

For the finite-horizon setting, Domingues et al. (2021a) provided tight lower bounds, building on the sketch by Jin et al. (2018). The minimax lower bound for regret is

$$\min_{\text{algorithm } (\boldsymbol{\pi}_t)_{t\in[T]}} \max_{\text{MDP } \mathcal{M}} \mathfrak{R}_{\mathcal{M}}(T) \geqslant \Omega\left(\sqrt{H^3SAT}\right), \tag{1.8}$$

where $\mathfrak{R}_{\mathcal{M}}(T)$ is the regret suggered in an MDP $\mathcal{M}$ with horizon H, S states, and A actions. The minimum is taken over all possible regret-minimization algorithms $(\boldsymbol{\pi}_t)_{t\in[T]}$ (that is, sequences of non-stationary policies), and the maximum is taken over all MDPs with the specified parameters. This lower bound characterizes the worst-case scaling of regret with respect to all parameters of interest.

The corresponding lower bound for the sample complexity for best policy identification is given by

$$\min_{\text{algorithm } ((\boldsymbol{\pi}_t)_{t\in\mathbb{N}},\mathfrak{t},\widehat{\boldsymbol{\pi}})} \max_{\text{MDP } \mathcal{M}} \mathbb{E}_{\mathcal{M}}[\mathfrak{t}] \geqslant \Omega\left(\frac{H^3SA}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right), \tag{1.9}$$

where $\mathfrak{t}$ represents the sample complexity, defined as the number of episodes required by the algorithm to output a policy $\widehat{\boldsymbol{\pi}}$ that is ε-optimal with probability at least $1-\delta$ in the MDP $\mathcal{M}$ with horizon H, S states, and A actions. The minimum, as in the case of regret minimization, is taken over all possible BPI algorithms, each specified by a triple $((\boldsymbol{\pi}_t)_{t\in\mathbb{N}}, \mathfrak{t}, \widehat{\boldsymbol{\pi}})$, and the maximum is taken over all MDPs with the given parameters.

Both bounds are derived using a "needle-in-a-haystack" argument, where the hardness comes from identifying a single rewarding state-action path among many suboptimal ones. These bounds are known to be tight, and in the next section, we will describe an algorithm that matches them.

The remainder of the thesis focuses on regret minimisation and BPI; PAC-MDP is not considered further.

1.1.4 Upper Confidence Bound Value Iteration (UCBVI)

To introduce algorithms for reinforcement learning, we first focus on the regret minimization objective. A key challenge in this setting is balancing exploration and exploitation. One of the most well-studied principles to address this is *optimism in the face of uncertainty*. The idea is to act as if the environment is as favorable as plausibly possible given the current knowledge. In practice, this is often implemented by adding *exploration bonuses* to the rewards and performing *optimistic planning* with these modified rewards.

One of the first algorithms to achieve minimax regret bounds up to poly-logarithmic factors and a second-order term in T is `UCBVI` (Azar et al. 2017). It is a model-based optimistic algorithm that builds an empirical estimate of the transition probabilities and solves the Bellman optimality equations using these estimates, augmented with exploration bonuses. This approach has inspired numerous extensions (Dann et al. 2017; Zanette and Brunskill 2019; Zhang et al. 2021b; Zhang et al. 2024). However, `UCBVI` did not achieve a minimax optimality over all possible values of parameters S, A, H because of large second-order term in the regret bound. This gap was a motivation for a line of further research (Menard et al. 2021; Li et al. 2024; Zhang et al. 2024). In particular, a recent extension of `UCBVI`, Monotonic Value Propagation (Zhang et al. 2024), achieves the minimax regret bound for all problem parameter values, fully closing the gap between upper and lower minimax regret bounds (up to poly-logarithmic factors).

We briefly describe the `UCBVI` algorithm below, as it provides a foundation for methods developed later in this thesis.

Algorithm description. We introduce the following empirical counts, for all episodes $t \in [T]$, stages $h \in [H-1]$, state-action pairs $(s,a) \in \mathcal{S}\times\mathcal{A}$, and states $s' \in \mathcal{S}$:

$$n_h^t(s,a,s') \triangleq \sum_{t'=1}^{t} \mathbb{1}\left\{(S_h^{t'}, A_h^{t'}, S_{h+1}^{t'}) = (s,a,s')\right\}, \qquad n_h^t(s,a) \triangleq \sum_{s'\in\mathcal{S}} n_{t,h}(s,a,s'), \tag{1.10}$$

where $(S_1^{t'}, A_1^{t'}, S_2^{t'}, A_2^{t'}, \ldots, S_H^{t'}, A_H^{t'})$ is the trajectory generated by the algorithm in the episode t'.

We can now define the estimated transition kernels $\widehat{\boldsymbol{P}}^t$: for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $s' \in \mathcal{S}$, first for all $h \in [H-1]$,

$$\widehat{P}_h^t(s' \mid s,a) \triangleq \begin{cases} 1/S & \text{if } n_{t,h}(s,a) = 0\,, \\ \dfrac{n_{t,h}(s,a,s')}{n_{t,h}(s,a)} & \text{if } n_{t,h}(s,a) \geqslant 1\,. \end{cases} \tag{1.11}$$

With this definition in hand, we define the optimistic planning for `UCBVI` as follows for all $h \in [H]$

$$\begin{aligned} \widehat{Q}_h^t(s,a) &\triangleq r_h(s,a) + \widehat{P}_h^t \widehat{V}_{h+1}^t(s,a) + b_h^t(s,a)\,, \\ \widehat{V}_h^t(s) &\triangleq \min\Big\{ \max_{a \in \mathcal{A}} \widehat{Q}_h^t(s,a), H-h+1 \Big\}\,, \\ \pi_h^{t+1}(s) &\triangleq \arg\max_{a \in \mathcal{A}} \widehat{Q}_h^t(s,a)\,, \end{aligned} \tag{1.12}$$

where $\widehat{Q}_{H+1}^t(s,a) \equiv \widehat{V}_{H+1}^t(s) \equiv 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $b_h^t(s,a)$ is an exploration bonus that is added to the reward function to ensure the optimism condition with high probability: $\widehat{Q}_h^t(s,a) \geqslant Q_h^\star(s,a)$ for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$. Ties in $\arg\max$ are resolved arbitrarily; that is, $\pi_h^{t+1}(s)$ can be any element of $\arg\max_{a \in \mathcal{A}} \widehat{Q}_h^t(s,a)$ without further specification.

Analysis. For a proper choice of exploration bonuses, the `UCBVI` algorithm achieves the following regret bound:

$$\mathfrak{R}(T) = \widetilde{\mathcal{O}}\Big(\sqrt{H^3SAT} + H^3S^2A\Big)\,. \tag{1.13}$$

The main idea of the proof is to *couple* the part that corresponds to the optimal value and the part that corresponds to the value of the current policy. There are two main ingredients:

- Optimistic planning: $\widehat{V}_h^{t-1}(s) \geqslant V_h^\star(s)$ for any $(h,s) \in [H] \times \mathcal{S}$.
- The policy is *greedy* with respect to the optimistic value: $\pi_h^t(s) = \arg\max_{a \in \mathcal{A}} \widehat{Q}_h^{t-1}(s,a)$.

With these ingredients, we can proceed with the following bound on one-step simple regret:

$$\begin{aligned} V_h^\star(S_h^t) - V_h^{\boldsymbol{\pi}_t}(S_h^t) &\overset{(1)}{\leqslant} \underbrace{\widehat{V}_h^{t-1}(S_h^t) - V_h^{\boldsymbol{\pi}_t}(S_h^t)}_{\text{simple optimistic regret for a step } h} \overset{(2)}{\leqslant} \widehat{Q}_h^{t-1}(S_h^t, A_h^t) - Q_h^{\boldsymbol{\pi}_t}(S_h^t, A_h^t) \\ &\overset{(3)}{=} \widehat{P}_h^{t-1} \widehat{V}_{h+1}^{t-1}(S_h^t, A_h^t) - P_h V_{h+1}^{\boldsymbol{\pi}_t}(S_h^t, A_h^t) + b_h^{t-1}(S_h^t, A_h^t) \\ &\overset{(4)}{=} \underbrace{\widehat{V}_{h+1}^{t-1}(S_{h+1}^t) - V_{h+1}^{\boldsymbol{\pi}_t}(S_{h+1}^t)}_{\text{simple optimistic regret for a step } h+1} + \underbrace{\Big[\widehat{P}_h^{t-1} - P_h\Big] \widehat{V}_{h+1}^{t-1}(S_h^t, A_h^t)}_{\text{transition estimation error}} \\ &\quad + \underbrace{P_h \widehat{V}_{h+1}^{t-1}(S_h^t, A_h^t) - \widehat{V}_{h+1}^{t-1}(S_{h+1}^t)}_{\text{martingale noise } \xi_h^t} + \underbrace{b_h^{t-1}(S_h^{t+1}, A_h^{t+1})}_{\text{exploration bonus}}\,, \end{aligned}$$

where (1) follows from optimistic planning, which ensures $\widehat{V}_h^t(s) \geqslant V_h^\star(s)$ for any $(h,s) \in [H] \times \mathcal{S}$; (2) follows because the policy $\boldsymbol{\pi}_t$ is greedy with respect to the previous Q-value estimate $\widehat{Q}_h^{t-1}$; (3) follows from optimistic planning and the Bellman equations. The reward function cancels out automatically since both Q-values are evaluated at the same action. This cancellation is the reason for calling this argument a "coupling": rather than comparing the reward of the optimal action and the chosen one, the analysis uses an auxiliary sequence that is naturally coupled with the sequence of values $V_h^{\boldsymbol{\pi}_t}(s)$. The last step (4) results from rearranging the terms according to their role in the analysis.

Overall, ξ_h^{t+1} forms a martingale difference sequence, and its sum over T and H can be controlled easily using concentration of measure results such as the Azuma-Hoeffding inequality (Hoeffding 1963; Azuma 1967). The bonus terms are also straightforward to bound when summed over T and H. Thus, only first two terms are needed to be contolled.

The main challenge in the analysis is in how to control the transition estimation error; a naive approach with a bound on ℓ_1 norm of the transition estimation error will automatically lead to a suboptimal $\sqrt{H^3S^2AT}$ regret. The idea behind the `UCBVI` analysis is to use the following relation:

$$\Big[\widehat{P}_h^{t-1} - P_h\Big]\widehat{V}_{h+1}^{t-1}(s,a) = \underbrace{\Big[\widehat{P}_h^{t-1} - P_h\Big]V_{h+1}^\star(s,a)}_{V_{h+1}^\star \text{ is a fixed vector}} + \Big[\widehat{P}_h^{t-1} - P_h\Big][\widehat{V}_{h+1}^{t-1} - V_{h+1}^\star](s,a)\,,$$

where for the first term we have that $V_{h+1}^\star \in \mathbb{R}^S$ is a fixed vector and thus $\Big[\widehat{P}_h^{t+1} - P_h\Big]V_{h+1}^\star(s,a)$ is easy to control using Bernstein's inequality (Bernstein 1926), or, in a simplified versions, Hoeffding's inequality (Hoeffding 1963). The second term is more challenging; there are several approaches to handle it and all of them are connected to achiving a uniform Bernstein-style inequality of the form

$$\Big[\widehat{P}_h^{t-1} - P_h\Big]f(s,a) \leqslant \frac{1}{H}P_h f(s,a) + \widetilde{\mathcal{O}}\left(\frac{H^2 S}{n^{t-1}(s,a)}\right),$$

where $f\colon \mathcal{S}\times\mathcal{A} \to [0,H]$ is an arbitrary function. The original approach by Azar et al. (2017) considers using Bernstein inequality directly on estimated probabilities $\widehat{P}_h^{t-1}(s'|s,a) - P_h(s'|s,a)$ and an additional re-groupping argument. An alternative approach is to provide a uniform Bernstein-style inequality in terms of KL-divergence that follows from Donsker-Varadhan variation formula (Donsker and Varadhan 1983) combined with Bernstein-style bound on exponential moments (Boucheron et al. 2013), see Talebi and Maillard (2018):

$$\Big[\widehat{P}_h^{t-1} - P_h\Big]f(s,a) \leqslant \sqrt{2\mathrm{Var}_{P_h(s,a)}[f]\cdot \mathrm{KL}(\widehat{P}_h^{t-1}(s,a)\|P_h(s,a))} + \frac{2H}{3}\,\mathrm{KL}(\widehat{P}_h^{t-1}(s,a)\|P_h(s,a))\,,$$

where $P_h(s,a)$ denotes $P_h(\cdot|s,a)$. The next step is to obtain a concentration bound for KL-divergence between an empirical and , which is of order $\widetilde{\mathcal{O}}(\frac{S}{n_h^t(s,a)})$. With this bound, Young's inequality can be used to separate the variance term into $1/HP_hf(s,a)$ and a second-order term. For a concrete example, see Tiapkin et al. (2022a). This approach seems to be related to PAC-Bayesian theorems in statistical learning theory (McAllester 1999) because it uses the Donsker-Varadhan variational formula in a similar context. Another, more direct, approach builds an ε-net over all functions $f\colon \mathcal{S}\times\mathcal{A}\to[0,H]$, applies Bernstein's inequality to each point in the net, and then adjusts for the ε-approximation.

Overall, this uniform bound after plugging-in $f = \widehat{V}_{h+1}^{t+1} - V_{h+1}^\star$ results in the following inequality

$$\Big[\widehat{P}_h^{t+1} - P_h\Big][\widehat{V}_{h+1}^{t+1} - V_{h+1}^\star](s,a) \leqslant \frac{1}{H}P_h[\widehat{V}_{h+1}^{t+1} - V_{h+1}^\star](s,a) + \widetilde{\mathcal{O}}\left(\frac{H^2 S}{n_h^t(s,a)}\right),$$

where the second term results in a H^3S^2A second-order term in the final regret bound, but does not affect the main (first-order) term. The additional $1/H$ factor to an optimistic regret does not harm the first-order term since it accumulates over no more than H steps, resulting only in $(1+1/H)^H \leqslant \mathrm{e}$ additional factor for the first-order term.

On the second-order term. The standard UCBVI analysis yields a suboptimal second-order term of order H^3S^2A. Recently, the open problem of achieving an optimal second-order term was resolved in the work of Zhang et al. (2024). This result was notable because the prevailing belief

was that such improvements would come from refining variance-reduced versions (Zhang et al. 2020; Li et al. 2024; Menard et al. 2021) of Q-learning (Watkins and Dayan 1992; Jin et al. 2018), rather than from a model-based approach.

The main idea of the approach in Zhang et al. (2024) is a more careful analysis of the transition estimation error. Rather than bounding the error against any possible value function difference $\widehat{V}_{h+1}^{t+1} - V_{h+1}^{\star}$, their analysis leverages the fact that the optimistic value function $\widehat{V}_{h+1}^{t+1}$ is constructed by the algorithm itself and, if the number of samples for all state-action pair is fixed, is independent from $\widehat{P}_h^{t+1}$. By applying a doubling trick, analyzing the structure of $\widehat{V}_{h+1}^{t+1}$ more precisely, and using a very careful union bound over possible algorithm "profiles" (that encode the number of visits for all state-action pairs and the same time) they derive a tighter bound that avoids the suboptimal H^3S^2A term.

Crucially for this thesis, their refined analysis avoids an additional recursive term that appears in the standard UCBVI proof. This property is extremely important, as this recursive term would be difficult to control in the adversarial reward setting (Chapter 4).

Best policy identification. In the case of the best-policy identification problem, it is possible to adapt `UCBVI` to provide BPI guarantees, as shown by Ménard et al. (2021). The core idea is to define a stopping rule based on a high-probability upper bound on the suboptimality gap, $G_1^t(s_1) \geqslant V_1^{\star}(s_1) - V_1^{\boldsymbol{\pi}_t}(s_1)$. The stopping time $\mathfrak{t}$ is then defined as the first episode t where this gap estimate is smaller than the target precision ε:

$$\mathfrak{t} \triangleq \min\{t \in \mathbb{N} : G_1^t(s_1) \leqslant \varepsilon\}\,.$$

By definition, for any episode $t < \mathfrak{t}$, we have $G_1^t(s_1) > \varepsilon$. Summing over the episodes before stopping yields $\sum_{t=1}^{\mathfrak{t}-1} G_1^t(s_1) > (\mathfrak{t}-1)\varepsilon$. This inequality connects the sample complexity $\mathfrak{t}$ to the cumulative sum of the gap estimates. The structure of these gap estimates is such that they are almost the same as optimistic regret, and this allows us to derive an upper bound on the stopping time $\mathfrak{t}$ with high probability, which corresponds to the sample complexity for BPI.

1.1.5 Beyond Standard Objectives

The classical RL objective of maximizing expected cumulative reward is already addressed by minimax-optimal algorithms described above, but it does not cover several important scenarios in modern applications. Reinforcement learning is often used in contexts where additional objectives, constraints, or dynamic environments are central. This motivates the study of alternative or extended RL objectives, such as:

- **Explore without rewards.** Rather than focusing solely on reward maximization, the agent may be tasked with exploring the environment as broadly as possible. For example, maximizing the entropy of the induced state-action or trajectory distribution is particularly relevant for pretraining policies in complex domains, where comprehensive coverage is crucial for subsequent downstream tasks (Seo et al. 2021; Zhang et al. 2021a; Mutti et al. 2022). This setting is studied in Chapter 2.
- **Leverage expert demonstrations.** In domains like robotics (Atkeson and Schaal 1997) or language model alignment (Ouyang et al. 2022), access to large datasets of expert demonstrations is common. Integrating these demonstrations into the reinforcement-learning pipeline can accelerate training and improve sample efficiency; we examine this question in Chapter 3.
- **Adapt to adversarial rewards.** In many real-world settings, the reward function may evolve over time due to inherent non-stationarity or adversarial changes. The questions of adaptivity of RL algorithms to these changes is studied in Chapter 4.

These extensions lie outside the standard RL paradigm and raise new algorithmic and theoretical challenges. The next sections examine each direction in turn, state the key problems, and highlight this thesis's main contributions.

1.2 Maximum Entropy Exploration (Summary of Chapter 2)

This section and Chapter 2 are based on: D. Tiapkin, D. Belomestny, D. Calandriello, É. Moulines, R. Munos, A. Naumov, P. Perrault, Y. Tang, M. Valko, and P. Ménard, "Fast Rates for Maximum Entropy Exploration", Proceedings of the 40th International Conference on Machine Learning (ICML 2023) (Tiapkin et al. 2023a).

The central question addressed in this section is: *How can an agent learn a policy that explores an MDP effectively in the absence of any reward signal?* At first glance, this may seem at contradiction with the reinforcement learning foundational — namely, that "reward is enough" (Silver et al. 2021). However, reward-free exploration may be important for RL pipelines, particularly for pretraining agents in environments where the reward function is either unknown, task-specific, or not yet defined (Seo et al. 2021; Zhang et al. 2021a; Mutti et al. 2022). In such settings, the agent's objective shifts from maximizing cumulative reward to achieving broad and informative coverage of the state-action or trajectory space, thereby facilitating efficient adaptation to downstream tasks once a reward signal becomes available.

1.2.1 Problem Definition

In this section, we consider interaction with a *reward-free MDP* $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \boldsymbol{P}, H, s_1)$, which is defined identically to a standard MDP but without a reward function: that is, during interaction with the environment, the agent does not observe any reward signal.

Next, our goal is to learn a policy that provides broad coverage of the MDP. Several approaches address this challenge. One line of work adds *intrinsic-motivation* rewards to encourage exploration (Meyer and Wilson 1991; Oudeyer et al. 2007; Bellemare et al. 2016). Another, called *goal-conditioned exploration*, trains the agent to reach a diverse set of goals (Lim and Auer 2012; Tarbouriech et al. 2020b; Ecoffet et al. 2021).

In this section and in Chapter 2, we consider a different strategy: directly maximising an entropy criterion. Two natural notions of entropy arise:

- **Visitation entropy:** This objective seeks to maximize the entropy of the occupancy measures (visitation distributions) across all steps:

$$\max_{\boldsymbol{\pi} \in \Pi} \mathcal{H}_{\text{visit}}(\boldsymbol{d}^{\boldsymbol{\pi}}) \triangleq \sum_{h=1}^{H} \mathcal{H}(d_h^{\boldsymbol{\pi}}) \,, \tag{1.14}$$

  wher $\boldsymbol{d}^{\boldsymbol{\pi}}$ is defined in (1.2). Maximizing visitation entropy encourages the agent to visit as diverse a set of state-action pairs as possible throughout the episode. This promotes broad coverage of state and actions of the environment, which is crucial to open unexplored parts of the MDP.

- **Trajectory entropy:** The goal is to maximize the entropy of the trajectory distribution induced by the policy.

$$\max_{\boldsymbol{\pi} \in \Pi} \mathcal{H}_{\text{traj}}(q^{\boldsymbol{\pi}}) \triangleq \mathcal{H}(q^{\boldsymbol{\pi}}) = \sum_{\tau \in \mathcal{T}} q^{\boldsymbol{\pi}}(\tau) \log \frac{1}{q^{\boldsymbol{\pi}}(\tau)} \,, \tag{1.15}$$

where $q^{\boldsymbol{\pi}}$ is defined in (1.1). This approach incentivizes the agent to generate a diverse set of entire trajectories. This objective focuses on covering the whole trajectory space instead of just state-action pairs.

When collecting data for subsequent RL training, the choice of exploration objective often depends on the downstream algorithm. For example, Path Consistency Learning (Nachum et al. 2017) benefits from gathering diverse trajectories, while Deep Q-Networks (Mnih et al. 2015) typically prioritize coverage of a wide range of state-action pairs.

In the following, we discuss both objectives and the ways to optimize them.

1.2.2 Maximum Trajectory Entropy Exploration (MTEE)

We start from maximizing a trajectory-wise objective since its analysis contains very useful insights on the optimization of the visitation entropy.

The main property of the trajectory entropy is the following decomposability:

$$\begin{aligned}\mathcal{H}_{\mathrm{traj}}(q^{\boldsymbol{\pi}}) &= \mathbb{E}_{\tau\sim q^{\boldsymbol{\pi}}}[-\log q^{\boldsymbol{\pi}}(\tau)] = \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{P}}\left[\sum_{h=1}^{H-1} -\log P_h(S_{h+1}\mid S_h, A_h) + \sum_{h=1}^{H} -\log \pi_h(A_h\mid S_h)\right]\\ &= \sum_{h=1}^{H}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{P}}[-\log P_h(S_{h+1}\mid S_h, A_h)\mathbb{1}\{h<H\}] + \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{P}}[-\log \pi_h(A_h|S_h)]\\ &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{P}}\left[\sum_{h=1}^{H}\mathcal{H}(P_h(\cdot\mid S_h, A_h))\mathbb{1}\{h<H\} + \mathcal{H}(\pi_h(S_h))\right],\end{aligned}$$

where $\mathcal{H}(p)$ is a Shannon entropy function, and the last equality follows from the properties of conditional expectation. In particular, it implies that the trajectory entropy is a value function in a usual MDP with rewards $r_h(s,a) = \mathcal{H}(P_h(\cdot\mid s,a))\mathbb{1}\{h<H\}$ and an additional *entropy regularization*.

Regularized MDPs. A regularized MDP augments the standard RL objective by adding a regularization term to the reward, which can promote exploration, encode prior knowledge, or enforce constraints (Neu et al. 2017; Geist et al. 2019). Specifically, let $\Omega_h\colon \mathcal{S}\times\Delta(\mathcal{A})\to\mathbb{R}$ be a Legendre-type[2] convex function (in the second argument) that may depend on the state at each step h. For a Markovian[3] policy $\boldsymbol{\pi} = (\pi_h)_{h=1}^H$, the regularized value function at step h and state s is defined as

$$V_h^{\boldsymbol{\pi},\Omega}(s) \triangleq \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{P}}\left[\sum_{h'=h}^{H}\left(r_{h'}(S_{h'},A_{h'}) + \Omega_{h'}(S_{h'},\pi_{h'}(S_{h'}))\right)\middle| S_h = s\right], \tag{1.16}$$

and the corresponding Q-value function is

$$Q_h^{\boldsymbol{\pi},\Omega}(s,a) \triangleq \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{P}}\left[r_h(s,a) + \sum_{h'=h+1}^{H}\left(r_{h'}(S_{h'},A_{h'}) + \Omega_{h'}(S_{h'},\pi_{h'}(S_{h'}))\right)\middle| S_h = s,\ A_h = a\right]. \tag{1.17}$$

Note that the regularization term $\Omega_h(s,\pi_h(s))$ is included in the value function but not in the Q-value at the current step, reflecting the fact that regularization is applied after the action is chosen. This leads to the following regularized Bellman equations:

$$\begin{aligned}Q_h^{\boldsymbol{\pi},\Omega}(s,a) &= r_h(s,a) + P_h V_{h+1}^{\boldsymbol{\pi},\Omega}(s,a)\,,\\ V_h^{\boldsymbol{\pi},\Omega}(s) &= \pi_h Q_h^{\boldsymbol{\pi},\Omega}(s) + \Omega_h(s,\pi_h(s))\,.\end{aligned} \tag{1.18}$$

[2]We refer to Rockafellar (1997, Section 26) for the definition of Legendre-type convex functions.

[3]While regularized value functions can be defined for history-dependent policies, we focus on Markovian policies for clarity.

When $\Omega_h(s,\pi) = \lambda \cdot \mathcal{H}(\pi)$ is the Shannon entropy and $\lambda > 0$ is a regularization coefficient, this gives the standard entropy-regularized RL objective, which is the basis for maximum trajectory entropy exploration and algorithms such as Soft Q-Learning (Fox et al. 2016; Haarnoja et al. 2017; Schulman et al. 2018) and Soft Actor-Critic (Haarnoja et al. 2018). Our general formulation of the regularizer Ω_h also covers other important cases, such as KL-regularization $\Omega_h(s,\pi) = \lambda \cdot \mathrm{KL}(\pi \| \widetilde{\pi}_h(s))$ for a reference policy $\widetilde{\boldsymbol{\pi}}$ and $\lambda > 0$, which will be particularly relevant in the context of learning from demonstrations (see Chapter 3).

In this setting, the optimal value and Q-value functions are defined as suprema over all policies, and satisfy the regularized Bellman optimality equations:

$$\begin{aligned} Q_h^{\star,\Omega}(s,a) &= r_h(s,a) + P_h V_{h+1}^{\star,\Omega}(s,a)\,, \\ V_h^{\star,\Omega}(s) &= \max_{\pi \in \Delta(\mathcal{A})} \Big\{ \mathbb{E}_{A\sim\pi}[Q_h^{\star,\Omega}(s,A)] + \Omega_h(s,\pi) \Big\}\,, \end{aligned} \tag{1.19}$$

and the optimal policy $\boldsymbol{\pi}^{\star,\Omega}$, i.e., a policy such that $V_h^{\boldsymbol{\pi}^{\star,\Omega},\Omega}(s) = V_h^{\star,\Omega}(s)$, is given by

$$\pi_h^{\star,\Omega}(s) = \arg\max_{\pi \in \Delta(\mathcal{A})} \Big\{ \mathbb{E}_{A\sim\pi}[Q_h^{\star,\Omega}(s,A)] + \Omega_h(s,\pi) \Big\}. \tag{1.20}$$

To simplify notation, we denote $Q_h^{\boldsymbol{\pi},\lambda}(s,a)$ and $V_h^{\boldsymbol{\pi},\lambda}(s)$ for the entropy-regularized case with coefficient $\lambda > 0$, instead of $Q_h^{\boldsymbol{\pi},\Omega}(s,a)$ and $V_h^{\boldsymbol{\pi},\Omega}(s)$.

For certain choices of the regularizer Ω_h, the optimal policy admits a closed-form solution. In particular, if $\Omega_h(s,\pi) = \lambda\mathcal{H}(\pi)$ is the Shannon entropy regularizer with $\lambda > 0$, the optimal policy at each step h and state s is the Boltzmann (softmax) distribution:

$$\pi_h^{\star,\lambda}(a \mid s) = \frac{\exp\Big(\lambda^{-1} Q_h^{\star,\lambda}(s,a)\Big)}{\sum_{a' \in \mathcal{A}} \exp\Big(\lambda^{-1} \cdot Q_h^{\star,\lambda}(s,a')\Big)}\,. \tag{1.21}$$

Substituting this policy back into the regularized Bellman optimality equation for the value function (1.19) yields a closed-form expression for the optimal value, often referred to as the *soft-max* or *log-sum-exp* operator:

$$V_h^{\star,\lambda}(s) = \lambda \log\left(\sum_{a' \in \mathcal{A}} \exp\Big(\lambda^{-1} Q_h^{\star,\lambda}(s,a')\Big) \right). \tag{1.22}$$

Properties of the regularized Bellman operator. To facilitate the analysis of regularized MDPs, we now introduce an operator that maps a Q-function to its corresponding regularized value. Let us define the following function for any $Q_h \in \mathbb{R}^{S\times A}$:

$$F_{h,\Omega}(Q_h)(s) \triangleq \max_{\pi \in \Delta(\mathcal{A})} \{ \mathbb{E}_{A\sim\pi}[Q_h(s,A)] + \Omega_h(s,\pi) \}\,. \tag{1.23}$$

With this notation, the optimal value function can be expressed as $V_h^{\star,\Omega}(s) = F_{h,\Omega}(Q_h^{\star,\Omega})(s)$. This operator has several important properties derived from convex duality, which we summarize below.

Since the function $\Omega_h(s,\cdot)$ is of Legendre-type, the function $F_{h,\Omega}(Q_h)(s)$ is the convex conjugate of $-\Omega_h(s,\cdot)$ evaluated at $Q_h(s,\cdot)$, i.e., $F_{h,\Omega}(Q_h)(s) = (-\Omega_h(s,\cdot))^*(Q_h(s,\cdot))$. This duality has two key consequences:

- **Smoothness.** If $-\Omega_h(s,\cdot)$ is λ-strongly convex with respect to the ℓ_1-norm, then its conjugate $F_{h,\Omega}(Q_h)(s)$ is $1/\lambda$-smooth with respect to the ℓ_∞-norm (Kakade et al. 2009). Formally, for

any $Q_h, Q'_h \in \mathbb{R}^{S\times A}$:

$$F_{h,\Omega}(Q_h)(s) \leqslant F_{h,\Omega}(Q'_h)(s) + \langle \nabla_{Q'_h(s,\cdot)} F_{h,\Omega}(Q'_h)(s), Q_h(s,\cdot) - Q'_h(s,\cdot)\rangle + \frac{1}{2\lambda}\|Q_h(s,\cdot) - Q'_h(s,\cdot)\|_\infty^2 .$$

This property holds for both entropy and KL-divergence regularizers.

- **Greedy Policy as Gradient.** The policy that attains the maximum in (1.23), which we call the regularized greedy policy with respect to Q_h, can be obtained as the gradient of $F_{h,\Omega}$ with respect to $Q_h(s,\cdot)$:

$$\pi_h^{Q_h,\Omega}(s) = \arg\max_{\pi\in\Delta(\mathcal{A})}\{\mathbb{E}_{A\sim\pi}[Q_h(s,A)] + \Omega_h(s,\pi)\} = \nabla_{Q_h(s,\cdot)} F_{h,\Omega}(Q_h)(s) .$$

Consequently, the optimal policy is given by $\pi_h^{\star,\Omega}(s) = \nabla_{Q_h^{\star,\Omega}(s,\cdot)} F_{h,\Omega}(Q_h^{\star,\Omega})(s)$.

Solving regularized MDPs. Solving entropy-regularized MDPs has been extensively studied in the policy gradient literature because regularization enhances the smoothness of the value function. For example, Agarwal et al. (2020b) proved that the softmax policy gradient method achieves polynomial convergence rates. Later, Mei et al. (2020) and Cen et al. (2022) established that the convergence is actually linear and global.

However, these guarantees are purely algorithmic: they presuppose that the transition dynamics are known and depend on potentially large constants that depend on the initial parameters of the policy gradient algorithm. Neither assumption is realistic in more applied cases: we rarely know the transition dynamics beforehand, and the hidden constants can be extremly large. Therefore, in this thesis, we ask a complementary question: how efficiently can we solve entropy-regularized MDPs when the kernel must be learned from data? Chapter 2 answers this question and provides the first sample-complexity guarantees for the task. In the following, we outline the approach and all main techniques.

A naive approach would be to adapt an algorithm like `UCBVI` by replacing standard optimistic planning with its regularized counterpart. However, this method, which we call `UCBVI-Ent`, achieves a sample complexity of $\widetilde{\mathcal{O}}(\mathrm{poly}(H,S,A)/\varepsilon^2)$ for best-policy identification, which turns out to be suboptimal for this problem in contrast to the unregularized counterpart.

To achieve a better sample complexity, we must look deeper into the problem's structure. The key insight lies in the relationship between the accuracy of the estimated Q-function and the suboptimality of the resulting policy. Let $\widehat{Q}$ be an estimate of the optimal Q-function $Q^{\star,\lambda}$, and let $\widehat{\boldsymbol{\pi}}$ be the regularized greedy policy with respect to $\widehat{Q}$ (as defined in (1.20) with $Q^{\star,\lambda}$ replaced by $\widehat{Q}$). From the properties of the function $F_{h,\Omega}$ (defined in (1.23)), one can derive the following bound on the suboptimality gap:

$$V_1^{\star,\lambda}(s_1) - V_1^{\widehat{\boldsymbol{\pi}},\lambda}(s_1) \leq \frac{1}{2\lambda}\mathbb{E}_{\widehat{\boldsymbol{\pi}},\boldsymbol{P}}\left[\sum_{h=1}^{H}\|Q_h^{\star,\lambda}(S_h,\cdot) - \widehat{Q}_h(S_h,\cdot)\|_\infty^2\right]. \tag{1.24}$$

This inequality shows that the suboptimality of $\widehat{\boldsymbol{\pi}}$ is controlled by the *squared* ℓ_∞-error of the Q-function estimate, weighted by the visitation distribution of $\widehat{\boldsymbol{\pi}}$. If we can guarantee that $\|Q_h^{\star,\lambda} - \widehat{Q}_h\|_\infty \leq \delta_Q$ for all h, the total error is bounded by $\frac{H}{2\lambda}\delta_Q^2$. Thus, to achieve an ε-optimal policy, we need an estimate $\widehat{Q}$ with accuracy $\delta_Q \approx \sqrt{2\lambda\varepsilon/H}$.

In a generative model setting[4], results from Azar et al. (2013) suggest that achieving such accuracy requires $\widetilde{\mathcal{O}}(\mathrm{poly}(H,S,A)/\delta_Q^2) = \widetilde{\mathcal{O}}(\mathrm{poly}(H,S,A)/(\lambda\varepsilon))$ samples per state-action pair. However, in the standard online RL setting, we do not have a generative model. A simple approach

[4]When there is a sampling oracle for each state-action pair.

is to use a reward-free exploration algorithm (Jin et al. 2020c) to collect a dataset that effectively simulates a generative model.

In our work, which we discuss in detail in Chapter 2, we implemented this approach and named the resulting algorithm `RL-Explore-Ent`. We show that `RL-Explore-Ent` achieves a sample complexity of order $\widetilde{\mathcal{O}}(\mathrm{poly}(H,S,A)/\lambda\varepsilon)$ for best-policy identification in entropy-regularized MDPs. When considering maximum trajectory entropy exploration, which corresponds to the case $\lambda = 1$, the sample complexity becomes of order $\widetilde{\mathcal{O}}(\mathrm{poly}(H,S,A)/\varepsilon)$. This result improves the dependence on ε compared to the more naive `UCBVI-Ent` approach, which requires $\widetilde{\mathcal{O}}(\mathrm{poly}(H,S,A)/\varepsilon^2)$ samples. However, the sample complexity's dependence on H, S, and A onder poly increases, leaving some room for additional improvements.

Discussion. The problem of maximum trajectory entropy exploration was an initial goal of this subsection. However, eventually, the analysis for this objective led to general methods that achieve a sample complexity of order $\widetilde{\mathcal{O}}\big(\mathrm{poly}(H,S,A)/(\lambda\varepsilon)\big)$ for a broader class of entropy-regularized RL problems. These techniques play a crucial role in improving convergence results for maximum visitation entropy exploration, as discussed in the next subsection, and they serve as a foundation for the learning-from-demonstrations analysis in Chapter 3. One limitation of the current results is the high dependence of the sample complexity upper bound on H, S, and A for the final algorithm `RL-Explore-Ent`. The next section (Section 1.3) introduces refinements that address this issue.

1.2.3 Maximum Visitation Entropy Exploration (MVEE)

We now turn to the maximum visitation entropy exploration (MVEE) problem. The objective is to find a policy that maximizes the sum of entropies of its state-action visitation distributions:

$$\max_{\boldsymbol{\pi}\in\Pi} \mathcal{H}_{\mathrm{visit}}(\boldsymbol{d}^{\boldsymbol{\pi}}) = \sum_{h=1}^{H} \mathcal{H}(d_h^{\boldsymbol{\pi}})\,,$$

where $d_h^{\boldsymbol{\pi}}$ is the visitation distribution at step h (see Equation (1.2)). This objective encourages the agent to visit a diverse set of state-action pairs at each step of the episode, which is beneficial for building a comprehensive model of the environment.

Unlike a trajectory entropy, which can be decomposed into a sum of local rewards, the visitation entropy objective does not admit such a simple reformulation. Furthermore, the mapping from the policy $\boldsymbol{\pi}$ to the visitation distributions $\boldsymbol{d}^{\boldsymbol{\pi}}$ is complex, and the objective function $\mathcal{H}_{\mathrm{visit}}(\boldsymbol{d}^{\boldsymbol{\pi}})$ is non-concave with respect to the policy parameters and does not admit a dynamic-programming structure, like a trajectory entropy. This makes direct optimization challenging.

To overcome this difficulty, we shift our perspective from the space of policies to the space of occupancy measures (or visitation distributions). The set of all achievable visitation distributions for Markovian policies forms a convex polytope, which we denote by $\mathcal{K}_{\boldsymbol{P}}$:

$$\begin{aligned}\mathcal{K}_{\boldsymbol{P}} \triangleq \Big\{ \boldsymbol{d} = (d_h)_{h\in[H]} : d_h(s,a) \geqslant 0,\ \forall (h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}\,,\\ \sum_{a\in\mathcal{A}} d_1(s,a) = \mathbb{1}\{s = s_1\}\ \forall s\in\mathcal{S}\\ \sum_{a\in\mathcal{A}} d_{h+1}(s,a) = \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} P_h(s|s',a')d_h(s',a')\ \forall s\in\mathcal{S}, \forall h\in[H-1]\Big\}\,.\end{aligned}$$

This formulation is standard in the literature on a linear-programming perspective on MDPs (e.g., Altman 2021). Since the entropy function is concave, the MVEE problem can be recast

as a concave maximization problem over this convex set:

$$\max_{\boldsymbol{d}\in\mathcal{K}_{\boldsymbol{P}}} \sum_{h=1}^{H} \mathcal{H}(d_h)\,.$$

This convex reformulation is the starting point for developing efficient algorithms. For the subsequent analysis, it is also convenient to define $\mathcal{K}$ as the larger set of all valid sequences of distributions, without the dynamic constraints imposed by the MDP:

$$\mathcal{K} \triangleq \left\{\boldsymbol{d} = (d_h)_{h\in[H]} : d_h(s,a) \geqslant 0,\ \forall (h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}\,,\ \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h(s,a) = 1,\ \forall h\in[H]\right\}.$$

Solving MVEE via Frank-Wolfe algorithm. A common approach to solving the MVEE problem relies on its convex formulation. The set of valid occupancy measures $\mathcal{K}_{\boldsymbol{P}}$ is a convex polytope, and the entropy objective is concave. A key property of this polytope is that linear optimization over it is equivalent to solving a standard RL problem:

$$\max_{\boldsymbol{d}\in\mathcal{K}_{\boldsymbol{P}}} \langle \boldsymbol{d}, \boldsymbol{r}\rangle = \max_{\boldsymbol{\pi}\in\Pi} \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{P}}\left[\sum_{h=1}^{H} r_h(S_h, A_h)\right].$$

This structure makes the problem well-suited for the Frank-Wolfe algorithm (Hazan et al. 2019), which iteratively finds an optimal policy for a linear objective (the gradient of the entropy) and takes a step in that direction. A sketch of the algorithm is as follows:

- Initialize $\boldsymbol{d}^0 \in \mathcal{K}_{\boldsymbol{P}}$.
- For $t = 1, 2, \ldots, T$:
  - Compute the reward vector $\boldsymbol{r}_t = \nabla\mathcal{H}_{\text{visit}}(\boldsymbol{d}^{t-1})$.
  - Find the vertex $\widetilde{\boldsymbol{d}}^t = \arg\max_{\boldsymbol{d}\in\mathcal{K}_{\boldsymbol{P}}}\langle \boldsymbol{d}, \boldsymbol{r}_t\rangle$ by solving an RL problem.
  - Update the solution: $\boldsymbol{d}^t = (1-\alpha_t)\boldsymbol{d}^{t-1} + \alpha_t\widetilde{\boldsymbol{d}}^t$ for a step-size α_t.

However, a naive application of this method in the online RL setting faces several challenges:

- **Non-smoothness:** The entropy objective is not smooth, which complicates the analysis;
- **Expensive updates:** Each Frank-Wolfe iteration requires solving a complete RL problem to find the optimal policy for the current gradient, which can be computationally prohibitive;
- **Occupancy Measure Computation:** The algorithm operates on occupancy measures $\boldsymbol{d}$, but an RL oracle returns a policy $\boldsymbol{\pi}$. Computing the occupancy measure of a policy requires knowing the transition model $\boldsymbol{P}$, which is typically unknown and must be learned.

Prior work has addressed these issues with varying degrees of success. The non-smoothness can be handled by smoothing the objective by adding a small regularization term under the logarithm, as proposed by Hazan et al. (2019) and later refined by Cheung (2019). To avoid solving a full RL problem at each step, Zahavy et al. (2021) employed a game-theoretic interpretation of the Frank-Wolfe algorithm (Abernethy and Wang 2017) and replaced the complete solution of RL problem with only one step of regret minimization algorithm. However, their analysis assumes access to an oracle that can efficiently compute the occupancy measure of any given policy, sidestepping the statistical challenge of estimating these quantities from data in an unknown MDP, resulting in an additional $\widetilde{\mathcal{O}}(1/\varepsilon^3)$ factor in the sample complexity.

Solving MVEE via `EntGame` algorithm. Next, we describe our approach to solve the MVEE problem, desrcibed further in Chapter 2 of this thesis. The main idea takes inspiration from a similar

idea as in Zahavy et al. (2021), but we use a different game to solve the problem (as proposed by Grünwald and Dawid (2002)). The game is defined as follows:

$$\max_{\boldsymbol{d}\in\mathcal{K}_p} \mathcal{H}_{\text{visit}}(\boldsymbol{d}) = \max_{\boldsymbol{d}\in\mathcal{K}_p} \min_{\bar{\boldsymbol{d}}\in\mathcal{K}} \sum_{(h,s,a)} d_h(s,a) \log \frac{1}{\bar{d}_h(s,a)} = \min_{\bar{\boldsymbol{d}}\in\mathcal{K}} \max_{\boldsymbol{d}\in\mathcal{K}_{\boldsymbol{P}}} \sum_{(h,s,a)} d_h(s,a) \log \frac{1}{\bar{d}_h(s,a)},$$

This minimax game is equivalent to an online learning problem with two interacting players. The first player, the *forecaster*, aims to predict the state-action visitation distribution of the second player. The second player, the *sampler*, is rewarded for visiting state-action pairs that the forecaster assigned low probability, effectively encouraging exploration of less predictable parts of the state-action space.

The interaction proceeds in episodes. At each episode t, the forecaster forms a prediction $\bar{\boldsymbol{d}}^t$ based on the history of visits. It uses a standard strategy for online prediction with logarithmic loss known as a mixture forecaster (Cesa-Bianchi and Lugosi 2006). Specifically, after $t-1$ episodes, it has observed counts $n_h^{t-1}(s,a)$ for each state-action pair (s,a) at step h. The forecaster's prediction for episode t is a smoothed empirical distribution:

$$\bar{d}_h^t(s,a) = \frac{n_h^{t-1}(s,a) + n_0}{t-1+SAn_0},$$

where $n_0 > 0$ is a prior count (e.g., $n_0 = 1$). This is equivalent to the posterior mean of a Dirichlet-multinomial model. The sampler then acts in the environment by executing a policy that maximizes the rewards $r_h(s,a) = \log(1/\bar{d}_h^t(s,a))$. This policy is computed using an optimistic planning algorithm, similar to `UCBVI`, based on the currently estimated transition model. After T episodes, the algorithm `EntGame` returns a policy $\widehat{\pi}$ that is a uniform mixture of the policies returned by the sampler-player.

Analysis. To analyze the performance of `EntGame`, we start with an error decomposition. We first define the regret of each player, obtained by playing the game T times. For the forecaster-player, for any $\bar{\boldsymbol{d}} \in \mathcal{K}$, we define

$$\mathfrak{R}_{\text{Fore}}^T(\bar{\boldsymbol{d}}) \triangleq \sum_{t=1}^T \sum_{h,s,a} \widetilde{d}_h^t(s,a) \left(\log \frac{1}{\bar{d}_h^t(s,a)} - \log \frac{1}{\bar{d}_h(s,a)} \right)$$

where $\widetilde{d}_h^t(s,a) \triangleq \mathbb{1}\left\{(s_h^t, a_h^t) = (s,a)\right\}$ is a sample from $d_h^{\pi^t}(s,a)$. Similarly, for the sampler-player, for any $\boldsymbol{d} \in \mathcal{K}_{\boldsymbol{P}}$, we define

$$\mathfrak{R}_{\text{Samp}}^T(\boldsymbol{d}) \triangleq \sum_{t=1}^T \sum_{h,s,a} \left(d_h(s,a) - d_h^{\pi^t}(s,a) \right) \log \frac{1}{\bar{d}_h^t(s,a)}.$$

Recall that the policy returned by `EntGame` is a uniform mixture of the policies played by the sampler. Thus, the corresponding visitation distribution is the average of the sampler's visitation distributions: $d_h^{\widehat{\pi}}(s,a) = \widehat{d}_h^T(s,a) \triangleq (1/T)\sum_{t=1}^T d_h^{\pi^t}(s,a)$. We also denote by $\mathring{d}_h^T(s,a) \triangleq (1/T)\sum_{t=1}^T \widetilde{d}^t(s,a)$ the average of the empirical visitation distributions. Also recall that $\bar{\boldsymbol{d}}^t$ is a prediction by a forecaster-player.

We now relate the difference between the optimal visitation entropy and the visitation entropy of the output policy $\widehat{\pi}$ with the regrets of the two players. Indeed, using $\mathcal{H}(p) = \sum_{i\in[n]} p_i \log(1/q_i) -$

$\mathrm{KL}(p,q) \leqslant \sum_{i\in[n]} p_i \log(1/q_i)$ for all $(p,q) \in (\Delta([n]))^2$, we obtain

$$\mathcal{H}_{\mathrm{visit}}(\boldsymbol{d}^{\boldsymbol{\pi}^\star}) - \mathcal{H}_{\mathrm{visit}}(\boldsymbol{d}^{\widehat{\boldsymbol{\pi}}}) \leqslant \frac{1}{T}\mathfrak{R}^T_{\mathrm{Samp}}(\boldsymbol{d}^{\boldsymbol{\pi}^\star}) + \frac{1}{T}\mathfrak{R}^T_{\mathrm{Fore}}(\mathring{\boldsymbol{d}}^T) + \underbrace{\mathcal{H}_{\mathrm{visit}}(\mathring{\boldsymbol{d}}^T) - \mathcal{H}_{\mathrm{visit}}(\widehat{\boldsymbol{d}}^T)}_{\text{Bias term}} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\sum_{h,s,a}\left(d_h^{\pi^t}(s,a) - \widetilde{d}_h^t(s,a)\right)\log\frac{1}{\bar{d}_h^t(s,a)}}_{\text{Noise term}} . \tag{1.25}$$

Standard game-theoretic analysis bounds the objective by the sum of the regrets of the max and min players. In our case, we have also extracted noise and bias terms, which are easy to bound:

$$\text{Noise term} = \widetilde{\mathcal{O}}\left(\sqrt{H/T}\right), \quad \text{Bias term} = \widetilde{\mathcal{O}}\left(\sqrt{H^2SA/T}\right).$$

The main part of the analysis is to bound the regret of the sampler and the forecaster, which can be done using standard techniques. We achieve

$$\mathfrak{R}^T_{\mathrm{Samp}}(\boldsymbol{d}^{\boldsymbol{\pi}^\star}) = \widetilde{\mathcal{O}}(\sqrt{H^4S^2AT}), \quad \mathfrak{R}^T_{\mathrm{Fore}}(\mathring{d}^T) = \widetilde{\mathcal{O}}(HSA).$$

In particular, the high regret of the sampler-player follows from the fact that the new reward functions are dependent on the previously generated samples, and thus the standard `UCBVI`-like analysis is not applicable, since $V^\star$ also becomes random. We will discuss this phenomenon more in the section devoted to adversarial reinforcement learning (see also Chapter 4).

Overall, the analysis above shows that `EntGame` achieves a sample complexity of $\widetilde{\mathcal{O}}(\frac{H^4S^2A}{\varepsilon^2})$ for best-policy identification in MVEE, which improves upon the previously known $\widetilde{\mathcal{O}}(1/\varepsilon^3)$ dependence on the desired accuracy. However, this analysis is not tight and can be improved further.

Faster algorithm for MVEE. To improve the sample complexity of `EntGame` further, we replace the sampler's problem with a regularized one. It is possible since the bound (1.25) still holds if we replace the definition of the sampler's regret with the following:

$$\mathfrak{R}^T_{\mathrm{Samp}}(\boldsymbol{d}) \triangleq \sum_{t=1}^{T}\left(\sum_{h,s,a}\left[d_h(s,a) - d_h^{\pi^t}(s,a)\right]\log\frac{1}{\bar{d}_h^t(s)} - \sum_{h,s}\left[d_h(s)\mathcal{H}(\pi_h(s)) - d_h^{\pi^t}(s)\mathcal{H}(\pi_h^t(s))\right]\right),$$

where $\boldsymbol{\pi} = (\pi_h)_{h\in[H]}$ is the policy that corresponds to $\boldsymbol{d}$, i.e., $\pi_h(a|s) = d_h(s,a)/d_h(s)$, and we define $d_h(s) \triangleq \sum_{a\in\mathcal{A}} d_h(s,a)$. This leads to an entropy-regularized problem with time-varying rewards, which can be solved by the `RL-Explore-Ent` algorithm (see Section 1.2.2), using a warm-up phase to estimate the transition model.

Overall, the resulting algorithm is called `RegEntGame` and it achieves a regret of $\mathfrak{R}^T_{\mathrm{Samp}}(\boldsymbol{d}^{\boldsymbol{\pi}^\star}) = \mathcal{O}(\varepsilon \cdot T)$ after $\widetilde{\mathcal{O}}(H^8S^4A/\varepsilon)$ episodes. This reduces both average regret terms of sampler and forecaster players to second-order terms with respect to the desired accuracy ε, resulting in $\widetilde{\mathcal{O}}(\frac{H^2SA}{\varepsilon^2})$ sample complexity for the maximum visitation entropy exploration problem.

Discussion. The main contribution of this section (and the corresponding part of Chapter 2) is a game-theoretic approach that achieves a $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity for the maximum visitation entropy exploration problem, which was an open question in the literature. We also show that visitation entropy maximization is *statistically simpler* than reward-free exploration: the former can be solved with a sample complexity of $\widetilde{\mathcal{O}}(H^2SA/\varepsilon^2)$, while the latter requires at least $\Omega(H^3S^2A/\varepsilon^2)$ samples. This raises the question of whether a more formal connection can be established between these two problems. Both are related to exploration without a reward signal, yet their sample complexities differ.

1.2.4 Perspectives

The results presented in this section open several avenues for future research, which we outline below.

Connection with Mean-Field Games. It has been established that concave-utility RL, of which MVEE is a special case, is equivalent to potential mean-field games (Geist et al. 2022). While our techniques are readily applicable to a broader class of concave-utility RL problems, particularly those that are strongly convex with respect to the trajectory entropy, a compelling open question remains: can this framework be extended to solve more general types of mean-field games, such as strongly monotone games (Yardim et al. 2025)?

Connection with Reward-Free Exploration. A significant gap remains in our understanding of the relationship between different exploration paradigms. Specifically, it is still unclear how to formally connect the solution to the MVEE problem with the solution to the standard reward-free exploration problem (Jin et al. 2020c; Kaufmann et al. 2021; Ménard et al. 2021). Clarifying this connection would provide deeper insights into the trade-offs between different exploration objectives.

Optimality of the Achieved Sample Complexity. Another critical direction is to determine whether the sample complexity of our algorithm is optimal. We conjecture that the current bounds for both MVEE and MTEE are not tight. It may be possible to achieve a sample complexity of $\widetilde{\mathcal{O}}(\mathrm{poly}(H,S,A)/\varepsilon)$ for the MVEE problem and $\widetilde{\mathcal{O}}(H^3SA/\varepsilon)$ for the MTEE problem. Proving this, however, would require developing new lower-bound techniques. The classic "needle-in-a-haystack" construction from Domingues et al. (2021a) is difficult to apply here, as the entropy objective fundamentally alters the problem structure.

State-wise versus State-Action-wise Entropy. Our analysis highlights that the acceleration for MVEE actively uses the entropy of actions. This raises a fundamental question: is there a statistical or computational separation between maximizing state-wise entropy and state-action-wise entropy? The latter might be inherently easier to optimize because it directly regularizes the policy, but a formal understanding of this distinction is still lacking.

1.3 Learning from Demonstrations (Summary of Chapter 3)

This section and Chapter 3 are based on: D. Tiapkin, D. Belomestny, D. Calandriello, É. Moulines, A. Naumov, P. Perrault, M. Valko, and P. Ménard, "Demonstration-Regularized RL", The Twelfth International Conference on Learning Representations (ICLR 2024) (Tiapkin et al. 2024a).

In this section, we address how to leverage expert demonstrations to accelerate reinforcement learning. While standard RL focuses on learning from scalar rewards, many modern applications provide this richer source of information. We use the post-training of large language models (LLMs) as a central motivating example, as it represents a high-impact domain where learning from demonstrations is critical.

A standard pipeline for LLM post-training (Stiennon et al. 2020; Ouyang et al. 2022) consists of three stages:

1. **Supervised Fine-Tuning (SFT):** A base model is fine-tuned on a high-quality dataset of expert demonstrations to learn an initial policy, $\boldsymbol{\pi}_{\mathrm{SFT}}$.
2. **Reward Modeling:** The SFT policy generates pairs of responses, which humans then label by preference. A reward model, $\boldsymbol{r}_{\mathrm{RM}}$, is trained on this preference data to predict which responses are better.

3. **Reinforcement Learning from Human Feedback (RLHF):** The policy is further refined using reinforcement learning. The learned reward model $\boldsymbol{r}_{\mathrm{RM}}$ provides the reward signal, and a KL-divergence penalty is added to keep the policy $\boldsymbol{\pi}_{\mathrm{RL}}$ close to the SFT policy $\boldsymbol{\pi}_{\mathrm{SFT}}$.

The use of KL-divergence in the final RLHF stage is a crucial heuristic. Practitioners introduced it primarily to combat *reward hacking*, where the model learns to exploit inaccuracies in the reward model to achieve high rewards without genuinely improving its capabilities.

Recent theoretical studies on RLHF confirm that some form of regularization is necessary to prevent reward hacking (Zhu et al. 2023). However, they often analyze the RLHF stage in isolation and overlook the preceding SFT stage. This stage is responsible for producing the very policy $\boldsymbol{\pi}_{\mathrm{SFT}}$ that serves as the anchor for the KL penalty. This omission leaves a critical question unanswered: Is KL-regularization merely a defensive mechanism against a flawed reward model, or does it play a more constructive role? Specifically, does it help to integrate knowledge from the initial demonstrations throughout the entire learning process? This section analyzes the full pipeline to provide a more complete theoretical understanding of demonstration-regularized reinforcement learning.

1.3.1 Problem Formulation

To isolate the role of demonstration data from the complexities of reward modeling, we analyze a simplified yet foundational problem: *best-policy identification (BPI) with demonstrations*. In this setting, we assume direct access to the true reward function $\boldsymbol{r}$, bypassing the reward modeling stage. The learning agent is provided with a static dataset of expert demonstrations and can interact with the environment to collect additional data.

Formally, the agent has access to:

- A dataset $\mathcal{D}_{\mathrm{E}}$ of N^{E} trajectories sampled from an expert policy $\boldsymbol{\pi}^{\mathrm{E}}$. We assume the expert policy is near-optimal with respect to the true reward function.
- An interactive learning environment where it can execute policies and observe outcomes.

The central research question is: *How can an agent leverage the offline expert data in $\mathcal{D}_{\mathrm{E}}$ to accelerate the online learning process?* Specifically, our goal is to design an algorithm that integrates information from the expert demonstrations to find an ε-optimal policy with significantly lower sample complexity than standard RL algorithms that learn from scratch.

The learning protocol combines offline and online phases. First, the agent can process the demonstration dataset $\mathcal{D}_{\mathrm{E}}$. Then, it enters an online interaction loop, which follows the standard BPI protocol:

- At each episode $t = 1, 2, \ldots$, the agent selects a policy $\boldsymbol{\pi}_t$ based on all information gathered so far, including $\mathcal{D}_{\mathrm{E}}$ and the history of online interactions.
- It executes $\boldsymbol{\pi}_t$ for one episode, collecting a new trajectory.
- The agent decides whether to continue exploring or to stop.

Upon stopping at episode $\mathfrak{t}$, the agent outputs a final policy π^{RL}. The objective is to ensure that this policy is ε-optimal with high probability, while minimizing the number of online interactions $\mathfrak{t}$.

We formalize this goal using the PAC framework. An algorithm, defined by the sequence of exploration policies $(\boldsymbol{\pi}_t)_{t\in\mathbb{N}}$, a stopping time $\mathfrak{t}$, and an output policy $\boldsymbol{\pi}^{\mathrm{RL}}$, is said to solve the BPI with demonstrations problem with sample complexity $\mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta)$ if for any given precision $\varepsilon > 0$ and confidence $1 - \delta$:

$$\mathbb{P}\Big[V_1^\star(s_1) - V_1^{\boldsymbol{\pi}^{\mathrm{RL}}}(s_1) \leqslant \varepsilon \quad \text{and} \quad \mathfrak{t} \leqslant \mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta)\Big] \geqslant 1 - \delta.$$

The sample complexity $\mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta)$ measures the number of online episodes required to find a near-optimal policy. We expect this complexity to decrease as the number of demonstrations N^{E} increases.

Also, we define the following notion of distance between two policies:

$$\mathrm{KL}_{\mathrm{traj}}(\boldsymbol{\pi}, \boldsymbol{\pi}') \triangleq \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{P}}\left[\sum_{h=1}^{H} \mathrm{KL}(\pi_h(S_h), \pi'_h(S_h))\right].$$

This KL-divergence is closely related to trajectory entropy (see (1.15)), as it serves as the Bregman divergence induced by trajectory entropy as a Bregman generator (Neu et al. 2017). We also introduce the following notation for the KL-regularized value function:

$$V^{\boldsymbol{\pi}}_{1,\lambda,\widetilde{\boldsymbol{\pi}}}(s_1) \triangleq V^{\boldsymbol{\pi}}_1(s_1) - \lambda \mathrm{KL}_{\mathrm{traj}}(\boldsymbol{\pi}, \widetilde{\boldsymbol{\pi}}),$$

which corresponds to the definition of the regularized value function (1.16) with $\Omega_h(s, \pi_h) = -\lambda\,\mathrm{KL}(\pi_h(s), \widetilde{\pi}_h(s))$.

1.3.2 Demonstration-regularized RL

To tackle BPI with demonstrations, we focus on a natural and simple approach that mirrors the LLM post-training pipeline. The analysis of this approach is one of the main contributions of this thesis, presented in Chapter 3.

First, we use the expert dataset $\mathcal{D}_{\mathrm{E}}$ to train a policy $\boldsymbol{\pi}^{\mathrm{BC}}$ via behavior cloning.[5] We assume that this policy imitates an ε_{E}-optimal expert policy $\boldsymbol{\pi}^{\mathrm{E}}$ up to a certain error $\varepsilon_{\mathrm{KL}}$ in terms of trajectory KL-divergence: $\mathrm{KL}_{\mathrm{traj}}(\boldsymbol{\pi}^{\mathrm{E}}, \boldsymbol{\pi}^{\mathrm{BC}}) \leqslant \varepsilon^2_{\mathrm{KL}}$. Second, we use $\boldsymbol{\pi}^{\mathrm{BC}}$ as a reference policy to regularize the BPI problem. Specifically, we solve the following *regularized* BPI problem up to an error $\varepsilon_{\mathrm{RL}}$:

$$V^{\star}_{1,\lambda,\boldsymbol{\pi}^{\mathrm{BC}}}(s_1) - V^{\boldsymbol{\pi}^{\mathrm{RL}}}_{1,\lambda,\boldsymbol{\pi}^{\mathrm{BC}}}(s_1) \leqslant \varepsilon_{\mathrm{RL}},$$

where $\lambda > 0$ is a regularization parameter and $\boldsymbol{\pi}^{\mathrm{RL}}$ is the resulting policy. This policy is then returned as the solution to the original BPI problem. We call this approach *demonstration-regularized RL*.

The performance of the final policy $\boldsymbol{\pi}^{\mathrm{RL}}$ can be bounded through a straightforward analysis. The suboptimality gap decomposes as follows:

$$\begin{aligned}
V^{\star}_1(s_1) - V^{\boldsymbol{\pi}^{\mathrm{RL}}}_1(s_1) &\leqslant \varepsilon_{\mathrm{E}} + V^{\boldsymbol{\pi}^{\mathrm{E}}}_1(s_1) - V^{\boldsymbol{\pi}^{\mathrm{RL}}}_1(s_1) \\
&= \varepsilon_{\mathrm{E}} + \Big(V^{\boldsymbol{\pi}^{\mathrm{E}}}_{1,\lambda,\boldsymbol{\pi}^{\mathrm{BC}}}(s_1) - V^{\boldsymbol{\pi}^{\mathrm{RL}}}_{1,\lambda,\boldsymbol{\pi}^{\mathrm{BC}}}(s_1)\Big) + \lambda\Big(\mathrm{KL}_{\mathrm{traj}}(\boldsymbol{\pi}^{\mathrm{E}}, \boldsymbol{\pi}^{\mathrm{BC}}) - \mathrm{KL}_{\mathrm{traj}}(\boldsymbol{\pi}^{\mathrm{RL}}, \boldsymbol{\pi}^{\mathrm{BC}})\Big) \\
&\leqslant \varepsilon_{\mathrm{E}} + \Big(V^{\star}_{1,\lambda,\boldsymbol{\pi}^{\mathrm{BC}}}(s_1) - V^{\boldsymbol{\pi}^{\mathrm{RL}}}_{1,\lambda,\boldsymbol{\pi}^{\mathrm{BC}}}(s_1)\Big) + \lambda \mathrm{KL}_{\mathrm{traj}}(\boldsymbol{\pi}^{\mathrm{E}}, \boldsymbol{\pi}^{\mathrm{BC}}) \\
&\leqslant \varepsilon_{\mathrm{E}} + \varepsilon_{\mathrm{RL}} + \lambda\varepsilon^2_{\mathrm{KL}}.
\end{aligned}$$

The first inequality uses the suboptimality of the expert, since by assumption, $\boldsymbol{\pi}^{\mathrm{E}}$ is ε_{E}-suboptimal. The second line follows by definition of the regularized value function. The third line uses the optimality of $V^{\star}_{1,\lambda,\boldsymbol{\pi}^{\mathrm{BC}}}$ and the non-negativity of KL-divergence. The final bound combines the assumptions on the behavior cloning and regularized RL steps.

This decomposition highlights how the final error accumulates from the three sources: the expert's suboptimality (ε_{E}), the behavior cloning error ($\varepsilon^2_{\mathrm{KL}}$), and the RL training error ($\varepsilon_{\mathrm{RL}}$). While regularization introduces the term $\lambda\varepsilon^2_{\mathrm{KL}}$, it makes the RL problem easier to solve. As established in the previous section, an $\varepsilon_{\mathrm{RL}}$-optimal policy for a λ-regularized MDP can be found with a sample complexity of $\widetilde{\mathcal{O}}(\mathrm{poly}(S, A, H)/(\varepsilon_{\mathrm{RL}} \cdot \lambda))$, for instance, with the `RL-Explore-Ent` algorithm. This is a significant improvement over the $\widetilde{\mathcal{O}}(1/\varepsilon^2_{\mathrm{RL}})$ dependency for standard BPI for large values of λ.

[5] We left a formal definition of behavior cloning for the next part of this subsection, it is not important for the current discussion.

To fully specify the sample complexity of demonstration-regularized RL, two key components require further analysis:

- **Behavior Cloning:** We need to establish statistical guarantees for the behavior cloning procedure to bound $\varepsilon_{\mathrm{KL}}$ as a function of the number of demonstrations N^{E}.
- **Regularized BPI:** The analysis of the regularized BPI problem might be refined to improve the dependence on S, A and H.

Behavior cloning. The most direct way to leverage expert demonstrations into a single policy is through *behavior cloning* (BC). This approach reframes imitation learning as a supervised learning problem, where the goal is to train a policy $\boldsymbol{\pi}^{\mathrm{BC}}$ that mimics the expert's actions from the demonstration dataset $\mathcal{D}_{\mathrm{E}}$. Formally, the policy is learned by minimizing the regularized negative log-likelihood of the expert's state-action pairs:

$$\boldsymbol{\pi}^{\mathrm{BC}} \in \arg\min_{\pi \in \mathcal{F}} \sum_{h=1}^{H} \left(\sum_{i=1}^{N^{\mathrm{E}}} \log \frac{1}{\pi_h(A_h^i | S_h^i)} + \mathcal{R}_h(\pi_h) \right),$$

where $(\tau_i = (S_1^i, A_1^i, \ldots, S_H^i, A_H^i))_{i=1}^{N^{\mathrm{E}}}$ are the expert trajectories in $\mathcal{D}_{\mathrm{E}}$, $\mathcal{R}_h$ is a regularizer, and $\mathcal{F}$ is a suitable class of policies.

For this procedure, it is possible to derive sharp statistical guarantees. Specifically, for finite and linear MDPs, we show that the learned policy $\boldsymbol{\pi}^{\mathrm{BC}}$ provably converges to the expert policy $\boldsymbol{\pi}^{\mathrm{E}}$ with the following rate on the trajectory KL-divergence:

$$\mathrm{KL}_{\mathrm{traj}}(\boldsymbol{\pi}^{\mathrm{E}}, \boldsymbol{\pi}^{\mathrm{BC}}) = \widetilde{\mathcal{O}}\left(\frac{SAH}{N^{\mathrm{E}}}\right).$$

This rate is minimax optimal up to logarithmic factors, which we show by proving a matching minimax lower bound. The proof uses a reduction from the expert policy estimation in KL-divergence to a hypothesis testing problem with finitely many hypotheses, following the approach of Tsybakov (2008, Chapter 2).

Our upper bound holds under a mild regularity assumption: the action probabilities under any policy in the class $\mathcal{F}$ must be bounded away from zero. The proof builds on standard techniques from statistical learning theory, with two key ingredients. The first is a careful management of the dependence on the minimal action probability, a term known to be unavoidable (Zhang 2006). Since this minimal probability can be as small as $1/N^{\mathrm{E}}$ in our setting, tight bounds are required. The second ingredient is the verification of the Bernstein's condition (Bartlett and Mendelson 2006) for KL-divergence, which allows for faster, $1/N^{\mathrm{E}}$-type convergence rates.

Regularized BPI problem. Next, we need to improve the sample complexity for the regularized BPI problem. Let $\overline{Q}$ and $\underline{Q}$ be optimistic and pessimistic estimates of the optimal Q-function $Q^{\star,\lambda}$, and let $\bar{\boldsymbol{\pi}}$ be the regularized greedy policy with respect to $\overline{Q}$. The bound on suboptimality (1.24) of $\bar{\boldsymbol{\pi}}$ can be upper bounded as:

$$V^{\star}_{1,\lambda,\widetilde{\boldsymbol{\pi}}}(s_1) - V^{\bar{\boldsymbol{\pi}}}_{1,\lambda,\widetilde{\boldsymbol{\pi}}}(s_1) \leqslant \frac{1}{2\lambda} \mathbb{E}_{\bar{\boldsymbol{\pi}}, \boldsymbol{P}} \left[\sum_{h=1}^{H} \left(\max_{a \in \mathcal{A}} \left(\overline{Q}_h(S_h, a) - \underline{Q}_h(S_h, a) \right) \right)^2 \right].$$

This bound highlights a critical challenge: the term inside the expectation involves a maximum over all actions, seeking out the most uncertain action where the gap $\overline{Q}_h - \underline{Q}_h$ is largest. However, the expectation is taken over trajectories generated by $\bar{\boldsymbol{\pi}}$, a policy that aims to converge to $\boldsymbol{\pi}^{\star}$ and is unlikely to sample these highly uncertain but potentially very suboptimal actions. A naive

approach of bounding the uncertainty $\overline{Q}_h - \underline{Q}_h$ uniformly over all reachable state-action pairs leads to a sample complexity of $\widetilde{\mathcal{O}}(H^8S^4A^2/(\varepsilon\lambda))$, which is highly suboptimal in the problem parameters despite the improved dependence on ε.

To overcome this issue, we must design an exploration policy that explicitly targets these uncertain actions. Instead of using $\bar{\boldsymbol{\pi}}$ for exploration, we construct a policy $\widehat{\boldsymbol{\pi}}$ that mixes $\bar{\boldsymbol{\pi}}$ with targeted exploration. Specifically, $\widehat{\boldsymbol{\pi}}$ is a uniform mixture of H policies $(\widehat{\boldsymbol{\pi}}^{(h')})_{h' \in [H]}$, where each $\widehat{\boldsymbol{\pi}}^{(h')}$ follows $\bar{\boldsymbol{\pi}}$ at all steps except h', at which it deterministically selects the most uncertain action:

$$\widehat{\boldsymbol{\pi}}_h^{(h')}(a|s) = \begin{cases} \bar{\boldsymbol{\pi}}_h(a|s) & \text{if } h \neq h' \\ \mathbb{1}\{a = \arg\max_{a' \in \mathcal{A}} \big(\overline{Q}_h(s,a') - \underline{Q}_h(s,a')\big)\} & \text{if } h = h' \,, \end{cases}$$

where ties in arg max-operation are resolved arbitrarily. This allows us to bound the suboptimality of $\bar{\boldsymbol{\pi}}$ in terms of the uncertainty of the *visited* state-action pairs:

$$V^{\star}_{1,\lambda,\widetilde{\boldsymbol{\pi}}}(s_1) - V^{\bar{\boldsymbol{\pi}}}_{1,\lambda,\widetilde{\boldsymbol{\pi}}}(s_1) \leqslant \frac{1}{2\lambda} \mathbb{E}_{\widehat{\boldsymbol{\pi}},\boldsymbol{P}} \left[\sum_{h=1}^{H} \big(\overline{Q}_h(S_h,A_h) - \underline{Q}_h(S_h,A_h)\big)^2 \right].$$

This new bound is easier to analyze, as each visit of state-action pair (S_h, A_h) reduced the uncertainty that is represented exactly by the term $(\overline{Q}_h(S_h,A_h) - \underline{Q}_h(S_h,A_h))^2$. While we use $\widehat{\boldsymbol{\pi}}$ for exploration, we return $\bar{\boldsymbol{\pi}}$ as the final policy for the BPI task.

This approach, which results in an algorithm referred to as `UCBVI-Ent+`, improves the sample complexity of the regularized BPI problem to $\widetilde{\mathcal{O}}(H^5S^2A/(\varepsilon\lambda))$. The S^2 dependence arises for reasons similar to the second-order term in the regret bound of `UCBVI`, which dominates what would otherwise be a first-order term. A similar effect has been observed in RL in metric spaces (Domingues et al. 2021b; Sinclair et al. 2023; Tiapkin et al. 2023b). We conjecture that the dependence on S could be improved to $\widetilde{\mathcal{O}}(S)$ by using a more refined analysis inspired by Zhang et al. (2024) or by employing a Q-learning-style algorithm (Jin et al. 2018; Zhang et al. 2020; Li et al. 2024). However, improving the dependence on the horizon H is less straightforward, as our bound accumulates an additional factor of H that is not present in standard analyses.

Final sample complexity. Combining the results from the preceding analyses of behavior cloning and regularized BPI yields the overall online sample complexity for our demonstration-regularized RL framework. By optimally selecting a value of the regularization parameter λ that balances the error from the behavior cloning stage with the error from the regularized BPI stage, we arrive at the following bound:

$$\mathcal{E}(\varepsilon, N^{\mathrm{E}}, \delta) = \widetilde{\mathcal{O}}\left(\frac{H^6S^3A^2}{\varepsilon^2 \cdot N^{\mathrm{E}}}\right).$$

The $1/N^{\mathrm{E}}$ term shows that the required number of online interactions decreases as more expert demonstrations are provided. This provides a principled theoretical justification for leveraging offline data to accelerate online learning.

However, this benefit is counterbalanced by a worse dependency on the problem parameters (H, S, and A) compared to the standard BPI sample complexity of $\widetilde{\mathcal{O}}(H^3SA/\varepsilon^2)$. Our approach becomes more sample-efficient only when a substantial number of demonstrations are available, specifically when $N^{\mathrm{E}} > \widetilde{\mathcal{O}}(H^3S^2A)$. This highlights that while demonstrations are helpful, a large amount of high-quality data is needed to overcome the analytical overhead of the pipeline.

1.3.3 Demonstration-regularized RLHF

Next, we extend the demonstration-regularized RL framework to the full post-training setup, where the true reward function is unknown and must be learned from human feedback. This setting, known

as Reinforcement Learning from Human Feedback (RLHF), is central to aligning large language models. Instead of observing rewards, the agent queries a preference oracle (e.g., a human labeler) that, given two trajectories, indicates a preference for one over the other.

This approach, that we call demonstration-regularized RLHF, integrates expert demonstrations into a three-stage pipeline that closely mirrors the standard practice for LLM alignment:

1. **Behavior Cloning (SFT):** First, we perform behavior cloning on the expert demonstration dataset $\mathcal{D}_{\mathrm{E}}$ to obtain an initial policy $\boldsymbol{\pi}^{\mathrm{BC}}$. This stage is analogous to supervised fine-tuning (SFT) in the LLM literature.

2. **Reward Modeling:** Second, we use the policy $\boldsymbol{\pi}^{\mathrm{BC}}$ to generate dataset of trajectories. These are presented to the preference oracle to build a preference dataset $\mathcal{D}_{\mathrm{RM}} = \{(\tau_0^k, \tau_1^k, o^k)\}_{k=1}^{N^{\mathrm{RM}}}$, where τ_0^k, τ_1^k are trajectories sampled with the policy $\boldsymbol{\pi}^{\mathrm{BC}}$ and o^k is the preference for τ_1^k over τ_0^k that follows a so-called *Bradley-Terry* model (Bradley and Terry 1952) widely used in the literature (Wirth et al. 2017; Saha et al. 2023):

$$\mathbb{P}[o^k = 1 | \tau_0^k, \tau_1^k] = \sigma(\boldsymbol{r}^\star(\tau_1^k) - \boldsymbol{r}^\star(\tau_0^k)),$$

   where σ is a sigmoid function, and $\boldsymbol{r}^\star$ is the true reward function. A reward model $\widehat{\boldsymbol{r}}$ is then trained on this dataset via maximum likelihood estimation (MLE) to capture the underlying preferences.

3. **Demonstration-Regularized RL:** Finally, we perform reinforcement learning using the learned reward model $\widehat{\boldsymbol{r}}$ as the objective. Crucially, we apply the same KL-regularization as in the previous section, penalizing the KL-divergence to the behavior-cloned policy $\boldsymbol{\pi}^{\mathrm{BC}}$. The final policy, $\boldsymbol{\pi}^{\mathrm{RLHF}}$, is the solution to this regularized BPI problem.

Now, essentially, we are going to study how regularization helps to avoid reward hacking; without regularization it is known that the final RLHF-policy trained on reward-proxy $\boldsymbol{\pi}^{\mathrm{RLHF}}$ can be highly suboptimal (Zhu et al. 2023). Before proceeding, we need to derive an analysis for reward modeling.

Reward modeling. Given the preference dataset $\mathcal{D}_{\mathrm{RM}} = \{(\tau_0^k, \tau_1^k, o^k)\}_{k=1}^{N^{\mathrm{RM}}}$, we learn a reward model $\widehat{\boldsymbol{r}}$ by maximizing the log-likelihood of the observed preferences. This is a standard maximum likelihood estimation (MLE) problem over a function class $\mathcal{G}$ for the trajectory reward functions:

$$\widehat{\boldsymbol{r}} \triangleq \arg\max_{\boldsymbol{r} \in \mathcal{G}} \sum_{k=1}^{N^{\mathrm{RM}}} o^k \log\left(\sigma\Big(\boldsymbol{r}(\tau_1^k) - \boldsymbol{r}(\tau_0^k)\Big)\right) + (1 - o^k) \log\left(1 - \sigma\Big(\boldsymbol{r}(\tau_1^k) - \boldsymbol{r}(\tau_0^k)\Big)\right).$$

The statistical properties of this estimator can be analyzed using techniques from M-estimation theory (Geer 2000) and PAC-Bayes analysis (Zhang 2006; Agarwal et al. 2020a). These tools allow us to derive the following high-probability bound on the variance of the estimation error:

$$\mathrm{Var}_{q^{\boldsymbol{\pi}^{\mathrm{BC}}}}[\boldsymbol{r}^\star(\tau_1) - \widehat{\boldsymbol{r}}(\tau_1)] \leqslant \frac{\zeta^2 d_{\mathcal{G}} \log(R_{\mathcal{G}} / N^{\mathrm{RM}}) + \zeta^2 \log(\mathrm{e}^2/\delta)}{N^{\mathrm{RM}}},$$

where $\mathrm{d}_{\mathcal{G}}$ is the *bracketing dimension* of the function class $\mathcal{G}$ (equals to SAH for finite MDPs), $R_{\mathcal{G}}$ is a related range factor, and $\zeta = 1/(\inf_{x \in [-H,H]} \sigma'(x))$ captures the non-linearity of the link function σ. For the sigmoid function, ζ scales as $\exp\{\Theta(H)\}$. This exponential dependence on the horizon is known to be unavoidable even in simpler linear settings (Hazan et al. 2014). Finally, expressing the bound in terms of variance correctly implies that the estimation is invariant under constant shifts in the reward function, as such shifts do not affect preferences nor the MLE procedure.

Final analysis. The analysis of the demonstration-regularized RLHF pipeline is more complex, as it must account for the uncertainty in the learned reward model $\widehat{\boldsymbol{r}}$. Our goal is to bound the suboptimality of the final policy, $\boldsymbol{\pi}^{\mathrm{RLHF}}$, with respect to the true reward function $\boldsymbol{r}^\star$. This bound aggregates the distinct sources of error from the behavior cloning, reward modeling, and reinforcement learning stages.

We assume the following conditions:

1. An expert policy $\boldsymbol{\pi}^{\mathrm{E}}$ that is ε_{E}-suboptimal with respect to the true reward $\boldsymbol{r}^\star$: $V_1^\star(s_1;\boldsymbol{r}^\star) - V_1^{\boldsymbol{\pi}^{\mathrm{E}}}(s_1;\boldsymbol{r}^\star) \leqslant \varepsilon_{\mathrm{E}}$.
2. A behavior-cloned policy $\boldsymbol{\pi}^{\mathrm{BC}}$ that approximates the expert policy with a trajectory KL-divergence error of at most $\varepsilon_{\mathrm{KL}}^2$: $\mathrm{KL}_{\mathrm{traj}}(\boldsymbol{\pi}^{\mathrm{E}}, \boldsymbol{\pi}^{\mathrm{BC}}) \leqslant \varepsilon_{\mathrm{KL}}^2$.
3. An estimated reward function $\widehat{\boldsymbol{r}}$ whose error is bounded by $\varepsilon_{\mathrm{RM}}$ in terms of standard deviation, evaluated over trajectories from $\boldsymbol{\pi}^{\mathrm{BC}}$: $\sqrt{\mathrm{Var}_{q^{\boldsymbol{\pi}^{\mathrm{BC}}}}[\boldsymbol{r}^\star(\tau) - \widehat{\boldsymbol{r}}(\tau)]} \leqslant \varepsilon_{\mathrm{RM}}$.

The final policy, $\boldsymbol{\pi}^{\mathrm{RLHF}}$, is obtained by solving the regularized BPI problem with the estimated reward $\widehat{\boldsymbol{r}}$ and the reference policy $\boldsymbol{\pi}^{\mathrm{BC}}$, achieving an $\varepsilon_{\mathrm{RL}}$ error:

$$V^\star_{1,\lambda,\boldsymbol{\pi}^{\mathrm{BC}}}(s_1;\widehat{\boldsymbol{r}}) - V^{\boldsymbol{\pi}^{\mathrm{RLHF}}}_{1,\lambda,\boldsymbol{\pi}^{\mathrm{BC}}}(s_1;\widehat{\boldsymbol{r}}) \leqslant \varepsilon_{\mathrm{RL}}\,.$$

Under these conditions, for a large enough regularization parameter $\lambda \geqslant H$, the suboptimality of the final policy $\boldsymbol{\pi}^{\mathrm{RLHF}}$ in the true environment is bounded as follows:

$$V_1^\star(s_1;\boldsymbol{r}^\star) - V_1^{\boldsymbol{\pi}^{\mathrm{RLHF}}}(s_1;\boldsymbol{r}^\star) \leqslant 3(\varepsilon_{\mathrm{E}} + \varepsilon_{\mathrm{RL}}) + \frac{9}{\lambda}\varepsilon_{\mathrm{RM}}^2 + (\lambda + 4H)\varepsilon_{\mathrm{KL}}^2\,.$$

This result shows how each error source contributes to the final performance. The regularization, controlled by λ, helps mitigating the impact of reward model uncertainty ($\varepsilon_{\mathrm{RM}}^2$) at the cost of introducing an additional dependence on the behavior cloning error ($\varepsilon_{\mathrm{KL}}^2$).

A key tool in our proof is an inequality that connects the difference in expected values under two distributions to their KL-divergence. This inequality, which is a consequence of the Donsker-Varadhan variational formula and a Bernstein-style bound on exponential moments (see, e.g., Boucheron et al. (2013) and Talebi and Maillard (2018)), states that for any two distributions p and q and a random variable X with range b:

$$\mathbb{E}_p[X] - \mathbb{E}_q[X] \leqslant \sqrt{2\mathrm{Var}_q[X]\cdot \mathrm{KL}(p\|q)} + \frac{2}{3}b\,\mathrm{KL}(p\|q)\,. \tag{1.26}$$

This analysis results in the sample complexity for the demonstration-regularized RLHF pipeline. For finite MDPs, provided the number of preference labels satisfies $N^{\mathrm{RM}} \geqslant \widetilde{\mathcal{O}}(\zeta SA/\varepsilon)$ and the number of expert demonstrations is at least $N^{\mathrm{E}} \geqslant \widetilde{\mathcal{O}}(H^2SA/\varepsilon)$, our framework finds an ε-optimal policy with an online sample complexity of

$$\mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta) = \widetilde{\mathcal{O}}\left(\frac{H^6S^3A^2}{N^{\mathrm{E}}\varepsilon^2}\right).$$

In summary, our approach matches the sample complexity of the standard BPI with demonstrations problem discussed in (1.3.2). However, achieving this guarantee requires sufficiently large datasets for both reward modeling and expert demonstrations to effectively address the reward hacking issue.

Discussion. Our analysis provides the first theoretical demonstration that the standard RLHF pipeline can mitigate reward hacking, provided that the supervised fine-tuning and reward modeling stages are supported by sufficiently large datasets. However, it is not clear whether our final bound

is tight or not: it is only known that the bound for behaviour cloning is tight since we have a corresponding lower bound.

Additionally, we want to underline that the condition on the expert dataset size $N^{\mathrm{E}} \geqslant \widetilde{\mathcal{O}}(H^2SA/\varepsilon)$ does not make our result redundant: the convergence in KL divergence does not transfer to the RL performance so easily (however, it is true in the case of deterministic expert policy, as we discuss in the complement to Chapter 3).

1.3.4 Perspectives

Overall, the results described above are first steps towards the understanding of the RLHF pipeline. However, there are still many open questions:

- **Regularized Best-Policy Identification:** We believe that there is some room for improvement for the upper bound for the regularized best-policy identification problem. Indeed, the current dependence on the number of states is suboptimal and can be improved by using Q-learning-type algorithms (Jin et al. 2018; Zhang et al. 2020; Li et al. 2024) or the improved analysis for the model-based approaches by Zhang et al. (2024). Additionally, there is no lower bound for this problem, which is a natural question to ask.

- **Better utilization of the expert demonstrations:** WWe wonder whether we do not use our expert demonstrations in a suboptimal way. Indeed, we only use the actions of the expert policy, while the transitions (S_h, A_h, S_{h+1}) are not used at all. Therefore, a natural question to ask is: can we use the transitions to improve the behavior cloning phase? However, the main challenge is an additional dependence between regularization and an estimate of the transition kernel.

- **χ^2 regularization:** Recently, Huang et al. (2025) proposed to use a χ^2 regularization instead of KL-regularization in the RL stage since it provides much stronger regularization signal. Indeed, compared to (1.26), the consideration of the χ^2-divergence is useful to derive an inequality between moments without a second-order term:

$$\mathbb{E}_p[X] - \mathbb{E}_q[X] \leqslant \sqrt{2\mathrm{Var}_q[X] \cdot \chi^2(p\|q)}\,.$$

  This is important as the absence of a second-order would allow us to achieve a much better bound on the final performance and avoid the need of a large number of expert demonstrations. Thus, this approach is very natural to combine with our theoretical framework; however, it is less clear how to establish convergence of the behavior cloning phase in χ^2 divergence since it is a much stronger notion of convergence.

1.4 Reinforcement Learning with Adversarial Rewards (Summary of Chapter 4)

This section and Chapter 4 are based on: D. Tiapkin, E. Chzhen, and G. Stoltz, "Narrowing the Gap between Adversarial and Stochastic MDPs via Policy Optimization", Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS 2025) (Tiapkin et al. 2025a).

Finally, we address the question of how to adapt reinforcement learning to the adversarial setting, where the reward function is chosen by an adversary. This setting is very natural given two applications:

- **Maximum Visitation Entropy Exploration:** We have already discussed the algorithm EntGame, where we need to design an algorithm that minimizes the regret given the adversarily chosen rewards;
- **Training the destruction process for sampling:** In the applications of RL to (diffusion) sampling, we have an additional degree of freedom: we can choose not only the way how to partially generate the next object, but also how to partially *destroy* some of the current objects. It was shown (Gritsaev et al. 2025a; Gritsaev et al. 2025b) that this additional degree of freedom can improve the final quality, and thus, we need to understand how to adapt RL algorithms to this setting.

1.4.1 Problem Formulation

The main difference between the standard RL setting and the adversarial setting is that the reward function can be changed at each step. We consider the setting of *oblivious* adversarial MDPs, where the adversary chooses the sequence of reward functions $(\boldsymbol{r}_t)_{t\geqslant 1}$ in advance.

Then, the interaction protocol is as follows:

1. Reset state $S_1^t = s_1$;
2. Start new episode — for each stage $h = 1, \dots, H$:
   - Pick a policy $\pi_{t,h}$ and sample an action $A_h^t \sim \pi_{t,h}(\cdot \mid S_h^t)$;
   - If $h \leqslant H-1$, move to the next state $S_{h+1}^t \sim P_h(\cdot \mid S_h^t, A_h^t)$;
3. Observe the reward function $\boldsymbol{r}_t = (r_{t,h})_{h\in[H]}$.

We compare the performance of the policies $\boldsymbol{\pi}_t = (\pi_{t,h})_{h\in[H]}$ picked to the one achieved by resorting to a static policy $\boldsymbol{\pi} = (\pi_h)_{h\in[H]}$ in each episode, in terms of value functions given the current reward $\boldsymbol{r}_t$. To underline this notation, we modify the notation for the value function:

$$V_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s) \triangleq \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{P}}\left[\sum_{j=h}^{H} r_{t,j}(S_j^t, A_j^t) \;\middle|\; S_h^t = s\right];$$

we recall the environment (reward functions, transition kernels) in the notation for value functions and expectations, as environments will vary in the algorithm and analysis.

The regret of the learner is defined as the difference between the accumulated value of the best static policy in hindsight and the gained value of the learner, that is,

$$\mathfrak{R}(T) \triangleq \max_{\boldsymbol{\pi}} \sum_{t=1}^{T} \left(V_1^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s_1) - V_1^{\boldsymbol{\pi}_t,\boldsymbol{r}_t,\boldsymbol{P}}(s_1)\right).$$

The goal of the learner is to design policies $(\boldsymbol{\pi}_t)_{t\in[T]}$ minimizing the above-defined regret. Our goal is to improve the existing upper bound on the regret bound by Rosenberg and Mansour (2019a) that is of order $\widetilde{\mathcal{O}}(\sqrt{\mathrm{poly}(H)S^2AT})$ to a bound of order $\widetilde{\mathcal{O}}(\sqrt{\mathrm{poly}(H)SAT})$, which is the same as the upper bound for the stochastic case, i.e., the case when there is only one reward function.

Other regret definitions. Also, we can consider other regret definitions, such as *dynamic regret*, where the regret is defined as the difference between the accumulated value of the best static policy in hindsight and the gained value of the learner, that is,

$$\mathfrak{R}_{\mathrm{dyn}}(T) \triangleq \sum_{t=1}^{T} \left(\max_{\boldsymbol{\pi}} V_1^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s_1) - V_1^{\boldsymbol{\pi}_t,\boldsymbol{r}_t,\boldsymbol{P}}(s_1)\right).$$

From the online learning literature, it is known that this type of regret is controllable only if the accumulated differences between $\boldsymbol{r}_t$ and $\boldsymbol{r}_{t+1}$ are small (Zhang et al. 2018); this is exactly the case for sampling problems (Gritsaev et al. 2025b).

1.4.2 Algorithm and Main Result

Next, we describe our algorithm and the main result, which will be presented in more detail in Chapter 4.

Preliminaries: Online Linear Optimization. As a preliminary to the study of online learning in MDPs, we introduce a basic concepts of online linear optimization.

At each round $t \geqslant 1$ and based on the past, a learning strategy $\varphi = (\varphi_t)_{t\geqslant 1}$ picks a convex combination $\boldsymbol{w}_t = (w_{t,1}, \dots, w_{t,K}) \in \Delta\Big([K]\Big)$ while an opponent player picks, possibly at random, a vector $\boldsymbol{g}_t = (g_{t,1}, \dots, g_{t,K})$ of signed rewards. Both $\boldsymbol{w}_t$ and $\boldsymbol{g}_t$ are revealed at the end of the round. By "based on the past", we mean, for the learning strategy, that $\boldsymbol{w}_t = \varphi_t\Big((\boldsymbol{g}_\tau)_{\tau\leqslant t-1}\Big)$. The initial vector $\boldsymbol{w}_1$ is constant.

Then, we say that a learning strategy φ controls the regret in the adversarial setting with rewards bounded by $M > 0$ if there exists a sequence $(B_{T,K})_{T\geqslant 1}$ with $B_{T,K}/T \to 0$ such that, against all opponent players sequentially picking reward vectors $\boldsymbol{g}_t$ in $[-M, M]^K$, for all $T \geqslant 1$,

$$\max_{k\in[K]} \sum_{t=1}^{T} g_{t,k} - \sum_{t=1}^{T} \sum_{j\in[K]} w_{t,j}\, g_{t,j} \leqslant 2M\, B_{T,K}\,.$$

The optimal orders of magnitude of $B_{T,K}$ are $\sqrt{T\ln(K)}$. We assume that the strategy may know M and rely on its value. Also, the strategy should work for any optimization horizon T (see the final "for all $T \geqslant 1$" in the definition above), the reason will be revealed later.

There exist several strategies meeting the requirements of the definition above. One prominent example is the multiplicative weights algorithm (Littlestone and Warmuth 1994), also known as `Hedge` (Freund and Schapire 1997) or Online Mirror Descent (Kivinen and Warmuth 1997). In particular, its version with a decaying learning rate (Auer et al. 2002) is designed to achieve a regret bound that holds uniformly for all $T \geqslant 1$. The algorithm updates the weights as follows:

$$w_{t,k} \propto \exp\left\{\eta_t \left(\sum_{\tau=1}^{t-1} g_{\tau,k} - \sum_{\tau=1}^{t-1}\sum_{j\in[K]} w_{\tau,j} g_{\tau,j}\right)\right\}, \qquad \eta_t = \frac{1}{M}\sqrt{\frac{\log(K)}{t}}\,.$$

This strategy achieves a regret bound of order $B_{T,K} = \sqrt{T\log(K)}$. Adaptive versions of this strategy, such as `AdaHedge`, have also been developed (Erven et al. 2011; Rooij et al. 2014; Orabona and Pál 2015). For another example, see Chapter 4.

Known transition kernel: `Adv2`. As a warm-up, we consider the case when the transition kernel is known. In this case, there exist multiple approaches (e.g., `REPS` by Zimin and Neu (2013)), but we will focus on the black-box approach `Adv2` of Jonckheere et al. (2025) that generalizes an observations first made by Shani et al. (2020).

The main idea of `Adv2` is to consider online learning in MDP as a set independent problems for each state and step $(s, h) \in \mathcal{S} \times [H]$, with an adversarial cost given by the advantage function:

$$A_h^{\boldsymbol{\pi}^t, \boldsymbol{r}_t, \boldsymbol{P}}(s, a) \triangleq Q_h^{\boldsymbol{\pi}^t, \boldsymbol{r}_t, \boldsymbol{P}}(s, a) - V_h^{\boldsymbol{\pi}^t, \boldsymbol{r}_t, \boldsymbol{P}}(s)\,.$$

Then, after each round, we re-compute the advantage function for a newly observed reward function and use it to update the policy by an online linear optimization [OLO] strategy $\varphi = (\varphi_t)_{t\geqslant 1}$:

$$\pi_h^t(\,\cdot \mid s) = \varphi_t\left(\left(A_h^{\boldsymbol{\pi}^\tau, \boldsymbol{r}^\tau, \boldsymbol{P}}(s, \,\cdot\,)\right)_{\tau\in[t-1]}\right).$$

In particular, one can use φ that corresponds to multiplicative weights update and get the following form

$$\pi_h^t(a \mid s) \propto \exp\left(\eta_t \sum_{\tau=1}^{t-1} A_h^{\boldsymbol{\pi}^\tau, \boldsymbol{r}^\tau, \boldsymbol{P}}(s,a)\right)$$

for an appropriate choice of the learning rate η_t. This update rule is interesting since it can be related to *natural policy gradient* method under softmax policy parameterization (Kakade 2001; Agarwal et al. 2021), which uses the following form of updates: $\pi_h^t(a|s) \propto \pi_h^{t-1}(a|s) \cdot \exp(\eta_t A_h^{\boldsymbol{\pi}^\tau, \boldsymbol{r}^\tau, \boldsymbol{P}}(s,a))$.

Given this strategy, we can achieve the following regret bound:

$$\mathfrak{R}(T) = \widetilde{\mathcal{O}}\left(H^2\sqrt{T \log A}\right).$$

Unknown transition kernel: `APO-MVP`. Next, we study how to adapt the `Adv2` algorithm to the case when the transition kernel is unknown. This is one of the main contributions of the current thesis and will be presented in more detail in Chapter 4.

When the transition kernel is unknown, we face two primary challenges:

- **Exploration:** The agent must efficiently explore the environment to learn the transition dynamics. We address this by incorporating exploration bonuses into the reward signal, a standard technique used in algorithms like `UCBVI` (Azar et al. 2017).
- **A non-stationary model:** The analysis of `Adv2` relies on a fixed, known transition kernel. In our setting, the estimated kernel $\widehat{\boldsymbol{P}}$ is continuously updated as more data is collected. These updates create a non-stationary model, which makes the analysis of Jonckheere et al. (2025) non-applicable.

To address the second challenge, we employ a *doubling trick*. Instead of updating the transition model after every episode, we group episodes into *epochs*. The model estimate is held fixed within an epoch and is only recomputed at the start of a new one. An epoch ends when the number of visits to any state-action pair doubles. This technique also was used in the analysis of Monotonic Value Propagation (`MVP`) by Zhang et al. (2024).

We now formalize the idea. The algorithm proceeds in a series of random epochs, $\mathcal{E}_e \subseteq [T]$. At the beginning of each epoch e, we compute an estimate of the transition kernel, $\widehat{\boldsymbol{P}}^{(e)}$, and a set of exploration bonuses, $\boldsymbol{b}^{(e)}$. Both are held constant for the duration of the epoch.

Within each epoch, the algorithm proceeds as follows. At each episode t, the agent selects a policy $\boldsymbol{\pi}^t$. At the end of the episode, after the reward function $\boldsymbol{r}_t$ is revealed, we construct optimistic estimates of the Q-value and value functions. For $h \in [H]$, with $\widehat{Q}_{H+1}^t \equiv 0$ and $\widehat{V}_{H+1}^t \equiv 0$:

$$\widehat{Q}_h^t(s,a) \triangleq r_{t,h}(s,a) + b_h^{(e_t)}(s,a) + \widehat{P}_h^{(e_t)}\widehat{V}_{h+1}^t(s,a), \qquad \widehat{V}_h^t(s) \triangleq \pi_h^t \widehat{Q}_h^t(s),$$

where $\widehat{V}_h^t$ is the value of the current policy $\boldsymbol{\pi}^t$ under the optimistic Q-function $\widehat{Q}_h^t$. The estimated advantage function is then $\widehat{A}_h^t(s,a) \triangleq \widehat{Q}_h^t(s,a) - \widehat{V}_h^t(s)$.

Finally, the policy for the next episode is updated. We maintain $S \times H$ independent instances of an Online Linear Optimization (OLO) strategy φ. For each state-space pair (s,h), the policy is updated based on the sequence of advantage functions observed within the current epoch:

$$\pi_h^{t+1}(\cdot \mid s) = \varphi_{t'}\left(\left(\widehat{A}_h^\tau(s,\cdot)\right)_{\tau \in \mathcal{E}_{e_t} \cap [t]}\right), \qquad t' = |\mathcal{E}_{e_t} \cap [t]|,$$

where t' is the number of episodes passed in the current epoch e_t. Note that this update only uses information available at the beginning of episode $t+1$.

Analysis Overview. Our analysis hinges on a regret decomposition that separates the problem into adversarial and stochastic components. This allows us to apply the most suitable and powerful tools to each part. Following standard techniques in optimistic algorithms (Auer et al. 2008; Azar et al. 2017), we let $\boldsymbol{\pi}^\star$ be the best fixed policy in hindsight and decompose the total regret as $\mathfrak{R}(T) = (\mathbf{A}) + (\mathbf{B}) + (\mathbf{C}) + (\mathbf{D})$, where:

$$(\mathbf{A}) \triangleq \sum_{t=1}^{T}\left(V_1^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s_1) - V_1^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t+\boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_1)\right) \qquad \text{(Optimism Gap)}$$

$$(\mathbf{B}) \triangleq \sum_{t=1}^{T}\left(V_1^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t+\boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_1) - V_1^{\boldsymbol{\pi}_t, \boldsymbol{r}_t+\boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_1)\right) \qquad \text{(Adversarial Regret)}$$

$$(\mathbf{C}) \triangleq \sum_{t=1}^{T}\left(V_1^{\boldsymbol{\pi}_t, \boldsymbol{r}_t+\boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_1) - V_1^{\boldsymbol{\pi}_t, \boldsymbol{r}_t+\boldsymbol{b}_t, \boldsymbol{P}}(s_1)\right) \qquad \text{(Model Error)}$$

$$(\mathbf{D}) \triangleq \sum_{t=1}^{T} V_1^{\boldsymbol{\pi}_t, \boldsymbol{b}_t, \boldsymbol{P}}(s_1) \qquad \text{(Exploration Cost)}$$

Here, $\widehat{\boldsymbol{P}}_t$ and $\boldsymbol{b}_t$ refer to the model estimate and exploration bonuses used during the epoch containing episode t.

The novelty of our work does not lie in the control of any single term, but in this decomposition between adversarial and stochastic parts. It allows us to overcome the challenges that have led prior analyses of adversarial MDPs to suboptimal regret bounds with an extra $\sqrt{S}$ factor.

$(\mathbf{B})$ is the core adversarial component. It measures the regret of our OLO strategies within a fixed, optimistic version of the MDP. Because the transition model $\widehat{\boldsymbol{P}}_t$ is held constant within each epoch, this term can be bounded using standard black-box guarantees for any suitable OLO algorithm, following Jonckheere et al. (2025), such as one based on exponential weights. Here the use of the doubling trick is crucial to ensure that the transition model is held constant within each epoch.

$(\mathbf{C})$ is the core stochastic component and the most technically involved. It quantifies the error incurred by our agent's policies due to inaccuracies in the estimated transition model. To bound this term, we adapt the powerful machinery developed by Zhang et al. (2024). Our use of the doubling trick is critical, as it aligns our algorithm's structure with the analytical requirements of their method. A crucial feature of the analysis from Zhang et al. (2024) is that it allows us to bound the model error term, $(\mathbf{C})$, directly, without it recursively depending on other error terms. This is a critical difference from the standard analysis of `UCBVI` (Azar et al. 2017). In that framework, the proof relies on the policy being greedy with respect to the optimistic value estimates. Since our policy $\boldsymbol{\pi}^t$ is chosen by an OLO strategy and is not necessarily optimistic with respect to the current Q-function, a standard analysis would fail, making the model error term uncontrollable.

Finally, $(\mathbf{A})$ and $(\mathbf{D})$ are controlled with more standard arguments. By construction, the exploration bonuses ensure that our optimistic value functions are an upper bound on the true optimal values. This makes $(\mathbf{A})$ non-positive with high probability. The total exploration cost, $(\mathbf{D})$, is bounded by a straightforward calculation.

All this analysis results in the following regret bound:

$$\mathfrak{R}(T) = \widetilde{\mathcal{O}}\left(\sqrt{H^7SAT}\right).$$

Why obliviousness is important. The assumption of an oblivious adversary is critical, and its importance is concentrated entirely in the analysis of the optimism gap, $(\mathbf{A})$. While the analysis of terms $(\mathbf{B})$, $(\mathbf{C})$, and $(\mathbf{D})$ is agnostic to the adversary's nature, ensuring optimism for $(\mathbf{A})$ requires

that the following concentration inequality holds with high probability:

$$[\widehat{P}_h^t - P_h]V_{h+1}^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s,a) + b_h^t(s,a) \geqslant 0\,.$$

For an oblivious adversary, the value function $V_{h+1}^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}$ is a deterministic vector, allowing this bound to hold with small exploration bonuses. Specifically, our choice of bonuses $b_h^t(s,a) \approx \sqrt{H/n_h^t(s,a)}$ is sufficient, and ensures that the exploration cost (**D**) scales as $\sqrt{S \cdot T}$ in the number of states and episodes.

However, in the case of an adaptive adversary, the reward function $\boldsymbol{r}_t$ can depend on the history, including the estimated model $\widehat{P}_h^t$. This makes $V_{h+1}^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}$ a random variable correlated with our model estimate, breaking standard concentration arguments. Consequently, controlling this term individually becomes impossible without introducing an additional $\sqrt{S}$ factor into the regret. We conjecture that this is a fundamental limitation and that a matching lower bound could be established by extending the constructions from reward-free exploration (Jin et al. 2020c) to the adaptive adversarial setting.

On dynamic regret. Our analytical framework also extends naturally to the stronger notion of *dynamic regret*, where the learner competes against a different optimal policy in each round. This adaptability comes from our regret decomposition, which isolates the stochastic model-learning components from the adversarial online-learning problem. The analysis of the stochastic terms ((**A**), (**C**), and (**D**)) remains largely unchanged, as they depend on the accuracy of the model estimation, not the competitor. The only structural change occurs in the adversarial regret term, (**B**). However, our approach already handles this by reducing the problem to a series of independent online learning tasks within each epoch:

$$(\mathbf{B}) = \sum_{e=1}^{m(T)} \mathrm{regret}(\{\boldsymbol{r}_t + \boldsymbol{b}_t\}_{t\in\mathcal{E}_e} \mid \text{fixed } \widehat{\boldsymbol{P}}^{(e)})\,,$$

where the internal regret are controled by the regret of adversarial MDP algorithm for a fixed transition kernel, such as `Adv2` (Jonckheere et al. 2025). This modularity gives us a strong belief that our techniques can be readily extended to the dynamic regret setting.

1.4.3 Perspectives

This work opens several interesting directions for future research:

- **Closing the gap between adversarial and stochastic MDPs:** We narrow the gap between the standard RL problem and its counterpart with oblivious adversarial rewards by showing that a $\widetilde{\mathcal{O}}(\sqrt{\mathrm{poly}(H)SAT})$ regret bound is achievable for the adversarial MDP problem. However, our dependence on H is highly suboptimal and far from the $\sqrt{H^3}$ dependence in the lower bound for the stochastic setting. There are several sources for this suboptimality:
  - *Lack of clipping:* Typically, RL methods for a fixed reward function apply an additional clipping step to the value function to ensure that the range of the optimistic value matches that of the true optimal value. However, the analysis of `Adv2` does not accommodate a clipping procedure for advantages because it heavily relies on performance-difference lemma simply does not hold for clipped values due to lack of linearity. A possible alternative is to study the contraction argument by Cassel and Rosenberg (2024).
  - *Suboptimal exploration bonuses:* We used simple Hoeffding-type exploration bonuses to simplify the arguments. Applying Bernstein-style bonuses should improve the regret bound by a factor of $\sqrt{H}$.

  - *Inherited suboptimality from* `Adv2` *analysis:* The analysis inherits a suboptimal dependence on H from the analysis of `Adv2`. Even with known transition kernels, the existing analysis yields an $H^2\sqrt{T \log A}$ bound, which is known to be suboptimal. Adapting other algorithms for online learning in known MDPs, such as `REPS` (Zimin and Neu 2013), which achieves an $H\sqrt{T \log(SA)}$ regret bound, could lead to improvements.

- **Bandit feedback:** Another important question is how to adapt our approach to the case of bandit feedback from the reward model. This question is non-trivial even for known transition kernels when using an `Adv2`-type procedure.

- **Adaptive adversary:** Finally, it would be interesting to derive a lower bound of order $\sqrt{H^3 S^2 AT}$ for the adaptive adversary setup, possibly by reusing the lower-bound techniques for reward-free exploration from Jin et al. (2020c).

1.5 Other Contributions (not covered in the thesis)

In addition to the works presented in this thesis, I have contributed to a range of topics in reinforcement learning throughout and prior to my PhD studies. For completeness, these related publications are listed below, grouped by topic; note that they are *not* included as part of the main thesis.

1.5.1 Posterior Sampling for Reinforcement Learning

The following group of work are related to the randomized exploration via posterior sampling.

In (Tiapkin et al. 2022a), we proposed a new approach to optimistic exploration in RL that generalizes the work of Kaufmann et al. (2012) to the reinforcement learning setting. The main technical tool is an anti-concentration inequality for the Dirichlet distribution with parameter $\alpha = (\alpha_0, \dots, \alpha_d)$ and a vector $f \in \mathbb{R}^d_+$, which bounds from below the probability of the event $\sum_{i=1}^d w_i f_i \geqslant \mu$. This bound exploits properties of the Dirichlet distribution, its connection to the Gamma distribution, and techniques from complex analysis such as the saddle-point method. In (Tiapkin et al. 2022b), we sharpened this approach to show that the optimistic posterior sampling algorithm achieves minimax-optimal regret guarantees even with a logarithmic number of samples from the posterior distribution, thus solving the open problem of Agrawal and Jia (2017). In (Tiapkin et al. 2023b), we proposed a model-free version of the optimistic posterior sampling algorithm that uses learning rate randomization. In (Belomestny et al. 2023), we provided sharp upper and lower bounds for the Dirichlet distribution and applied them to the analysis of multinomial Thompson sampling.

- **Daniil Tiapkin**, Denis Belomestny, Éric Moulines, Alexey Naumov, Sergey Samsonov, Yunhao Tang, Michal Valko, Pierre Ménard. *"From Dirichlet to Rubin: Optimistic Exploration in RL without Bonuses"*. Proceedings of the 39th International Conference on Machine Learning (ICML 2022).

- **Daniil Tiapkin**, Denis Belomestny, Daniele Calandriello, Éric Moulines, Rémi Munos, Alexey Naumov, Mark Rowland, Michal Valko, Pierre Ménard. *"Optimistic Posterior Sampling for Reinforcement Learning with Few Samples and Tight Guarantees"*. Advances in Neural Information Processing Systems (NeurIPS 2022).

- **Daniil Tiapkin**, Denis Belomestny, Daniele Calandriello, Éric Moulines, Rémi Munos, Alexey Naumov, Pierre Perrault, Michal Valko, Pierre Ménard. *"Model-free Posterior Sampling via Learning Rate Randomization"*. Advances in Neural Information Processing Systems (NeurIPS 2023).

- Denis Belomestny, Pierre Ménard, Alexey Naumov, **Daniil Tiapkin**, Michal Valko. *"Sharp Deviations Bounds for Dirichlet Weighted Sums with Application to Analysis of Bayesian Algorithms"*. arXiv preprint arXiv:2304.03056 (2023), authors are listed alphabetically.

1.5.2 Reinforcement Learning and Sampling

This line of work is directly related to the work in Chapter 2, in particular to the regularized RL problem. Bengio et al. (2021) introduced *generative flow networks* (GFlowNets) to learn approximate samplers for distributions over complex combinatorial objects, such as molecules. GFlowNets construct these objects incrementally by adding components, reducing the task to sampling terminal states in a deterministic acyclic MDP. The goal is to match the *forward distribution* of trajectories, generated by the construction policy, with the *backward distribution*, obtained by sampling from the target distribution and applying a destruction process.

The GFlowNet community initially viewed the sampling problem as fundamentally different from RL, believing that RL techniques required significant adaptation. In (Tiapkin et al. 2024b), we demonstrated that GFlowNets with a fixed destruction process can be formulated as entropy-regularized RL with a specific reward function based on the destruction process. This insight enabled direct application of RL algorithms, such as Monte-Carlo Tree Search (MCTS, Silver et al. 2018) (Morozov et al. 2024), to GFlowNets. Further research addressed learning the destruction process, which differentiates GFlowNets from standard entropy-regularized RL (Gritsaev et al. 2025b; Gritsaev et al. 2025a), and relaxing the acyclicity assumption of the underlying MDP (Morozov et al. 2025).

- **Daniil Tiapkin***, Nikita Morozov*, Alexey Naumov, Dmitry P Vetrov. *"Generative Flow Networks as Entropy-Regularized RL"*. Proceedings of The 27th International Conference on Artificial Intelligence and Statistics (AISTATS 2024)
  * indicates equal contribution between the first two authors. The main idea was proposed by myself; both of us contributed equally to the implementation and experiments.

- Nikita Morozov, **Daniil Tiapkin**, Sergey Samsonov, Alexey Naumov, Dmitry Vetrov. *"Improving GFlowNets with Monte Carlo Tree Search"*. ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling (SPIGM 2024).
  I proposed the use of MENTS as the base algorithm for the Monte Carlo Tree Search procedure.

- Timofei Gritsaev, Nikita Morozov, Sergey Samsonov, **Daniil Tiapkin**. *"Optimizing Backward Policies in GFlowNets via Trajectory Likelihood Maximization"*. The Thirteenth International Conference on Learning Representations (ICLR 2025).
  I served as the main senior author and mentor for Timofei, guiding the development of the main ideas and contributing to the writing.

- Nikita Morozov, Ian Maksimov, **Daniil Tiapkin**, Sergey Samsonov. *"Revisiting Non-Acyclic GFlowNets in Discrete Environments"*. Forty-second International Conference on Machine Learning (ICML 2025).
  I contributed to the development of the proofs and the writing.

- Timofei Gritsaev, Nikita Morozov, Kirill Tamogashev, **Daniil Tiapkin**, Sergey Samsonov, Alexey Naumov, Dmitry Vetrov, Nikolay Malkin. *"Adaptive Destruction Processes for Diffusion Samplers"*. arXiv preprint arXiv:2506.01541 (2025), under review.
  I contributed to the development of stabilization methods and provided help in the low-dimensional experiments.

1.5.3 Federated Reinforcement Learning

This line of work examines the key distinctions between federated RL and standard RL. We identified several important structural differences that highlight why the low heterogeneity assumption is crucial for federated RL. This requirement stands in contrast to the typical assumptions in federated learning. Our research addressed these differences from the perspective of exploration (Labbi et al. 2025a) and policy gradient methods (Labbi et al. 2025b).

- Safwan Labbi, **Daniil Tiapkin**, Lorenzo Mancini, Paul Mangold, Éric Moulines. *"Federated UCBVI: Communication-Efficient Federated Regret Minimization with Heterogeneous Agents"*. Proceedings of The 28th International Conference on Artificial Intelligence and Statistics (AISTATS 2025).
  I contributed to the development of the main ideas, writing, and proofs.

- Safwan Labbi, Paul Mangold, **Daniil Tiapkin**, Éric Moulines. *"On Global Convergence Rates for Federated Policy Gradient under Heterogeneous Environment"*. arXiv preprint arXiv:2505.23459 (2025), under review.
  I contributed to the development of the main ideas, writing, and proofs.

1.5.4 Other relevant works

I have also contributed to the following works, which are broadly related to reinforcement learning. Three of these (Scheid et al. 2024; Ocello et al. 2025; Tiapkin et al. 2025b) focus on game theory and the study of different equilibrium concepts: Stackelberg equilibrium in the principal-agent framework, mean-field Nash equilibrium in mean-field RL, and standard Nash equilibrium in the context of large language models. The work (Tiapkin et al. 2025b) is connected to demonstration-regularized RL (Chapter 3) through the shared RLHF framework. The fourth work (Samsonov et al. 2024) investigates temporal difference learning with linear function approximation, using methods from linear stochastic approximation.

- Antoine Scheid, **Daniil Tiapkin**, Etienne Boursier, Aymeric Capitaine, Éric Moulines, Michael Jordan, El-Mahdi El-Mhamdi, Alain Oliviero Durmus. *"Incentivized Learning in Principal-Agent Bandit Games"*. Proceedings of the 41st International Conference on Machine Learning (ICML 2024).
  I contributed to the writing and the development of the proofs.

- Antonio Ocello, **Daniil Tiapkin**, Lorenzo Mancini, Mathieu Lauriere, Éric Moulines. *"Finite-Sample Convergence Bounds for Trust Region Policy Optimization in Mean Field Games"*. Forty-second International Conference on Machine Learning (ICML 2025).
  I contributed to the theoretical analysis and writing.

- **Daniil Tiapkin**, Daniele Calandriello, Denis Belomestny, Éric Moulines, Alexey Naumov, Kashif Rasul, Michal Valko, Pierre Ménard. *"Accelerating Nash Learning from Human Feedback via Mirror Prox"*. arXiv preprint arXiv:2505.19731 (2025), under review.
  I developed the main ideas, proofs, and implementation.

- Sergey Samsonov, **Daniil Tiapkin**, Alexey Naumov, Éric Moulines. *"Improved High-Probability Bounds for the Temporal Difference Learning Algorithm via Exponential Stability"*. Proceedings of Thirty Seventh Conference on Learning Theory (COLT 2024).
  I contributed to the reinforcement learning aspects of the proofs and the writing.

Chapter 2

Fast Rates for Maximum Entropy Exploration

Contents

We address the challenge of exploration in reinforcement learning (RL) when the agent operates in an unknown environment with sparse or no rewards. In this chapter, we study the maximum entropy exploration problem of two different types. The first type is *visitation entropy maximization* previously considered by Hazan et al. (2019) in the discounted setting. For this type of exploration, we propose a game-theoretic algorithm that has $\widetilde{\mathcal{O}}(H^3S^2A/\varepsilon^2)$ sample complexity thus improving the ε-dependence upon existing results, where S is a number of states, A is a number of actions, H is an episode length, and ε is a desired accuracy. The second type of entropy we study is the *trajectory entropy*. This objective function is closely related to the entropy-regularized MDPs, and we propose a simple algorithm that has a sample complexity of order $\widetilde{\mathcal{O}}(\text{poly}(S, A, H)/\varepsilon)$. Interestingly, it is the first theoretical result in RL literature that establishes the potential statistical advantage of regularized MDPs for exploration. Finally, we apply developed regularization techniques to reduce sample complexity of visitation entropy maximization to $\widetilde{\mathcal{O}}(H^2SA/\varepsilon^2)$, yielding a statistical separation between maximum entropy exploration and reward-free exploration.

2.1 Introduction

In reinforcement learning (RL), an agent interacts with an environment aiming to maximize the sum of rewards returned by the environment. When the reward signal is very sparse or completely absent, the agent may experience long periods without any feedback. In these periods exploration is the main challenge.

This work studies the problem of efficient *exploration in the absence of rewards*. Approaches to solve this problem can be roughly cast into three main groups: The *bonus-based exploration* where the agent maximizes self-defined bonuses or intrinsic rewards collected along trajectory (Meyer and Wilson 1991; Oudeyer et al. 2007; Bellemare et al. 2016). Typically these bonuses are related to

the variances of error-signals from some auxiliary tasks, such as learning the transition probability distributions (Meyer and Wilson 1991; Chentanez et al. 2004; Pathak et al. 2017; Savas et al. 2020), learning the optimal value function for all the possible rewards (Jin et al. 2020c; Kaufmann et al. 2021; Ménard et al. 2021), learning random generated features (Burda et al. 2019). A second approach is the *goal-conditioned exploration* where the agent learns to navigate to self-assigned states (or goals). A common goal-selection rule for this class of algorithms is to select as goals the states at the frontier of the visited states (Lim and Auer 2012; Tarbouriech et al. 2020b; Ecoffet et al. 2021). Other selection-goal rules include reaching each state a fixed number of times (Tarbouriech et al. 2021) or going to states where the estimation error for the transition probabilities is large (Tarbouriech et al. 2020a). The third approach, which has received relatively less attention thus far, is the *maximum entropy exploration* (Hazan et al. 2019; Lee et al. 2020; Mutti and Restelli 2020; Mutti et al. 2021). This approach involves learning a policy that aims to achieve a visitation distribution over state-action pairs that is as uniform as possible. One specific application of this approach is in unsupervised pretraining, where it helps to obtain a better initial policy (Seo et al. 2021; Zhang et al. 2021a; Mutti et al. 2022). To achieve this goal, the approach focuses on maximizing entropy-like functionals, which is the main focus of our study.

In this chapter, we focus on environments modeled by an episodic, finite, reward-free Markov Decision Process (MDP) with S states, A actions, horizon H and step-dependent transitions. We consider two types of entropy: the *visitation entropy* and the *trajectory entropy*. The visitation entropy of a policy is defined as the sum of the entropies of the visitation distributions induced by the given policy at each step. The trajectory entropy of a policy is given by the entropy of a trajectory, generated when one follows the given policy and seen as one random variable on the corresponding path space. We study maximum entropy exploration under the (ε, δ)-PAC framework, that is, we want to learn, with probability $1-\delta$, a policy leading to ε-optimal maximum visitation or trajectory entropy and using as few as possible interactions with the environment.

Visitation entropy. Hazan et al. (2019) study maximum visitation entropy[1] exploration (MVEE) in the more general framework of convex MDPs where the agent wants to maximize a convex function of the visitation distribution. The authors in Hazan et al. (2019) propose to apply the Frank-Wolfe algorithm (Frank and Wolfe 1956) to a smoothed version of the visitation entropy. Their algorithm, `MaxEnt`[2], has a sample complexity of order[3] $\widetilde{\mathcal{O}}(H^4S^2A/\varepsilon^3+S^3/\varepsilon^6)$, that is, it needs to sample that number of trajectories in order to find an ε-optimal policy for MVEE. Later, Cheung (2019) obtains a better rate of order $\widetilde{\mathcal{O}}(H^4S^2A/\varepsilon^2 + H^3S/\varepsilon^3)$ for the `Toc-UCRL2` algorithm. Then, building on the ideas introduced by Abernethy and Wang (2017), Zahavy et al. (2021) reinterpret the `MaxEnt` algorithm as a method to compute the equilibrium of a particular game induced by the Legendre-Fenchel transform of the smoothed entropy. Using this new point of view, they propose the `MetaEnt` algorithm[4] with a sample complexity of order $\widetilde{\mathcal{O}}(H^4S^2A/(\delta^2\varepsilon^2) + H^3S/\varepsilon^3)$. In this chapter, building on the ideas by Grünwald and Dawid (2002), we draw a connection between MVEE and another game. In this game, a *forecaster-player tries to predict the state-action pairs visited by a sampler-player* who aims at *surprising the forecaster-player by visiting not well predicted state-action pairs*. We propose the `EntGame` algorithm that tackles MVEE by solving this prediction game. We prove that `EntGame` has a sample complexity of order $\widetilde{\mathcal{O}}(H^4S^2A/\varepsilon^2 + HSA/\varepsilon)$, thus

[1] Note that Hazan et al. (2019) consider a slightly different entropy; see Remark 2.3.

[2] In this chapter we refer to `MaxEnt` as the algorithm by Hazan et al. (2019) applied to the visitation entropy and not to the reverse entropy as initially proposed by the authors.

[3] We adapt rates from the γ-discounted setting by replacing $1/(1-\gamma)$ with H. To take into account step-dependent transitions we multiply the first order term by H^2.

[4] We call `MetaEnt` the specialization of their general Meta-algorithm to the special case of MVEE. Note that `MaxEnt`, `Toc-UCRL2`, `MetaEnt` could be seen as variations of the same algorithm. We use different names to distinguish, at least, the associated sample complexity.

improving over the previous rate in terms of its dependence of ε, see Table 2.1. The key technical point leading to this improvement is that, contrary to the previous algorithms, `EntGame` does not need to estimate accurately the visitation distribution of a policy at each iteration but only needs *one trajectory generated by following this policy*. Moreover, we propose `RegEntGame`, the regularized version of `EntGame`, that achieves sample complexity of order $\widetilde{\mathcal{O}}(H^2SA/\varepsilon^2)$, additionally improving the previous rates in S. The main technique behind this improvement is exploiting a strong connection between visitation entropy and regularization in MDPs. As a result, we have shown that *MVEE is statistically strictly simpler than reward-free exploration* (Jin et al. 2020c).

Trajectory entropy. The second problem we consider is maximum trajectory entropy exploration (MTEE). The entropy of paths of a (Markovian) stochastic process was first introduced in Ekroot and Cover (1993). Intuitively maximizing the trajectory entropy of an MDP minimizes the predictability of paths. Also there is a close connection between MTEE and regularized RL, a very popular approach in practical applications of RL.

Contrary to MVEE, the optimal policy for MTEE can easily be obtained by solving *entropy-regularized Bellman equations* with *entropy of the transition probabilities as rewards*. Leveraging this observation, one can proceed in a similar way as for the best policy identification[5] (BPI, Fiechter 1994). Precisely, we propose two algorithms. The first one, `UCBVI-Ent` is the simplest one and computes a policy by solving optimistic version of the aforementioned Bellman equations and using it to interact with the environment. The algorithm stops when an upper confidence bound on the difference between the maximum trajectory entropy and the trajectory entropy of the current policy is small enough. The second algorithm, `RL-Explore-Ent`, is an adaptation of the reward-free exploration by Jin et al. (2020c) to our setting. This algorithm has two phases. In the first phase, we compute a *preliminary exploration policy* which is then used to generate independent trajectories (data). In the second phase, a nearly optimal MTEE policy is obtained by solving the empirical Bellman equations with transitions estimated from the data collected in the first phase.

Interestingly, we prove that `RL-Explore-Ent` enjoys a sample complexity of order $\widetilde{\mathcal{O}}\Big(\frac{\mathrm{poly}(S,A,H)}{\varepsilon}\Big)$. The key technical ingredients to obtain such fast rate are exploitation of the *smoothness introduced by the regularization* and the use of the explicit form of the optimal policy.

Regularized MDP. Notably we can adapt[6] our algorithms for MTEE to best policy identification in regularized MDPs (Neu et al. 2017; Geist et al. 2019). Especially, we consider the same entropy-regularized MDPs and associated Bellman equations as in `Soft Q-learning` (Fox et al. 2016; Schulman et al. 2018; Haarnoja et al. 2017) or `SAC` (Haarnoja et al. 2018), see Remark 2.6. We show that a variation of `RL-Explore-Ent` has a sample complexity of order $\widetilde{\mathcal{O}}(\mathrm{poly}(S,A,H)/(\varepsilon\lambda))$ for BPI and reward-free exploration in regularized MDPs, where λ is the regularization parameter. In particular, it exhibits a *statistical separation between BPI in regularized MDP and BPI in the original MDP* since in this case the optimal sample complexity is of order $\widetilde{\mathcal{O}}(H^3SA/\varepsilon^2)$ (Ménard et al. 2021; Domingues et al. 2021a). Thus, our analysis shows that regularization is an effective way to trade-off bias for sample complexity. Additionally, we show how to use entropy regularization to obtain a theoretically faster version of `EntGame` algorithm.

We highlight our main contributions:

- We propose the `EntGame` algorithm for MVEE with a sample complexity of order $\widetilde{\mathcal{O}}(H^4S^2A/\varepsilon^2)$ thus significantly improving the existing complexity bounds for MVEE.
- We introduce the new MTEE setting for exploration and provide two algorithm: the `UCBVI-Ent` algorithm for MTEE with a sample complexity of order $\widetilde{\mathcal{O}}(H^3SA/\varepsilon^2)$, and `RL-Explore-Ent` with

[5] Where in this problem the goal is to identify the optimal policy of a given MDP (equipped with a reward function).
[6] That is replace the entropy of the transition probability by an arbitrary reward function.

Algorithm	Setting	Sample complexity
`MaxEnt` (Hazan et al. 2019)	MVEE	$\widetilde{\mathcal{O}}(H^4S^2A/\varepsilon^3+S^3/\varepsilon^6)$
`Toc-UCRL2` (Cheung 2019)	MVEE	$\widetilde{\mathcal{O}}(H^4S^2A/\varepsilon^2+H^3S/\varepsilon^3)$
`MetaEnt` (Zahavy et al. 2021)	MVEE	$\widetilde{\mathcal{O}}(H^4S^2A/(\delta^2\varepsilon^2)+H^3S/\varepsilon^3)$
`EntGame` (this chapter)	MVEE	$\widetilde{\mathcal{O}}(H^4S^2A/\varepsilon^2+HSA/\varepsilon)$
`RegEntGame` (this chapter)	MVEE	$\widetilde{\mathcal{O}}(H^2SA/\varepsilon^2+H^8S^4A/\varepsilon)$
`UCBVI-Ent` (this chapter)	MTEE	$\widetilde{\mathcal{O}}(H^3SA/\varepsilon^2 + H^3S^2A/\varepsilon)$
`RL-Explore-Ent` (this chapter)	MTEE	$\widetilde{\mathcal{O}}(H^8S^4A/\varepsilon)$

Table 2.1: Sample complexities for MVEE and MTEE. We convert rates from the γ-discounted setting by replacing $1/(1-\gamma)$ with H or from the infinite horizon setting by replacing the diameter with H. To take into account step-dependent transitions we multiply the first order term by H^2. (For `MetaEnt` since they do not precisely specify the cost for estimating a visitation distribution we use the same $1/\varepsilon^3$ term as for `Toc-UCRL2`.)

a sample complexity of order $\widetilde{\mathcal{O}}(\text{poly}(S,A,H)/\varepsilon)$. Up to our knowledge, this is the first time that a fast rate (in $1/\varepsilon$) is obtained thanks to regularization.

- We adapt `UCBVI-Ent` and `RL-Explore-Ent` to solve the entropy-regularized MDPs with a sample complexity of order $\widetilde{\mathcal{O}}(H^3SA/\varepsilon^2)$ and $\widetilde{\mathcal{O}}(\text{poly}(S,A,H)/(\lambda\varepsilon))$ correspondingly, where λ is the regularization parameter.
- We combine `EntGame` algorithm with regularization techniques, resulting in a new algorithm `RegEntGame`. This algorithm improves a sample complexity of `EntGame` to $\widetilde{\mathcal{O}}(H^2SA/\varepsilon^2)$ and shows statistical separation of MVEE from reward-free exploration.

2.2 Setting

We consider a finite episodic *reward-free* MDP $\mathcal{M} = \left(\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h\in[H]}, s_1\right)$, where $\mathcal{S}$ is the set of states of size S, $\mathcal{A}$ is the set of actions of size A, H is the number of steps in one episode, $p_h(s'|s,a)$ is the probability transition from state s to state s' by performing action a in step h. And s_1 is the fixed initial state.

Policy & value functions. A general policy $\pi = (\psi_h)_{h\in[H]}$ is a collection of function $\psi_h : (\mathcal{S}\times\mathcal{A}\times[0,1])^{h-1}\times\mathcal{S}\times[0,1] \to \mathcal{A}$ that maps an history $I_h = (s_1, a_1, u_1, \dots, s_{h-1}, a_{h-1}, u_{h-1})$ where u_h are i.i.d. uniformly distributed on the unit interval, a state s_h and an auxiliary independent uniformly distributed random variable u_h to an action $a_h = \psi_h(I_h, s_h, u_h)$. A policy is Markovian if the action depends only on the previous state and the auxiliary noise $a_h = \psi_h(s_h, u_h)$. In this case the policy can be represented as $\pi = (\pi_h)_{h\in[H]}$ a collection of mappings from states to probability distributions over actions $\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}}$ for all $h \in [H]$ where $a_h = \psi_h(s_h, u_h) \sim \pi_h(s_h)$. A *stochastic* policy $\pi = (\pi_h)_{h\in[H]}$ is a collection of mappings from states to probability distributions over actions $\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}}$ for all $h \in [H]$, where each π_h maps each state to a distribution over actions. Furthermore, $p_h f(s,a) \triangleq \mathbb{E}_{s'\sim p_h(\cdot|s,a)}[f(s')]$ denotes the expectation operator with respect to the transition probabilities p_h and $(\pi_h g)(s) \triangleq \pi_h g(s) \triangleq \mathbb{E}_{a\sim\pi_h(s)}[g(s,a)]$ denotes the composition with policy π at step h. Also, for any distribution over actions $\pi \in \Delta_{\mathcal{A}}$ define $\pi g(s) \triangleq \mathbb{E}_{a\sim\pi}[g(s,a)]$. The value functions of π, denoted by V_h^π, as well as the optimal value functions, denoted by $V_h^\star$ are

given by the Bellman respectively optimal Bellman equations

$$Q_h^\pi(s,a) = r_h(s,a) + p_h V_{h+1}^\pi(s,a) \qquad V_h^\pi(s) = \pi_h Q_h^\pi(s)$$
$$Q_h^\star(s,a) = r_h(s,a) + p_h V_{h+1}^\star(s,a) \qquad V_h^\star(s) = \max_a Q_h^\star(s,a)$$

where by definition, $V_{H+1}^\star \triangleq V_{H+1}^\pi \triangleq 0$.

Visitation distribution. The *state-action visitation distribution* of policy π at step h is denoted by d_h^π, where $d_h^\pi(s,a)$ is the probability of reaching the state-action pair (s,a) at step h after policy π.

Visitation polytopes. All the admissible collection of visitation distributions belong to the following polytope

$$\mathcal{K}_p \triangleq \Big\{ d = (d_h)_{h\in[H]} : \sum_{a\in\mathcal{A}} d_1(s,a) = \mathbb{1}\{s = s_1\} \ \forall s \in \mathcal{S}$$
$$\sum_{a\in\mathcal{A}} d_{h+1}(s,a) = \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} p_h(s|s',a') d_h(s',a') \ \forall s \in \mathcal{S}, \forall h \geqslant 1 \Big\}.$$

We also denote by $\mathcal{K}$ the set of collections of probability distributions over state-action pairs without the constraint to be a valid visitation distribution, that is

$$\mathcal{K} \triangleq \Big\{ d = (d_h)_{h\in[H]} : d_h(s,a) \geqslant 0, \ \forall (h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h(s,a) = 1, \ \forall h \in [H] \Big\}.$$

Trajectory distribution. We denote by $\mathcal{T} \triangleq (\mathcal{S}\times\mathcal{A})^H = \Big\{(s_1,a_1,\dots,s_H,a_H) : (s_h,a_h) \in \mathcal{S}\times\mathcal{A}, \forall h \in [H]\Big\}$ the set of all possible trajectories. The probability to generate the trajectory $m = (s_1,a_1,\dots,s_H,a_H)$ with he policy π is $q^\pi(m) \triangleq \pi(a_1|s_1)\prod_{h=2}^H p_{h-1}(s_h|s_{h-1},a_{h-1})\pi_h(a_h|s_h)$. Note that the visitation distribution at step h of policy π is a marginal of the trajectory distribution $d_h^\pi(s,a) = \mathbb{E}_{(s_1,a_1,\dots,s_H,a_H)\sim q^\pi}[\mathbb{1}\{(s,a) = (s_h,a_h)\}]$.

Counts and empirical transition probability. the number of times the state action-pair (s,a) was visited in step h in the first t episodes are $n_h^t(s,a) \triangleq \sum_{i=1}^t \mathbb{1}\{(s_h^i,a_h^i) = (s,a)\}$. Next, we define $n_h^t(s'|s,a) \triangleq \sum_{i=1}^t \mathbb{1}\Big\{(s_h^i,a_h^i,s_{h+1}^i) = (s,a,s')\Big\}$ the number of transitions from s to s' at step h. The empirical distribution is defined as $\widehat{p}_h^t(s'|s,a) = n_h^t(s'|s,a)/n_h^t(s,a)$ if $n_h^t(s,a) > 0$ and $\widehat{p}_h^t(s'|s,a) \triangleq 1/A$ for all $s' \in \mathcal{S}$ else.

2.3 Visitation Entropy

In this section we focus on maximizing the visitation entropy defined below.

Visitation entropy. We define the visitation entropy of a policy π denoted by $\mathcal{H}_{\text{visit}}(d^\pi)$ as the sum of the visitation distribution entropies at each steps

$$\mathcal{H}_{\text{visit}}(d^\pi) \triangleq \sum_{h=1}^H \mathcal{H}(d_h^\pi)\,.$$

We denote by $\pi^{\star,\text{VE}} \in \arg\max_\pi \mathcal{H}_{\text{visit}}(d^\pi)$ a policy that maximizes the visitation entropy.

Maximum visitation entropy exploration. In MVEE the agent interacts with the reward-free MDP $\mathcal{M}$ as follows. At the beginning of episode t, the agent picks a policy π^t based only on the transitions collected up to episode $t-1$. Then a new reward-free trajectory is sampled following the policy π^t and observed by the agent. At the end of each episode the agent can decide to stop collecting new data, according to a random stopping time $\mathfrak{t}$, the stopping rule, and outputs a (general) policy $\widehat{\pi}$ based on the observed transitions. An agent for MVEE is therefore made of a triplet $((\pi^t)_{t\in\mathbb{N}}, \mathfrak{t}, \widehat{\pi})$.

Definition 2.1. (PAC algorithm for MVEE) An algorithm $((\pi^t)_{t\in\mathbb{N}}, \mathfrak{t}, \widehat{\pi})$ is (ε, δ)-PAC for MVEE if

$$\mathbb{P}\Big(\mathcal{H}_{\text{visit}}\big(d^{\pi^\star,\text{VE}}\big) - \mathcal{H}_{\text{visit}}(d^{\widehat{\pi}}) \leqslant \varepsilon\Big) \geqslant 1-\delta.$$

Our goal is to design an algorithm that is (ε, δ)-PAC for MVEE with as sample complexity $\mathfrak{t}$ as small as possible.

2.3.1 MVEE by Solving Game

Following the general framework of Hazan et al. (2019) and Zahavy et al. (2021), it is possible to solve MVEE by applying the Frank-Wolfe algorithm to a smoothed version of the visitation entropy. Interestingly, Abernethy and Wang (2017) showed that this procedure is equivalent to computing the Nash equilibrium of a particular game induced by the Legendre-Fenchel transform of the smoothed entropy. In fact, as noted by Grünwald and Dawid (2002), there exists another game naturally linked to MVEE, stated next.

Prediction game. Maximum visitation entropy is the value of the following prediction game

$$\max_{d\in\mathcal{K}_p} \mathcal{H}_{\text{visit}}(d) = \max_{d\in\mathcal{K}_p} \min_{\bar{d}\in\mathcal{K}} \sum_{(h,s,a)} d_h(s,a) \log \frac{1}{\bar{d}_h(s,a)} = \min_{\bar{d}\in\mathcal{K}} \max_{d\in\mathcal{K}_p} \sum_{(h,s,a)} d_h(s,a) \log \frac{1}{\bar{d}_h(s,a)},$$

see Lemma D.1 in Appendix D for a proof. This game can be interpreted as follows. On the one hand, the min player, or forecaster player, tries to predict which state-action pairs the max player will visit to minimize $\mathrm{KL}(d_h, \bar{d}_h)$. On the other hand, the max player, or sampler player, is rewarded for visiting state-action pairs that the forecaster player did not predict correctly.

We now describe the algorithm `EntGame` for MVEE. In this algorithm, we let a forecaster player and a sampler player compete for T episodes long. Let us first define the two players.

Forecaster-player. As forecaster-player we use the Mixture-Forecaster for a logarithmic loss, see Section 9 in (Cesa-Bianchi and Lugosi 2006). Fix a prior count n_0 and their sum $t_0 = SAn_0$. The forecaster-player predicts at episode t the distributions $\bar{d}^t \in \mathcal{K}$ with $\bar{d}^t_h(s,a) = \bar{n}^{t-1}_h(s,a)/(t+t_0)$ where the pseudo counts are $\bar{n}^t_h(s,a) = n^t_h(s,a) + n_0$ and $n^t_h(s,a)$ the counts of state-action pairs visited by the sampler-player. Note that $\bar{d}^t_h$ can be seen as the posterior mean under a Dirichlet distribution prior $\mathcal{D}\mathrm{ir}(\{n_0\}_{(s,a)\in\mathcal{S}\times\mathcal{A}})$ on $\mathcal{S}\times\mathcal{A}$.

Sampler-player. As sampler-player we choose the optimistic best-response. Define the optimistic Bellman equations

$$\begin{aligned} \overline{Q}^t_h(s,a) &= \log \frac{1}{\bar{d}^{t+1}_h(s,a)} + \widehat{p}^t_h \overline{V}^t_{h+1}(s,a) + b^t_h(s,a), \\ \overline{V}^t_h(s) &= \mathrm{clip}\Big(\max_{a\in\mathcal{A}} \overline{Q}^t_h(s,a), 0, \log(t/n_0 + SA)H\Big), \end{aligned} \tag{2.1}$$

where $V_{H+1}^t = 0$ and b_h^t are some Hoeffding-like bonuses defined in (A.1) of Appendix A.2. The sampler player then plays $d^{\pi^{t+1}}$ where π^{t+1} is greedy with respect to the optimistic Q-values, that is, $\pi_h^{t+1}(s) \in \arg\max_{\pi \in \Delta_A} \pi \overline{Q}_h^t(s)$.

Sampling rule. At each episode t the policy π^t of the sampler-player is used as a sampling rule to generate a new trajectory.

Decision rule. After T episodes we output a non-Markovian policy $\widehat{\pi}$ defined as the mixture of the policies $\{\pi^t\}_{t\in[T]}$, that is, to obtain a trajectory from $\widehat{\pi}$ we first sample uniformly at random $t \in [T]$ and then follow the policy π^t. Note that the visitation distribution of $\widehat{\pi}$ is exactly the average $d^{\widehat{\pi}} = (1/T)\sum_{t\in[T]} d^{\pi^t}$.

Remark that the stopping rule of `EntGame` is deterministic and equals to $\mathfrak{t} = T$. The complete procedure is detailed in Algorithm 1.

Algorithm 1 `EntGame`

Input: Number of episodes T, prior counts n_0.
for $t \in [T]$ **do**
  # Forecaster-player
  Update pseudo counts $\bar{n}_h^{t-1}(s,a)$ and predict $\bar{d}_h^t(s,a)$.
  # Sampler-player
  Compute π^t by optimistic planning (2.1) with rewards $\log\left(1/\bar{d}_h^t(s,a)\right)$.
  # Sampling
  for $h \in [H]$ **do**
    Play $a_h^t \sim \pi_h^t(s_h^t)$
    Observe $s_{h+1}^t \sim p_h(s_h^t, a_h^t)$
  end for
  Update counts and transition estimates.
end for
Output $\widehat{\pi}$ the uniform mixture of $\{\pi^t\}_{t\in[T]}$.

Theorem 2.2. *Fix $\varepsilon > 0$, $\delta \in (0,1)$ and $n_0 = 1$. Then under the choice*

$$T = \widetilde{\mathcal{O}}\left(\frac{H^4S^2A}{\varepsilon^2} + \frac{HSA}{\varepsilon}\right)$$

the algorithm `EntGame` *is (ε,δ)-PAC. See Theorem A.7 in Appendix A.2 for a precise bound.*

Thus the sample complexity of `EntGame` is of order $\widetilde{\mathcal{O}}(H^4S^2A/\varepsilon^2)$. In particular, this result significantly improves over the previous rate for MTEE, see Table 2.1. Note that, by using Bernstein-like bonuses (Azar et al. 2017) instead of Hoeffding-like ones for the sampler-player would give a sample complexity of order $\widetilde{\mathcal{O}}(H^3S^2A/\varepsilon^2)$ saving one factor H. However, in the Section 2.5 we present a way to use regularization techniques to achieve a sample complexity of order $\widetilde{\mathcal{O}}(H^2SA/\varepsilon^2)$.

Space and time complexity. Since `EntGame` relies on a model-based algorithm for the sampler-player, its space complexity is of order $\mathcal{O}(HS^2A)$. Because of the value iteration performed by the sampler-player, the time-complexity of one iteration of `EntGame` is of order $\mathcal{O}(HS^2A)$.

Remark 2.3. Note that our definition of the visitation entropy slightly differs from the one considered by Hazan et al. (2019). Indeed, their definition, translated to the episodic setting, is the entropy of

the average of the visitation distributions which is an upper bound on the average of the entropies by concavity of the entropy

$$\mathcal{H}\left(\frac{1}{H}\sum_{h=1}^{H} d_h^\pi\right) \geqslant \frac{1}{H}\mathcal{H}_{\text{visit}}(d^\pi)\,.$$

Even if both definitions make sense in the episodic setting, we think ours is slightly more appropriate in the case of step-dependent transition probabilities. Indeed, in this case we want visitation distributions to be close to the uniform distribution over state-action pairs *for all steps*. Nevertheless, `EntGame` can be adapted to optimize the visitation entropy used in Hazan et al. (2019) simply by predicting $\bar{d}_h^t(s,a) = \sum_{h'=1}^{H} \bar{n}_{h'}^{t-1}(s,a)/(H(t+t_0))$ for the forecaster-player. We conjecture that the sample complexity of this adaptation of `EntGame` for the alternative entropy is again of order $\widetilde{\mathcal{O}}(HS^2A/\varepsilon^2)$.

Comparison with `MaxEnt` and `MetaEnt`. All three algorithms, `EntGame`, `MetaEnt` (Zahavy et al. 2021), `MaxEnt` (Hazan et al. 2019) rely on the same principle of computing, implicitly or explicitly, the equilibrium of a well chosen game and deduce from it an optimal policy for MVEE. One first difference between `EntGame` and its competitors lies in the choice of the game. While `MetaEnt`, `MaxEnt` consider the game induced by the Legendre-Fenchel conjugate of a smoothed visitation entropy (Zahavy et al. 2021), `EntGame` leverages the prediction game which looks more natural for MVEE. One advantage of using this game, is that it allows to avoid the need to smooth the visitation entropy because it is done implicitly by the forecaster-agent with the pseudo-counts. More importantly, `MaxEnt` and `MetaEnt` both needs to accurately estimate at each episode the visitation distributions of the sampler-player $d_h^{\pi^t}$, leading to an extra $1/\varepsilon^3$ term in the sample complexity. Whereas `EntGame` needs one trajectory from π^t since it only involves the estimation of the averages $1/T\sum_{t=1}^{T} d_h^{\pi^t}$.

2.4 Trajectory Entropy

In this section we focus on another type of entropy, the trajectory entropy, that can be efficiently maximized. The entropy of paths of a (Markovian) stochastic process is introduced by Ekroot and Cover (1993). It quantifies the randomness of realizations with fixed initial and final states. Later it was extended (Savas et al. 2020) to realizations that reach a certain set of states, rather than a fixed final state. This type of entropy is also closely related to the so-called entropy rate of a stochastic process.

Trajectory entropy. We define the trajectory entropy of a policy π as the entropy of a trajectory generated with the policy π

$$\mathcal{H}_{\text{traj}}(q^\pi) \triangleq \mathcal{H}(q^\pi) = \sum_{m\in\mathcal{T}} q^\pi(m)\log\frac{1}{q^\pi(m)}\,.$$

We denote by $\pi^{\star,\text{TE}} \in \arg\max_\pi \mathcal{H}_{\text{traj}}(q^\pi)$ a policy that maximizes the trajectory entropy.

Maximum trajectory entropy exploration. MTEE differs from MVEE only in the choice of entropy. In particular an algorithm $((\pi^t)_{t\in\mathbb{N}_+}, \mathbf{t}, \widehat{\pi})$ for MTEE is also a combination of a time dependent policy $(\pi^t)_{t\in\mathbb{N}_+}$, a stopping rule $\mathbf{t}$, and a decision rule $\widehat{\pi}$.

Definition 2.4. (PAC algortihm for MTEE) An algorithm $((\pi^t)_{t\in\mathbb{N}}, \mathbf{t}, \widehat{\pi})$ is (ε,δ)-PAC for MTEE if

$$\mathbb{P}\Big(\mathcal{H}_{\text{traj}}\Big(q^{\pi^{\star,\text{TE}}}\Big) - \mathcal{H}_{\text{traj}}(q^{\widehat{\pi}}) \leqslant \varepsilon\Big) \leqslant 1-\delta\,.$$

As noted by Eysenbach and Levine (2019), MTEE can also be connected to a prediction game. In this game, the forecaster-player aims to predict the whole trajectory that the sampler-player will generate. Remark that predicting the trajectory implies to predict, in particular, the visited state-action pairs but the reverse is not true in general [7]. We could then apply the same strategy as in Section 2.3 to solve MTEE. Nevertheless, for trajectory entropy, there is a more direct way to proceed.

Entropy regularized Bellman equations. One big difference between MVEE and MTEE is that the optimal policy can be obtained by solving regularized Bellman equations. Indeed, thanks to the chain rule for the entropy, the trajectory entropy of a policy π is $\mathcal{H}_{\mathrm{traj}}(d^\pi) = V_1^\pi(s_1)$ and the maximum trajectory entropy is $\mathcal{H}_{\mathrm{traj}}\left(d^{\pi^\star,\mathrm{TE}}\right) = V_1^\star(s_1)$ where the value functions V^π and $V^\star$ satisfy

$$\begin{aligned}
Q_h^\pi(s,a) &= \mathcal{H}\Big(p_h(s,a)\Big) + p_h V_{h+1}^\pi(s,a)\,,\\
V_h^\pi(s) &= \pi_h Q_h^\pi(s) + \mathcal{H}\Big(\pi_h(s)\Big)\,,\\
Q_h^\star(s,a) &= \mathcal{H}\Big(p_h(s,a)\Big) + p_h V_{h+1}^\star(s,a)\,,\\
V_h^\star(s) &= \max_{\pi\in\Delta_A}\{\pi Q_h^\star(s) + \mathcal{H}(\pi)\}\,,
\end{aligned}$$

where by definition, $V_{H+1}^\star \triangleq V_{H+1}^\pi \triangleq 0$. In particular, the maximum trajectory entropy policy is given by $\pi_h^{\star,\mathrm{TE}}(s) = \arg\max_{\pi\in\Delta_A}(\pi Q^\star(s) + \mathcal{H}(\pi))$. It can be computed explicitly via $\pi_h^{\star,\mathrm{TE}}(a|s) = \exp\Big(Q_h^\star(s,a) - V_h^\star(s)\Big)$ as well as the optimal value function $V_h^\star(s) = \log\Big(\sum_{a\in\mathcal{A}} \mathrm{e}^{Q_h^\star(s,a)}\Big)$. We refer to Appendix A.3 for a complete derivation.

We now describe our algorithm `RL-Explore-Ent`, the description of `UCBVI-Ent` is postponed to Appendix A.4. The idea of the algorithm is rather simple: since we need to solve regularized Bellman equations to obtain a maximum trajectory entropy policy, we can 1) find a *preliminary exploration policy* π^{mix} allowing one to construct estimates of the transition probabilities which are uniformly good when computing expectations of arbitrary bounded functions over all policies (see Lemma D.20), and 2) solve the regularized Bellman equations based on the estimated model. A similar approach is used in reward-free exploration (Jin et al. 2020c; Kaufmann et al. 2021; Ménard et al. 2021), and, in particular, our algorithm is close to `RF-RL-Explore` by Jin et al. (2020c). However, the key difference is that in the presence of regularization a much smaller number of transitions (trajectories) needs to be collected in order to obtain a high quality policy.

Exploration phase. This phase is devoted to learn a simple (non-Markovian) preliminary exploration policy π^{mix} that could be used to construct a accurate enough estimates of transition probabilities. This policy is obtained, as in `RF-RL-Explore`, by learning for each state s and step h, a policy that reliably reaches state s at step h. This can be done by running for N_0 iterations any regret minimization algorithm, e.g. `EULER` (Zanette and Brunskill 2019), for the sparse reward function putting reward one at state s at step h and zero otherwise. The policy π^{mix} is defined as the uniform mixture of the aforementioned policies. Then the policy π^{mix} is used to collect N fresh independent trajectories from the MDP.

Planning phase. For the planning phase, the agent builds a transition model

$$\widehat{p}_h(s'|s,a) = \begin{cases} \frac{n_h(s'|s,a)}{n_h(s,a)} & n_h(s,a) > 0\\ \frac{1}{S} & n_h(s,a) = 0 \end{cases}, \tag{2.2}$$

[7] Indeed d_h^π are only the marginals of q^π.

where $n_h(s,a)$ is the number of visits of the state-action pair (s,a) at step h for these N sampled trajectories. The final policy is a solution to the empirical regularized Bellman equations

$$\begin{aligned}\widehat{Q}^\star_h(s,a) &= \mathcal{H}\Big(\widehat{p}_h(s,a)\Big) + \widehat{p}_h \widehat{V}^\star_{h+1}(s,a)\,,\\ \widehat{V}^\star_h(s) &= \max_{\pi\in\Delta_A}\Big\{\pi\widehat{Q}^\star_h(s) + \mathcal{H}(\pi)\Big\}\,,\\ \widehat{\pi}_h(s) &= \arg\max_{\pi\in\Delta_A}\Big\{\pi\widehat{Q}^\star_h(s) + \mathcal{H}(\pi)\Big\}\,.\end{aligned} \tag{2.3}$$

The complete procedure is described in Algorithm 2. We now prove that, for the well-calibrated choice of N and N_0 of order $\widetilde{\mathcal{O}}(\mathrm{poly}(S,A,H)/\varepsilon)$, the `RL-Explore-Ent` algorithm is (ε,δ)-PAC for MTEE and provide an upper bound on its sample complexity. For the proof we refer to Appendix A.5.

Theorem 2.5. *The algorithm* `RL-Explore-Ent` *with parameters* $N_0 = \Omega\Big(\frac{H^7S^3A\cdot L^3}{\varepsilon}\Big)$ *and* $N = \Omega\Big(\frac{H^6S^3AL^5}{\varepsilon}\Big)$ *is* (ε,δ)*-PAC for the MTEE problem, where* $L = \log(SAH/(\varepsilon\delta))$*. Its total sample complexity* $SHN_0 + N$ *is bounded by*

$$\widetilde{\mathcal{O}}\left(\frac{H^8S^4A}{\varepsilon}\right).$$

Algorithm 2 `RL-Explore-Ent`

1: **Input:** Target precision ε, target probability δ, number of episodes for simple exploration policy N_0, number of sampled trajectories N.
2: **for** $(s',h') \in \mathcal{S}\times[H]$ **do**
3: Form rewards $r_h(s,a) = \mathbb{1}\{s=s', h=h'\}$.
4: Run `EULER` (Zanette and Brunskill 2019) with rewards r_h over N_0 episodes and collect all policies $\Pi_{s',h'}$.
5: Modify $\pi\in\Pi_{s',h'}:\ \pi_{h'}(a|s') = 1/A$ for all $a\in\mathcal{A}$.
6: **end for**
7: Construct a uniform mixture policy π^{mix} over all $\{\pi\in\Pi_{s,h} : (s,h)\in\mathcal{S}\times[H]\}$.
8: Sample N independent trajectories $(z_n)_{n\in[N]}$ following π^{mix} in the original MDP.
9: Construct from $(z_n)_{n\in[N]}$ an empirical model $\widehat{p}_h$ as in (2.2).
10: Output policy $\widehat{\pi}$ as a solution to (2.3).

Remark 2.6. (On solving regularized MDPs) Interestingly, our approach for MTEE can be adapted to solve entropy-regularized MDPs. For a reward functions $(r_h)_{h\in[H]}$ and regularization parameter $\lambda > 0$, consider the regularized Bellman equations

$$\begin{aligned}Q^\pi_{\lambda,h}(s,a) &= r_h(s,a) + p_h V^\pi_{\lambda,h+1}(s,a)\,,\\ V^\pi_{\lambda,h}(s) &= \pi_h Q^\pi_{\lambda,h}(s) + \lambda\mathcal{H}\Big(\pi_h(s)\Big)\,,\\ Q^\star_{\lambda,h}(s,a) &= r_h(s,a) + p_h V^\star_{\lambda,h+1}(s,a)\,,\\ V^\star_{\lambda,h}(s) &= \max_{\pi\in\Delta_A}\pi Q^\star_{\lambda,h}(s) + \lambda\mathcal{H}(\pi),\end{aligned}$$

where $V^\pi_{\lambda,H+1} = V^\star_{\lambda,H+1} = 0$. Note that these are the Bellman equations used by `Soft Q-learning` (Fox et al. 2016; Schulman et al. 2018; Haarnoja et al. 2017) and `SAC` (Haarnoja et al. 2018) algorithms. We are interested in the best policy identification for this regularized MDP. That is finding an algorithm that will output an ε-optimal policy $\widehat{\pi}$ such that with probability $1-\delta$, it

holds $V^{\star}_{\lambda,1}(s_1) - V^{\widehat{\pi}}_{\lambda,1}(s_1) \leqslant \varepsilon$ after a minimal number $\mathfrak{t}$ of trajectories sampled from the MDP $\mathcal{M}^r = (\mathcal{S}, \mathcal{A}, H, (p_h)_{h\in[H]}, (r_h)_{h\in[H]}, s_1)$. By using similar exploration and planning phases as in RL-Explore-Ent, we get an algorithm for BPI in the entropy-regularized MDP that also enjoys the fast rate of order $\widetilde{\mathcal{O}}\Big(H^8 S^4 A/(\varepsilon\lambda)\Big)$. Moreover, this algorithm could be used for more general types of regularization and even in reward-free setting with the same order of the sample complexity. Refer to Appendix A.4-A.5 for precise statements and proofs.

We observe that the sample complexity for solving the regularized MDP is strictly smaller[8] than the sample complexity for solving the original MDP. Indeed, one needs at least $\widetilde{\mathcal{O}}(H^3SA/\varepsilon^2)$ trajectory (Domingues et al. 2021a) to learn a policy π which value in the (unregularized) MDP is ε-close to the optimal value. Nevertheless, regularizing the MDP introduces a bias in the value function. Precisely we have for all π, $0 \leqslant V^{\pi}_{\lambda,1}(s_1) - V^{\pi}_1(s_1) \leqslant \widetilde{\mathcal{O}}(\lambda H)$ where $V^{\pi}_1(s_1)$ is the value function of π at the initial state and MDP $\mathcal{M}^r$. Thus, to solve BPI in $\mathcal{M}^r$ through BPI in the regularized MDP, one needs to take $\lambda = \widetilde{\mathcal{O}}(1/(H\varepsilon))$, leading to a sample complexity of order $\widetilde{\mathcal{O}}(H^9S^4A/\varepsilon^2)$. In particular, our fast rate for BPI in regularized MDP does not contradict the lower bound for BPI in the original MDP. However, our analysis shows that regularization is an effective way to trade-off bias for sample complexity.

Visitation entropy vs trajectory entropy. We can compare the visitation entropy and the trajectory entropy with

$$\mathcal{H}_{\mathrm{traj}}(q^\pi) \leqslant \underbrace{\mathrm{KL}(q^\pi, \otimes_{h=1}^H d_h^\pi) + \mathcal{H}_{\mathrm{traj}}(q^\pi)}_{\mathcal{H}_{\mathrm{visit}}(d^\pi)} \leqslant H\mathcal{H}_{\mathrm{traj}}(q^\pi)\,,$$

where $\otimes_{h=1}^H d_h^\pi$ is a product measure, see Lemma D.3 in Appendix D for a proof. Note also that in general the visitation distributions of an optimal policy for maximum trajectory entropy will be less 'spread' than the one of an optimal policy for MTEE, see Section 2.6 for an example. In particular one can prove that the optimal policy for MTEE is the uniform policy if the transitions are deterministic, see Lemma D.4 of Appendix D.

2.4.1 Proof Sketch

In this section we sketch the proof of Theorem A.21.

Properties of entropy. We start from analysing several properties of the entropy function. First, we notice that the well-known log-sum-exp function is a convex conjugate to the negative entropy defined only on the probability simplex (Boyd and Vandenberghe 2004)

$$F(x) \triangleq \log\left(\sum_{a\in\mathcal{A}} \mathrm{e}^{x_a}\right) = \max_{\pi\in\Delta_A} \langle \pi, x\rangle + \mathcal{H}(\pi)$$

and extend its action to Q-functions

$$F(Q)(s) \triangleq \max_{\pi\in\Delta_A} \pi Q(s) + \mathcal{H}(\pi).$$

This definition is useful because we can rewrite the optimal value function for MTEE in real and empirical model as follows

$$V^{\star}_h(s) = F(Q^{\star}_h)(s), \quad \widehat{V}^{\widehat{\pi}}_h(s) = F(\widehat{Q}^{\widehat{\pi}}_h)(s). \tag{2.4}$$

[8] For small enough ε.

Additionally, we notice that the gradient of F is equal to the soft-max policy that maximizes the expressions above

$$\pi_h^\star(s) = \nabla F(Q_h^\star)(s), \quad \widehat{\pi}_h(s) = \nabla F(\widehat{Q}_h^{\widehat{\pi}})(s),$$

and, moreover, since the negative entropy $-\mathcal{H}(\pi)$ is 1-strongly convex with respect to ℓ_1 norm, gradients of F is 1-Lipschitz with respect to ℓ_∞ norm by the properties of the convex conjugate (Kakade et al. 2009). Combining the gradient properties with the smoothness definition to $Q^\star$ and $\widehat{Q}^{\widehat{\pi}}$ we obtain

$$F(Q_h^\star)(s) \leqslant F(\widehat{Q}_h^{\widehat{\pi}})(s) + \widehat{\pi}_h\Big(Q_h^\star - \widehat{Q}_h^{\widehat{\pi}}\Big)(s) + \frac{1}{2}\|Q_h^\star - \widehat{Q}_h^{\widehat{\pi}}\|_\infty^2(s), \tag{2.5}$$

where $\|Q_h^\star - \widehat{Q}_h^{\widehat{\pi}}\|_\infty(s) = \max_{a\in\mathcal{A}} |Q_h^\star(s,a) - \widehat{Q}_h^{\widehat{\pi}}(s,a)|$.

Bound on the policy error. Next we apply the key inequality (2.5) to analyze the error between the optimal policy and policy $\widehat{\pi}$. Using (2.4) yields

$$V_h^\star(s) - V_h^{\widehat{\pi}}(s) \leqslant \widehat{V}_h^{\widehat{\pi}}(s) - V_h^{\widehat{\pi}}(s) + \widehat{\pi}_h\Big(Q_h^\star - \widehat{Q}_h^{\widehat{\pi}}\Big)(s) + \frac{1}{2}\max_{a\in\mathcal{A}}\Big(\widehat{Q}_h^{\widehat{\pi}}(s,a) - Q_h^\star(s,a)\Big)^2.$$

Next, by definition of $\widehat{\pi}$ we have $\widehat{V}_h^{\widehat{\pi}}(s) = \mathcal{H}(\widehat{\pi}_h(s)) + \widehat{\pi}_h\widehat{Q}_h^{\widehat{\pi}}(s)$, therefore by the regularized Bellman equations

$$V_h^\star(s) - V_h^{\widehat{\pi}}(s) \leqslant \widehat{\pi}_h p_h\Big(V_{h+1}^\star - V_{h+1}^{\widehat{\pi}}\Big)(s) + \frac{1}{2}\max_{a\in\mathcal{A}}\Big(\widehat{Q}_h^{\widehat{\pi}}(s,a) - Q_h^\star(s,a)\Big)^2.$$

Finally, rolling out this expression we have

$$V_1^\star(s_1) - V_1^{\widehat{\pi}}(s_1) \leqslant \frac{1}{2}\mathbb{E}_{\widehat{\pi}}\left[\sum_{h=1}^H \max_{a\in\mathcal{A}}\Big(\widehat{Q}_h^{\widehat{\pi}} - Q_h^\star\Big)^2(s_h,a)\right].$$

Next we may notice that in the generative model setting[9] there is available results that tells us that $\widetilde{\mathcal{O}}(1/\varepsilon)$ samples are enough to obtain $\|\widehat{Q}_h^{\widehat{\pi}} - Q_h^\star\|_\infty \lesssim \sqrt{\varepsilon}$ (Azar et al. 2013), and we can conclude the statement. However, in the online setup the situation is more complicated, and we apply reward-free techniques developed by Jin et al. (2020c) to obtain a "surrogate" of the generative model.

2.5 Faster Rates for Visitation Entropy

In this section, we show how to combine the regularization techniques developed in Section 2.4 with `EntGame` algorithm presented in Section 2.3.1.

The new algorithm `RegEntGame` is based on exactly the same game-theoretical framework as `EntGame`, but uses a *regularized sampler player* instead of the usual one.

Regularized sampler-player. For the sampler player, we shall take advantage of strong convexity of the visitation entropy. Beforehand, we construct an estimate of the model $\{\widehat{p}_h\}_{h\in[H]}$ by reward-free exploration, using HSN_0 samples to compute a policy π^{mix} and N samples to estimate transitions. Next, define the empirical regularized Bellman equations

$$\widehat{Q}_h^t(s,a) = \log\left(\frac{1}{\widehat{d}_h^{t+1}(s)}\right) + \widehat{p}_h\widehat{V}_{h+1}^t(s,a)\,, \qquad \widehat{V}_h^t(s) = \max_{\pi\in\Delta_\mathcal{A}}\{\pi\widehat{Q}_h^t(s) + \mathcal{H}(\pi)\}\,,$$

where $\widehat{V}_{H+1}^t \equiv 0$. The sampler player then follows the distribution $d^{\pi^{t+1}}$ where π^{t+1} is greedy with respect to the regularized empirical Q-values, that is, $\pi_h^{t+1}(s) \in \arg\max_{\pi\in\Delta_A}\{\pi\widehat{Q}_h^t(s) + \mathcal{H}(\pi)\}$.

[9] When there is a sampling oracle for each state-action pair.

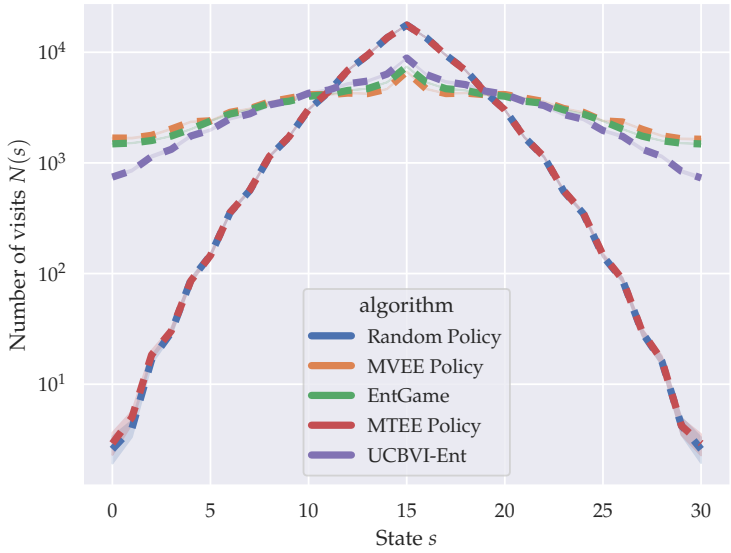

Figure 2.1: Number of state visits for $N = 100000$ samples in Double Chain MDP with $S = 31$ states, $A = 2$ actions, a horizon $H = 20$ and a 0.1 probability of moving to the opposite direction.

Theorem 2.7. *Fix $\varepsilon > 0$ and $\delta \in (0, 1)$. For $n_0 = 1$,*

$$N_0 = \Omega\left(\frac{H^7 S^3 A \cdot L^3}{\varepsilon}\right), \quad T = \Omega\left(\frac{H^3 SAL^3}{\varepsilon^2} + \frac{H^2 S^2 A^2 L^2}{\varepsilon}\right).$$

with $L = \log(SAH/\delta\varepsilon)$ the algorithm `RegEntGame` *is (ε, δ)-PAC. The total sample complexity is equal to $SH \cdot N_0 + N + T$, that is,*

$$\mathfrak{t} = \widetilde{\mathcal{O}}\left(\frac{H^2 SA}{\varepsilon^2} + \frac{H^8 S^4 A}{\varepsilon}\right).$$

Thus, the sample complexity of `RegEntGame` is of order $\widetilde{\mathcal{O}}(H^2 SA/\varepsilon^2)$ for ε large enough. In particular, this result significantly improves over the previous rates for MVEE, see Table 2.1. Moreover, this result shows a rate separation between reward-free exploration (Jin et al. 2020c), where the established lower bound on sample complexity $\Omega(H^3 S^2 A/\varepsilon^2)$ scales with S^2, and the visitation entropy maximization problem.

Proof idea. The main proof idea is to exploit not just strong convexity of the visitation entropy with respect to Euclidean distance but its strong convexity *with respect to trajectory entropy* (Bauschke et al. 2017). It allows us to use entropy regularization as described in Section 2.4.1 for the sampler player resulting in an averaged regret less than ε for only $\widetilde{\mathcal{O}}(\text{poly}(S, A, H)/\varepsilon)$ samples. Thus, the density estimation error becomes the leading term in the full error decomposition. For more details refer to Appendix A.6.

2.6 Experiments

In this section we report experimental results on simple tabular MDP for presented algorithms and show the difference between visitation and trajectory entropies. In particular, we compare `EntGame` and `UCBVI-Ent` algorithms with (a) random agent that takes all actions uniformly at random, (b)

an optimal MVEE policy computed by solving the convex program, and (c) an optimal MTEE policy computed by solving the regularized Bellman equations. As an MDP, we choose a stochastic environment called Double Chain as considered by Kaufmann et al. (2021).

Since the transition kernel for this environment is stage-homogeneous, for `EntGame` and `UCBVI-Ent` algorithms we joint counters over the different steps h. In particular, it changes the objective of the `EntGame` algorithm to the objective considered by Hazan et al. (2019) that makes more sense in the stage-independent setting [10].

In Figure 2.1 we present the number of state visits for our algorithms and baselines during $N = 100000$ interactions with the environment. For `UCBVI-Ent` algorithm the procedure was separated on two stages: at first we learn MDP with N-sample budget and extract the final policy, and then plot the number of visits for the final policy during another N samples.

In particular, we see that since this environment is almost deterministic the optimal MTEE policy is almost coincides with a random policy. Notably, the policy induced by `UCBVI-Ent` is more uniform over states because of $1/\sqrt{n^t(s,a)}$ bonuses, that make our algorithm close to `RF-UCRL` Kaufmann et al. 2021. Also we see that the optimal MVEE policy is the most uniform over states, that makes it an appropriate target for the pure exploration problem. For more details and additional experiments we refer to Appendix A.7.

2.7 Conclusion

In this chapter we studied MVEE for which we provided the `EntGame` algorithm with a sampling complexity significantly smaller than the existing complexity rates. We also introduced the MTEE problem where the optimal policy can be found using the dynamic programming. We proposed the `UCBVI-Ent` and `RL-Explore-Ent` algorithms for MTEE that can be adapted to BPI in regularized MDPs. We proved that, in both cases, `RL-Explore-Ent` and its variant enjoy a fast rate. In particular, we observed a statistical separation between BPI in regularized MDP and in the original MDP. Moreover, we show that the regularized version of `EntGame` called `RegEntGame` enjoys $\widetilde{\mathcal{O}}(H^2SA/\varepsilon^2)$ rates, making dependence in S smaller than in the reward-free exploration problem Jin et al. 2020c.

This work opens the following interesting future research directions:

Optimal rates for MVEE and MTEE. We are still lacking lower bounds for MTEE and MVEE problems enabling us to determine the optimal rates, especially what the number of states S and the horizon H is concerned. Note that one cannot apply directly the usual lower-bounds techniques for these two problems because of the entropy regularization in both cases. In particular, we conjecture that the optimal rate for MVEE is also of order $\widetilde{\mathcal{O}}(\mathrm{poly}(H,S,A)/\varepsilon)$.

Optimal rate for entropy-regularized RL. It would be interesting to obtain the optimal rate for BPI in a regularized MDP. In particular to recover the optimal rate for BPI in the original MDP by tuning the regularization parameter λ. We conjecture that the optimal rate is $\widetilde{\mathcal{O}}(H^2SA/(\lambda\varepsilon))$ for BPI in entropy-regularized MDP.

Other types of entropies. Our methodology can be applied to other types of entropies and even to other regularization penalties. One interesting case would be the goal-conditioned trajectory entropy (see Savas et al. 2020) where one considers only process realizations that reach a certain set of states at time H. This entropy can be applied to goal-conditioned RL. Another type of problem that could be of high interest is visitation entropy maximization under safety constraints (Yang and Spaan 2023).

[10] See Remark 2.3.

Chapter 3

Demonstration-Regularized RL

Contents

Incorporating expert demonstrations has empirically helped to improve the sample efficiency of reinforcement learning (RL). This chapter quantifies theoretically to what extent this extra information reduces RL's sample complexity. In particular, we study the demonstration-regularized reinforcement learning that leverages the expert demonstrations by KL-regularization for a policy learned by behavior cloning. Our findings reveal that using N^{E} expert demonstrations enables the identification of an optimal policy at a sample complexity of order $\widetilde{\mathcal{O}}(\mathrm{Poly}(S,A,H)/(\varepsilon^2 N^{\mathrm{E}}))$ in finite and $\widetilde{\mathcal{O}}(\mathrm{Poly}(d,H)/(\varepsilon^2 N^{\mathrm{E}}))$ in linear Markov decision processes, where ε is the target precision, H the horizon, A the number of action, S the number of states in the finite case and d the dimension of the feature space in the linear case. As a by-product, we provide tight convergence guarantees for the behavior cloning procedure under general assumptions on the policy classes. Additionally, we establish that demonstration-regularized methods are provably efficient for reinforcement learning from human feedback (RLHF). In this respect, we provide theoretical evidence showing the benefits of KL-regularization for RLHF in tabular and linear MDPs. Interestingly, we avoid pessimism injection by employing computationally feasible regularization to handle reward estimation uncertainty, thus setting our approach apart from the prior works.

3.1 Introduction

In reinforcement learning (RL, Sutton, Barto, et al. 1998), agents interact with an environment to maximize the cumulative reward they collect. While RL has shown remarkable success in mastering complex games (Mnih et al. 2015; Silver et al. 2018; OpenAI et al. 2019), controlling physical systems (Degrave et al. 2022), and enhancing computer science algorithms (Mankowitz et al. 2023), it does face several challenges. In particular, RL algorithms suffer from a large sample complexity, which is a hindrance in scenarios where simulations are impractical and struggle in environments with sparse rewards (Goecks et al. 2020).

A remedy found to handle these limitations is to incorporate information from a pre-collected offline dataset in the learning process. Specifically, leveraging demonstrations from experts, which

are trajectories without rewards, has proven highly effective in reducing sample complexity, especially in fields like robotics (Zhu et al. 2018; Nair et al. 2021) and guiding exploration (Nair et al. 2018; Aytar et al. 2018; Goecks et al. 2020).

However, from a theoretical perspective, little is known about the impact of this approach. Previous research has often focused on either offline RL (Rashidinejad et al. 2021; Xie et al. 2021; Yin et al. 2021; Shi et al. 2022) or online RL (Jaksch et al. 2010; Azar et al. 2017; Fruit et al. 2018; Dann et al. 2017; Zanette and Brunskill 2019; Jin et al. 2018). In this study, we aim to quantify how prior demonstrations from experts influence the sample complexity of various RL tasks, specifically two scenarios: best policy identification (BPI, Domingues et al. 2021a; Al Marjani et al. 2021) and reinforcement learning from human feedback (RLHF), within the context of finite or linear Markov decision processes (Jin et al. 2020b).

Imitation learning The case where the agent only observes expert demonstrations without further interaction with the environment corresponds to the well-known imitation learning problem. There are two primary approaches in this setting: *inverse reinforcement learning* (Ng and Russell 2000; Abbeel and Ng 2004; Ho and Ermon 2016) where the agent first infers a reward from demonstrations then finds an optimal policy for this reward; and *behavior cloning* (Pomerleau 1988; Ross and Bagnell 2010; Ross et al. 2011; Rajeswaran et al. 2018), a simpler method that employs supervised learning to imitate the expert. However collecting demonstration could be expansive and, furthermore, imitation learning suffers from the compounding errors effect, where the agent can diverge from the expert's policy in unvisited states (Ross and Bagnell 2010; Rajaraman et al. 2020). Hence, imitation learning is often combined with an online learning phase where the agent directly interacts with the environment.

BPI with demonstrations In BPI with demonstrations, the agent observes expert demonstrations like in imitation learning but also has the opportunity to collect new trajectories, including reward information, by directly interacting with the environment. There are three main method categories[1] for BPI with demonstration: one employs an off-policy algorithm augmented with a supervised learning loss and a replay buffer pre-filled the demonstrations (Hosu and Rebedea 2016; Lakshminarayanan et al. 2016; Vecerik et al. 2018; Hester et al. 2018); while a second uses reinforcement learning with a modified reward supplemented by auxiliary rewards obtained by inverse reinforcement learning (Zhu et al. 2018; Kang et al. 2018). The third class, demonstration-regularized RL, which is the one we study in this chapter, leverages behavior cloning to learn a policy that imitates the expert and then applies reinforcement learning with regularization toward this behavior cloning policy (Rajeswaran et al. 2018; Nair et al. 2018; Goecks et al. 2020; Pertsch et al. 2021).

Demonstration-regularized RL We introduce a particular demonstration-regularized RL method that consists of several steps. We start by learning with maximum likelihood estimation from the demonstrations of a behavior policy. This transfers the prior information from the demonstrations to a more practical representation: the behavior cloning policy. During the online phase, we aim to solve a trajectory Kullback-Leibler divergence regularized MDP (Neu et al. 2017; Vieillard et al. 2020; Tiapkin et al. 2023a), penalizing the policy for deviating too far from the behavior cloning policy. We use the solution of this regularized MDP as an estimate for the optimal policy in the unregularized MDP, effectively reducing BPI with demonstrations to regularized BPI.

Consequently, we propose two new algorithms for BPI in regularized MDPs: The `UCBVI-Ent+` algorithm, a variant of the `UCBVI-Ent` algorithm by Tiapkin et al. (2023a) with improved sample complexity, and the `LSVI-UCB-Ent` algorithm, its adaptation to the linear setting. When incorporated into the demonstration-regularized RL method, these algorithms yield sample complexity rates for

[1]The boundary between the above families of methods is not strict, since for example, one can see the regularization in the third family as a particular choice of auxiliary reward learned by inverse reinforcement learning that appears in the second class of methods.

BPI with N^{E} demonstrations of order[2] $\widetilde{\mathcal{O}}(\mathrm{Poly}(S, A, H)/(\varepsilon^2 N^{\mathrm{E}}))$ in finite and $\widetilde{\mathcal{O}}(\mathrm{Poly}(d, H)/(\varepsilon^2 N^{\mathrm{E}}))$ in linear MDPs, where ε is the target precision, H the horizon, A the number of action, S the number of states in the finite case and d the dimension of the feature space in the linear case. Notably, these rates show that leveraging demonstrations can significantly improve upon the rates of BPI without demonstrations, which are of order $\widetilde{\mathcal{O}}(\mathrm{Poly}(S, A, H)/\varepsilon^2)$ in finite MDPs (Kaufmann et al. 2021; Ménard et al. 2021) and $\widetilde{\mathcal{O}}(\mathrm{Poly}(d, H)/\varepsilon^2)$ in linear MDPs (Taupin et al. 2023). This work, up to our knowledge, represents the first instance of sample complexity rates for BPI with demonstrations, establishing the provable efficiency of demonstration-regularized RL.

Preference-based BPI with demonstration In RL with demonstrations, the assumption typically entails the observation of rewards in the online learning phase. However, in reinforcement learning from human feedback, such that recommendation system (Chaves et al. 2022), robotics (Jain et al. 2013; Christiano et al. 2017), clinical trials (Zhao et al. 2011) or large language models fine-tuning (Ziegler et al. 2020; Stiennon et al. 2020; Ouyang et al. 2022), the reward is implicitly defined by human values. Our focus centers on preference-based RL (PbRL, Busa-Fekete et al. 2014; Wirth et al. 2017; Novoseller et al. 2020; Saha et al. 2023) where the observed preferences between two trajectories are essentially noisy reflections of the value of a link function evaluated at the difference between cumulative rewards for these trajectories.

Existing literature on PbRL focuses either on the offline setting where the agent observes a pre-collected dataset of trajectories and preferences (Zhu et al. 2023; Zhan et al. 2023) or on the online setting where the agent sequentially samples a pair of trajectories and observes the associated preference (Saha et al. 2023; Xu et al. 2020; Wang et al. 2023).

In this chapter, we explore a hybrid setting that aligns more closely with what is done in practice (Ouyang et al. 2022). In this framework, which we call preference-based BPI with demonstration, the agent selects a sampling policy based on expert-provided demonstrations used to generate trajectories and associated preferences. The offline collection of preference holds particular appeal in RLHF due to the substantial cost and latency associated with obtaining preference feedback. Finally, in our setting, the agent engages with the environment by sequentially collecting *reward-free* trajectories and returns an estimate for the optimal policy.

Demonstration-regularized RLHF To address this novel setting, we follow a similar approach that was used in RL with demonstrations. We employ the dataset of preferences sampled using the behavior cloning policy to estimate rewards. Then, we solve the MDP regularized towards the behavior cloning policy, equipped with the estimated reward. Using the same regularized BPI solvers, we establish a sample complexity for the demonstration-regularized RLHF method of order $\widetilde{\mathcal{O}}((\mathrm{Poly}(S, A, H)/(\varepsilon^2 N^{\mathrm{E}}))$ in finite MDPs and $\widetilde{\mathcal{O}}((\mathrm{Poly}(d, H)/(\varepsilon^2 N^{\mathrm{E}}))$ in linear MDPs. Intriguingly, these rates mirror those of RL with demonstrations, illustrating that RLHF with demonstrations does not pose a greater challenge than RL with demonstrations. Notably, these findings expand upon the similar observation made by Wang et al. (2023) in the absence of prior information.

We highlight our main contributions:

- We establish that demonstration-regularized RL is an efficient solution method for RL with N^{E} demonstrations, exhibiting a sample complexity of order $\widetilde{\mathcal{O}}(\mathrm{Poly}(S, A, H)/(\varepsilon^2 N^{\mathrm{E}}))$ in finite MDPs and $\widetilde{\mathcal{O}}(\mathrm{Poly}(d, H)/(\varepsilon^2 N^{\mathrm{E}}))$ in linear MDPs.
- We provide evidence that demonstration-regularized methods can effectively address reinforcement learning from human feedback (RLHF) by collecting preferences offline and eliminating the necessity for pessimism (Zhan et al. 2023). Interestingly, they achieve sample complexities similar to those in RL with demonstrations.

[2] In the $\widetilde{\mathcal{O}}(\cdot)$ notation we ignore terms poly-log in $H, S, A, d, 1/\delta, 1/\varepsilon$ and the notation Poly indicates polynomial dependencies.

- We prove performance guarantees for the behavior cloning procedure in terms of Kullback-Leibler divergence from the expert policy. They are of order $\widetilde{\mathcal{O}}(\mathrm{Poly}(S, A, H)/N^{\mathrm{E}})$ for finite MDPs and $\widetilde{\mathcal{O}}(\mathrm{Poly}(d, H)/N^{\mathrm{E}})$ for linear MDPs.
- We provide novel algorithms for regularized BPI in finite and linear MDPs with sample complexities $\widetilde{\mathcal{O}}(H^5S^2A/(\lambda\varepsilon))$ and $\widetilde{\mathcal{O}}(H^5d^2/(\lambda\varepsilon))$, correspondingly, where λ is a regularization parameter.

3.2 Setting

MDPs. We consider an episodic MDP $\mathcal{M} = \big(\mathcal{S}, s_1, \mathcal{A}, H, \{p_h\}_{h\in[H]}, \{r_h\}_{h\in[H]}\big)$, where $\mathcal{S}$ is the set of states with s_1 the fixed initial state, $\mathcal{A}$ is the finite set of actions of size A, H is the number of steps in one episode, $p_h(s'|s,a)$ is the probability transition from state s to state s' by performing action a in step h. And $r_h(s,a) \in [0,1]$ is the reward obtained by taking action a in state s at step h.

We will consider two particular classes of MDPs.

Definition 3.1. (Finite MDP) An MDP $\mathcal{M}$ is *finite* if the state space $\mathcal{S}$ is finite with size denoted by S.

Definition 3.2. (Linear MDP) An MDP $\mathcal{M} = \big(\mathcal{S}, s_1, \mathcal{A}, H, \{p_h\}_{h\in[H]}, \{r_h\}_{h\in[H]}\big)$ is *linear* if the state space $\mathcal{S}$ is a measurable for a certain σ-algebra $\mathcal{F}_{\mathcal{S}}$, and there exists known feature map $\psi\colon \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$, and *unknown* parameters $\theta_h \in \mathbb{R}^d$ and an *unknown* family of signed measure $\mu_{h,i}, h \in [H], i \in [d]$ with its vector form $\mu_h\colon \mathcal{F}_{\mathcal{S}} \to \mathbb{R}^d$ such that for all $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$ and for any measurable set $B \in \mathcal{F}_{\mathcal{S}}$, it holds $r_h(s,a) = \psi(s,a)^\intercal\theta_h$, and $p_h(B|s,a) = \sum_{i=1}^d \psi(s,a)_i\mu_{h,i}(B) = \psi(s,a)^\intercal\mu_h(B)$. Without loss of generality, we assume $\|\psi(s,a)\|_2 \leqslant 1$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $\max\{\|\mu_h(\mathcal{S})\|_2, \|\theta_h\|_2\} \leqslant \sqrt{d}$ for all $h \in [H]$.

Policy & value functions. A policy π is a collection of functions $\pi_h\colon \mathcal{S} \to \Delta_{\mathcal{A}}$ for all $h \in [H]$, where every π_h maps each state to a probability over the action set. We denote by Π the set of policies. The value functions of policy π at step h and state s is denoted by V_h^π, and the optimal value functions, denoted by $V_h^\star$, are given by the Bellman and, respectively, optimal Bellman equations

$$\begin{aligned} Q_h^\pi(s,a) &= r_h(s,a) + p_hV_{h+1}^\pi(s,a) & V_h^\pi(s) &= \pi_hQ_h^\pi(s) \\ Q_h^\star(s,a) &= r_h(s,a) + p_hV_{h+1}^\star(s,a) & V_h^\star(s) &= \max_a Q_h^\star(s,a) \end{aligned}$$

where by definition, $V_{H+1}^\star \triangleq 0$. Furthermore, $p_hf(s,a) \triangleq \mathbb{E}_{s'\sim p_h(\cdot|s,a)}[f(s')]$ denotes the expectation operator with respect to the transition probabilities p_h and $\pi_hg(s) \triangleq \mathbb{E}_{a\sim\pi_h(\cdot|s)}[g(s,a)]$ denotes the composition with the policy π at step h.

Trajectory Kullback-Leibler divergence. We define the trajectory Kullback-Leibler divergence between policy π and policy π' as the average of the Kullback-Leibler divergence between policies at each step along a trajectory sampled with π,

$$\mathrm{KL}_{\mathrm{traj}}(\pi\|\pi') = \mathbb{E}_\pi\left[\sum_{h=1}^H \mathrm{KL}(\pi_h(s_h)\|\pi'_h(s_h))\right].$$

3.3 Behavior cloning

In this section, we analyze the complexity of behavior cloning for imitation learning in finite and linear MDPs.

Imitation learning. In imitation learning, a dataset $\mathcal{D}_{\mathrm{E}} \triangleq \{\mathring{\tau}_i = (s_1^i, a_1^i, \dots, s_H^i, a_H^i),\, i \in [N^{\mathrm{E}}]\}$ of N^{E} independent *reward-free* trajectories sampled from a fixed unknown expert policy π^{E} is provided. The objective is to learn from these demonstrations a policy close to optimal. In order to get useful demonstrations we assume that the expert policy is close to optimal, that is, $V_1^\star(s_1) - V_1^{\pi^{\mathrm{E}}}(s_1) \leqslant \varepsilon_{\mathrm{E}}$ for some small $\varepsilon_{\mathrm{E}} > 0$.

Behavior cloning. The simplest method for imitation learning is to directly learn to replicate the expert policy in a supervised fashion. Precisely the behavior cloning policy π^{BC} is obtained by minimizing the negative-loglikelihood over a class of policies $\mathcal{F} = \{\pi \in \Pi : \pi_h \in \mathcal{F}_h\}$ with $\mathcal{F}_h$ being a class of conditional distributions $\mathcal{S} \to \mathcal{P}(\mathcal{A})$ and where $\mathcal{R}_h$ is some regularizer,

$$\pi^{\mathrm{BC}} \in \arg\min_{\pi \in \mathcal{F}} \sum_{h=1}^{H} \left(\sum_{i=1}^{N^{\mathrm{E}}} \log \frac{1}{\pi_h(a_h^i | s_h^i)} + \mathcal{R}_h(\pi_h) \right). \tag{3.1}$$

In order to provide convergence guarantees for behavior cloning, we make the following assumptions. First, we assume some regularity conditions on the class of policies defined in terms of covering numbers of the class, see Appendix B.1 for a definition.

Assumption 1. For all $h \in [H]$, there are two positive constants $d_{\mathcal{F}}, R_{\mathcal{F}} > 0$ such that

$$\forall h \in [H], \forall \varepsilon \in (0,1) : \log \mathcal{N}(\varepsilon, \mathcal{F}_h, \|\cdot\|_\infty) \leqslant d_{\mathcal{F}} \log(R_{\mathcal{F}}/\varepsilon)\,.$$

Moreover, there is a constant $\gamma > 0$ such that for any $h \in [H]$, $\pi_h \in \mathcal{F}_h$ it holds $\pi_h(a|s) \geqslant \gamma$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$.

The Assumption 1 is a typical parametric assumption in density estimation, see, e.g., (Zhang 2002), with $d_{\mathcal{F}}$ being a covering dimension of the underlying parameter space. The part of the assumption on a minimal probability is needed to control KL-divergences (Zhang 2006).

Next, we assume that a smooth version of the expert policy belongs to the class of hypotheses.

Assumption 2. There is a constant $\kappa \in (0, 1/2)$ such that a κ-greedy version of the expert policy defined by $\pi_h^{\mathrm{E},\kappa}(a|s) = (1-\kappa)\pi_h^{\mathrm{E}}(a|s) + \kappa/A$ belongs to the hypothesis class of policies: $\pi^{\mathrm{E},\kappa} \in \mathcal{F}$.

Note that a deterministic expert policy verifies Assumption 2 provided that γ is small enough and the policy class is rich enough. For $\kappa = 0$, this assumption is never satisfied for any $\gamma > 0$.

In the sequel, we provide examples of the policy class $\mathcal{F}$ and regularizers $(\mathcal{R}_h)_{h\in[H]}$ for finite or linear MDPs such that the above assumptions are satisfied. We are now ready to state general performance guarantees for behavior cloning with KL regularization.

Theorem 3.3. *Let Assumptions 1-2 be satisfied and let $0 \leqslant \mathcal{R}_h(\pi_h) \leqslant M$ for all $h \in [H]$ and any policy $\pi \in \mathcal{F}_h$. Then with probability at least $1-\delta$, the behavior policy π^{BC} satisfies*

$$\mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}} \| \pi^{\mathrm{BC}}) \leqslant \frac{6 d_{\mathcal{F}} H \cdot (\log(A\mathrm{e}^3/(A\gamma \wedge \kappa)) \cdot \log(2HN^{\mathrm{E}} R_{\mathcal{F}}/(\gamma\delta))}{N^{\mathrm{E}}} + \frac{2HM}{N^{\mathrm{E}}} + \frac{18\kappa}{1-\kappa}\,.$$

This result shows that if the number of demonstrations N^{E} is large enough and $\gamma = 1/N^{\mathrm{E}}$, $\kappa = A/N^{\mathrm{E}}$ then the behavior cloning policy π^{BC} converges to the expert policy π^{E} at a fast rate of order $\widetilde{\mathcal{O}}((d_{\mathcal{F}}H + A)/N^{\mathrm{E}})$ where we measure the "distance" between two policies by the trajectory Kullback-Leibler divergence. The proof of this theorem is postponed to Appendix B.2 and it heavily relies on verifying the so-called Bernstein condition (Bartlett and Mendelson 2006).

3.3.1 Finite MDPs

For finite MDPs, we chose a logarithmic regularizer $\mathcal{R}_h(\pi_h) = \sum_{s,a} \log(1/\pi_h(a|s))$ and the class of policies $\mathcal{F} = \{\pi \in \Pi : \pi_h(a|s) \geqslant 1/(N^{\mathrm{E}} + A)\}$. One can check that Assumptions 1-2 hold and $0 \leqslant \mathcal{R}_h(\pi_h) \leqslant SA \log(N^{\mathrm{E}} + A)$. We can apply Theorem 3.3 to obtain the following bound for finite MDPs (see Appendix B.2.2 for additional details).

Corollary 3.4. *For all $N^{\mathrm{E}} \geqslant A$, for function class $\mathcal{F}$ and regularizer $(\mathcal{R}_h)_{h\in[H]}$ defined above, it holds with probability at least $1-\delta$,*

$$\mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}} \| \pi^{\mathrm{BC}}) \leqslant \frac{6SAH \cdot \log(2\mathrm{e}^4 N^{\mathrm{E}}) \cdot \log(12H(N^{\mathrm{E}})^2/\delta)}{N^{\mathrm{E}}} + \frac{18AH}{N^{\mathrm{E}}}\,.$$

Note that imitation learning with a logarithmic regularizer is closely related to the statistical problem of conditional density estimation with Kullback-Leibler divergence loss; see, for example, Section 4.3 by Hoeven et al. 2023 and references therein. Additionally, we would like to emphasize that the presented upper bound is optimal up to poly-logarithmic terms, see Appendix B.2.5 for a corresponding lower bound.

Remark 3.5. In fact, the constraint added by the class of policies $\mathcal{F}$ is redundant with the effect of regularization and one can directly optimize over the whole set of policies in (3.1). It is then easy to obtain a closed formula for the behavior cloning policy $\pi_h^{\mathrm{BC}}(a|s) = (N_h^{\mathrm{E}}(s,a) + 1)/(N_h^{\mathrm{E}}(s) + A)$, where we define the counts by $N_h^{\mathrm{E}}(s) = \sum_{a\in\mathcal{A}} N_h^{\mathrm{E}}(s,a)$ and $N_h^{\mathrm{E}}(s,a) = \sum_{i=1}^{N^{\mathrm{E}}} \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\}$.

Remark 3.6. Contrary to Ross and Bagnell (2010) and Rajaraman et al. (2020), our bound does not feature the optimality gap of the behavior policy but measures how close it is to the expert policy which is crucial to obtain the results of the next sections. Nevertheless, we can recover from our bound some of the results of the aforementioned references, see Appendix B.2.6 for details.

3.3.2 Linear MDPs

For the linear setting, we need the expert policy to belong to some well-behaved class of parametric policies. A first possibility would be to consider a greedy policy with respect to Q-value, linear in the feature space $\pi_h(s) \in \arg\max_{\pi\in\Delta_{\mathcal{A}}} (\pi\psi)(s)^\intercal w_h$ for some parameters w_h as it is done in the existing imitation learning literature (Rajaraman et al. 2021). However, under such a parametrization it would be almost impossible to learn an expert policy with a high quality since a small perturbation in the parameters w_h could lead to a completely different policy. We emphasize that if we assume that the expert policy is an optimal one, then Rajaraman et al. (2021) proposes a way to achieve a ε-optimal policy but not how to reconstruct the expert policy itself. That is why we consider another natural parametrization where the log probability of the expert policy is linear in the feature space.

Assumption 3. For all $h \in [H]$, there exists an *unknown* parameter $w_h^{\mathrm{E}} \in \mathbb{R}^d$ with $\|w_h^{\mathrm{E}}\|_2 \leqslant R$ for some known $R \geqslant 0$ such that $\pi_h^{\mathrm{E}}(a|s) = \exp(\psi(s,a)^\intercal w_h^{\mathrm{E}})/(\sum_{a'\in\mathcal{A}} \exp(\psi(s,a')^\intercal w_h^{\mathrm{E}}))$.

For instance, this assumption is satisfied for optimal policy in entropy-regularized linear MDPs, see Lemma B.3 in Appendix B.2.3. Under Assumption 3, a suitable choice of policy class is given by $\mathcal{F} = \{\pi \in \Pi : \pi_h \in \mathcal{F}_h\}$ where

$$\mathcal{F}_h = \left\{ \pi_h(a|s) = \frac{\kappa}{A} + (1-\kappa)\frac{\exp(\psi(s,a)^\intercal w_h)}{\sum_{a'\in\mathcal{A}} \exp(\psi(s,a')^\intercal w_h)} : w_h \in \mathbb{R}^d,\ \|w_h\|_2 \leqslant R \right\} \tag{3.2}$$

and $\kappa = A/(N^{\mathrm{E}} + A)$. Furthermore, for the linear setting, we do not need regularization, that is, $\mathcal{R}_h(\pi) = 0$. Equipped with this class of policies we can prove a similar result as in the finite setting with the number of states replaced by the dimension d of the feature space.

Corollary 3.7. *Under Assumption 3, function class $\mathcal{F}$ defined above and regularizer $\mathcal{R}_h = 0$ for all $h \in [H]$, it holds for all $N^{\mathrm{E}} \geqslant A$ with probability at least $1-\delta$,*

$$\mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}} \| \pi^{\mathrm{BC}}) \leqslant \frac{8dH \cdot (\log(2\mathrm{e}^3 A N^{\mathrm{E}}) \cdot (\log(48(N^{\mathrm{E}})^2 R) + \log(H/\delta)))}{N^{\mathrm{E}}} + \frac{18AH}{N^{\mathrm{E}}} .$$

Taking into account the fact that finite MDPs are a specific case within the broader category of linear MDPs, the lower bound presented in Appendix B.2.5 also shows the optimality of this result.

3.4 Demonstration-regularized RL

In this section, we study reinforcement learning when demonstrations from an expert are also available. First, we describe the regularized best policy identification framework that will be useful later.

Regularized best policy identification (BPI). Given some reference policy $\widetilde{\pi}$ and some regularization parameter $\lambda > 0$, we consider the trajectory Kullback-Leibler divergence regularized value function $V^{\pi}_{\widetilde{\pi},\lambda,1}(s_1) \triangleq V^{\pi}_1(s_1) - \lambda \mathrm{KL}_{\mathrm{traj}}(\pi, \widetilde{\pi})$. In this value function, the policy π is penalized for moving too far from the reference policy $\widetilde{\pi}$. Interestingly, we can compute the value of policy π with the regularized Bellman equations, (Neu et al. 2017; Vieillard et al. 2020)

$$Q^{\pi}_{\widetilde{\pi},\lambda,h}(s,a) = r_h(s,a) + p_h V^{\pi}_{\widetilde{\pi},\lambda,h+1}(s,a) , \quad V^{\pi}_{\widetilde{\pi},\lambda,h}(s) = \pi_h Q^{\pi}_{\widetilde{\pi},\lambda,h}(s) - \lambda \, \mathrm{KL}(\pi_h(s) \| \widetilde{\pi}_h(s)) ,$$

where $V^{\pi}_{\widetilde{\pi},\lambda,H+1} = 0$. Note that for the uniform policy $\widetilde{\pi}$ we recover the entropy-regularized Bellman equations (Fox et al. 2016; Schulman et al. 2018; Haarnoja et al. 2017; Haarnoja et al. 2018).

We are interested in the best policy identification for this regularized value. Precisely, in regularized BPI, the agent interacts with MDP as follows: at the beginning of episode t, the agent picks up a policy π^t based only on the transitions collected up to episode $t-1$. Then a new trajectory (with rewards) is sampled following the policy π^t and is observed by the agent. At the end of each episode, the agent can decide to stop collecting new data, according to a random stopping time $\mathfrak{t}$ ($\mathfrak{t} = t$ if the agent stops after the t-th episode), and output a policy $\widehat{\pi}$ based on the observed transitions. An agent for regularized BPI is therefore made of a triplet $((\pi^t)_{t\in\mathbb{N}}, \mathfrak{t}, \widehat{\pi})$.

Definition 3.8. (PAC algorithm for regularized BPI) An algorithm $((\pi^t)_{t\in\mathbb{N}}, \mathfrak{t}, \widehat{\pi})$ is (ε,δ)-PAC for BPI regularized with policy $\widetilde{\pi}$ and parameter λ with sample complexity $\mathcal{C}(\varepsilon,\lambda,\delta)$ if

$$\mathbb{P}\left(V^{\star}_{\widetilde{\pi},\lambda,1}(s_1) - V^{\widehat{\pi}}_{\widetilde{\pi},\lambda,1}(s_1) \leqslant \varepsilon , \quad \mathfrak{t} \leqslant \mathcal{C}(\varepsilon,\lambda,\delta) \right) \geqslant 1-\delta .$$

We can now describe the setting studied in this section.

BPI with demonstration. We assume, as in Section 3.3, that first the agent observes N^{E} independent trajectories $\mathcal{D}_{\mathrm{E}}$ sampled from an expert policy π^{E}. Then the setting is the same as in BPI. Precisely, the agent interacts with the MDP as follows: at episode t, the agent selects a policy π^t based on *the collected transitions and the demonstrations*. Then a new trajectory (with rewards) is sampled following the policy π^t and observed by the agent. At the end of each episode, the agent stops according to a stopping rule $\mathfrak{t}$ ($\mathfrak{t} = t$ if the agent stops after the t-th episode), and outputs a policy π^{RL}.

Definition 3.9. (PAC algorithm for BPI with demonstration) An algorithm $((\pi^t)_{t\in\mathbb{N}}, \mathfrak{t}, \pi^{\mathrm{RL}})$ is (ε,δ)-PAC for BPI with demonstration with sample complexity $\mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta)$ if

$$\mathbb{P}\left(V^{\star}_1(s_1) - V^{\pi^{\mathrm{RL}}}_1(s_1) \leqslant \varepsilon, \quad \mathfrak{t} \leqslant \mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta) \right) \geqslant 1-\delta .$$

To tackle BPI with demonstration we focus on the following natural and simple approach.

Demonstration-regularized RL. The main idea behind this method is to reduce BPI with demonstration to regularized BPI. Indeed, in demonstration-regularized RL, the agent starts by learning through behavior cloning from the demonstration of a policy π^{BC} that imitates the expert policy, refer to Section 3.3 for details. Then the agent computes a policy π^{RL} by performing regularized BPI with policy π^{BC} and some well-chosen parameter λ. The policy π^{RL} is then returned as the guess for an optimal policy. The whole procedure is described in Algorithm 3. Intuitively the prior information contained in the demonstration is compressed into a handful representation namely the policy π^{BC}. Then this information is injected into the BPI procedure by encouraging the agent to output a policy close to the behavior policy.

Algorithm 3 Demonstration-regularized RL

1: **Input:** Precision parameter $\varepsilon_{\mathrm{RL}}$, probability parameter δ_{RL}, demonstrations $\mathcal{D}_{\mathrm{E}}$, regularization parameter λ.
2: Compute behavior cloning policy $\pi^{\mathrm{BC}} = \texttt{BehaviorCloning}(\mathcal{D}_{\mathrm{E}})$.
3: Perform regularized BPI $\pi^{\mathrm{RL}} = \texttt{RegBPI}(\pi^{\mathrm{BC}}, \lambda, \varepsilon_{\mathrm{RL}}, \delta_{\mathrm{RL}})$
4: **Output:** policy π^{RL}.

Using the previous results for regularized BPI, we next derive guarantees for demonstration-regularized RL. We start from a general black-box result that shows how the final policy error depends on the behavior cloning error, parameter λ, and the quality of regularized BPI.

Theorem 3.10. *Assume that there are an expert policy π^{E} such that $V_1^\star(s_1) - V_1^{\pi^{\mathrm{E}}}(s_1) \leqslant \varepsilon_{\mathrm{E}}$ and a behavior cloning policy π^{BC} satisfying $\sqrt{\mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}} \| \pi^{\mathrm{BC}})} \leqslant \varepsilon_{\mathrm{KL}}$. Let π^{RL} be $\varepsilon_{\mathrm{RL}}$-optimal policy in λ-regularized MDP with respect to π^{BC}, that is, $V^\star_{\pi^{\mathrm{BC}},\lambda,1}(s_1) - V^{\pi^{\mathrm{RL}}}_{\pi^{\mathrm{BC}},\lambda,1} \leqslant \varepsilon_{\mathrm{RL}}$. Then π^{RL} fulfills*

$$V_1^\star(s_1) - V_1^{\pi^{\mathrm{RL}}}(s_1) \leqslant \varepsilon_{\mathrm{E}} + \varepsilon_{\mathrm{RL}} + \lambda \varepsilon_{\mathrm{KL}}^2 \,.$$

In particular, under the choice $\lambda^\star = \varepsilon_{\mathrm{RL}}/\varepsilon_{\mathrm{KL}}^2$, the policy π^{RL} is $(2\varepsilon_{\mathrm{RL}} + \varepsilon_{\mathrm{E}})$-optimal in the original (non-regularized) MDP.

Remark 3.11. We define an error in trajectory KL-divergence under the square root because the KL-divergence behaves quadratically in terms of the total variation distance by Pinsker's inequality.

Remark 3.12 (BPI with prior policy). We would like to underline that we exploit all the prior information only through one fixed behavior cloning policy. However, as it is observable from the bounds of Theorem 3.10, our guarantees are not restricted to such type of policies and potentially could work with any prior policy close enough to a near-optimal one in trajectory Kullback-Leibler divergence.

Proof. We start from the following observation that comes from the assumption on expert policy and a definition of regularized value

$$\begin{aligned} V_1^\star(s_1) - V_1^{\pi^{\mathrm{RL}}}(s_1) \leqslant \varepsilon_{\mathrm{E}} + V_1^{\pi^{\mathrm{E}}}(s_1) - V_1^{\pi^{\mathrm{RL}}}(s_1) \leqslant \varepsilon_{\mathrm{E}} + V^{\pi^{\mathrm{E}}}_{\pi^{\mathrm{BC}},\lambda,1}(s_1) + \lambda\, \mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}} \| \pi^{\mathrm{BC}}) \\ - V^{\pi^{\mathrm{RL}}}_{\pi^{\mathrm{BC}},\lambda,1}(s_1) - \lambda\, \mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{RL}} \| \pi^{\mathrm{BC}}) \leqslant \varepsilon_{\mathrm{E}} + \varepsilon_{\mathrm{RL}} + \lambda \varepsilon_{\mathrm{KL}}^2 \,, \end{aligned}$$

where in the last inequality we apply assumptions on π^{BC} and π^{RL}. □

In Appendix B.4 we propose and analyze the `UCBVI-Ent+` algorithm for regularized BPI. It is a modification of the algorithm `UCBVI-Ent` by Tiapkin et al. (2023a), that achieves a better sample complexity. The core idea leading to improvement of the sample complexity from $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ to $\widetilde{\mathcal{O}}(\varepsilon^{-1})$ is to use an additional randomization for an interaction policy. In particular, algorithm samples uniformly from one of $H+1$ policies $\pi^{t,(h')}$, where each one-step policy $\pi_h^{t,(h')}$ is optimistic for all

$h \neq h'$ and *greedy with respect to a width of a confidence interval* on the optimal Q-value for $h = h'$. This switching policy allows us to control the expected width of a confidence interval on the optimal Q-value over trajectories and, as a result, exploit strong convexity of the regularizer in a way similar to RL-Explore-Ent by Tiapkin et al. (2023a). In Appendix B.5, we also present the LSVI-UCB-Ent algorithm, a direct adaptation of the UCBVI-Ent+ to the linear setting. To apply this result and derive sample complexity for demonstration-regularized BPI, we present the UCBVI-Ent+ algorithm, a modification of the algorithm UCBVI-Ent proposed by Tiapkin et al. 2023a, that achieves better rates for regularized BPI in the finite setting. In Appendix B.5, we also present the LSVI-UCB-Ent algorithm, a direct adaptation of the UCBVI-Ent+ to the linear setting.

Notably, the use of UCBVI-Ent by Tiapkin et al. (2023a) within the framework of demonstration-regularized methods, fails to yield acceleration through expert data incorporation due to its $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity. In contrast, the enhanced variant, UCBVI-Ent+, exhibits a more favorable complexity of $\widetilde{\mathcal{O}}(1/(\varepsilon\lambda))$, where $\lambda = \varepsilon/\varepsilon_{\mathrm{KL}}^2 \gg \varepsilon$ under the conditions stipulated in Theorem 3.10, for a $\varepsilon_{\mathrm{KL}}$ sufficiently small. It is noteworthy that an alternative approach employing the RL-Explore-Ent algorithm, introduced by Tiapkin et al. (2023a), can also achieve such acceleration. However, RL-Explore-Ent is associated with inferior rates in terms of S and H and is challenging to extend beyond finite settings.

The UCBVI-Ent+ algorithm works by sampling trajectories according to an exploratory version of an optimistic solution for the regularized MDP which is characterized by the following rules.

UCBVI-Ent+ sampling rule. To obtain the sampling rule at episode t, we first compute a policy $\bar{\boldsymbol{\pi}}^t$ by optimistic planning in the regularized MDP,

$$\begin{aligned}
\overline{Q}_h^t(s,a) &= \mathrm{clip}\Big(r_h(s,a) + \widehat{p}_h^t \overline{V}_{h+1}^t(s,a) + b_h^{p,t}(s,a), 0, H\Big)\,, \\
\bar{\pi}_h^{t+1}(s) &= \arg\max_{\pi \in \Delta_{\mathcal{A}}} \Big\{\pi \overline{Q}_h^t(s) - \lambda\,\mathrm{KL}(\pi \| \widetilde{\pi}_h(s))\Big\}\,, \\
\overline{V}_h^t(s) &= \bar{\pi}_h^{t+1}\overline{Q}_h^t(s) - \lambda\,\mathrm{KL}(\bar{\pi}_h^{t+1}(s) \| \widetilde{\pi}_h(s))\,.
\end{aligned}$$

with $\overline{V}_{H+1}^t = 0$ by convention, where $\widetilde{\pi}$ is a reference policy, $\widehat{p}^t$ is an estimate of the transition probabilities, and b^t some bonus term taking into account estimation error for transition probabilities. Then, we define a family of policies that aim to explore actions for which Q-value is not well estimated at a particular step. That is, for $h' \in [0, H]$, the policy $\pi^{t,(h')}$ first follows the optimistic policy $\bar{\boldsymbol{\pi}}^t$ until step h where it selects an action leading to the largest width of a confidence interval for the optimal Q-value,

$$\pi_h^{t,(h')}(a|s) = \begin{cases} \bar{\pi}_h^t(a|s) & \text{if } h \neq h'\,, \\ \mathbb{1}\Big\{a = \arg\max_{a' \in \mathcal{A}} (\overline{Q}_h^t(s,a') - \underline{Q}_h^t(s,a'))\Big\} & \text{if } h = h'\,, \end{cases}$$

where $\underline{Q}^t$ is a lower bound on the optimal regularized Q-value function, see Appendix B.4.4. In particular, for $h' = 0$ we have $\pi^{t,(0)} = \bar{\boldsymbol{\pi}}^t$. The sampling rule is obtained by picking uniformly at random one policy among the family $\pi^t = \pi^{t,(h')}$, $h' \in [0, H]$ in each episode. Note that it is equivalent to sampling from a uniform mixture policy $\pi^{\mathrm{mix},t}$ over all $h' \in [0, H]$, see Appendix B.4 for more details. This algorithmic choice allows us to exploit strong convexity of the KL-divergence and control the properties of a stopping rule, defined in Appendix B.4.4, that depends on the gap $(\overline{Q}_h^t(s,a) - \underline{Q}_h^t(s,a))^2$.

Stopping rule and decision rule. To define the stopping rule, we first recursively build an upper-bound on the difference between the value of the optimal policy and the value of the current

optimistic policy $\bar{\boldsymbol{\pi}}^t$,

$$W_h^t(s,a) = \left(1 + \frac{1}{H}\right)\widehat{p}_h^t G_{h+1}^t(s) + b_h^{\mathrm{gap},t}(s,a),$$
$$G_h^t(s) = \mathrm{clip}\left(\bar{\pi}_h^{t+1} W_h^t(s) + \frac{1}{2\lambda}\max_{a\in\mathcal{A}}\left(\overline{Q}_h^t(s,a) - \underline{Q}_h^t(s,a)\right)^2, 0, H\right),$$

where $b_h^{\mathrm{gap},t}$ is a bonus defined in (B.10), Appendix B.4, $\underline{V}^t$ is a lower-bound on the optimal value function defined in Appendix B.4.4 and $G_{H+1}^t = 0$ by convention. Then the stopping time $\mathfrak{t} = \inf\{t \in \mathbb{N} : G_1^t(s_1) \leqslant \varepsilon\}$ corresponds to the first episode when this upper-bound is smaller than ε. At this episode we return the policy $\widehat{\pi} = \boldsymbol{\pi}^\tau$.

The complete procedure is described in Algorithm 8 in Appendix B.4. We prove that for the well-calibrated bonus functions $b^{p,t}$ and a stopping rule defined in Appendix B.4.4, the `UCBVI-Ent+` algorithm is (ε,δ)-PAC for regularized BPI and provide a high-probability upper bound on its sample complexity. Additionally, a similar result holds for `LSVI-UCB-Ent` algorithm. The next result is proved in Appendix B.4.5 and Appendix B.5.5.

Theorem 3.13. *For all $\varepsilon > 0$, $\delta \in (0,1)$, the* `UCBVI-Ent+` */* `LSVI-UCB-Ent` *algorithms defined in Appendix B.4.4 /Appendix B.5.4 are (ε,δ)-PAC for the regularized BPI with sample complexity*

$$\mathcal{C}(\varepsilon,\delta) = \widetilde{\mathcal{O}}\left(\frac{H^5 S^2 A}{\lambda\varepsilon}\right) \text{ (finite)} \qquad \mathcal{C}(\varepsilon,\delta) = \widetilde{\mathcal{O}}\left(\frac{H^5 d^2}{\lambda\varepsilon}\right) \text{ (linear)}.$$

Additionally, assume that the expert policy is $\varepsilon_{\mathrm{E}} = \varepsilon/2$-optimal and satisfies Assumption 3 in the linear case. Let π^{BC} be the behavior cloning policy obtained using corresponding function sets described in Section 3.3. Then demonstration-regularized RL based on `UCBVI-Ent+` */*`LSVI-UCB-Ent` *with parameters $\varepsilon_{\mathrm{RL}} = \varepsilon/4$, $\delta_{\mathrm{RL}} = \delta/2$ and $\lambda = \widetilde{\mathcal{O}}\left(N^{\mathrm{E}}\varepsilon/(SAH)\right)/\widetilde{\mathcal{O}}\left(N^{\mathrm{E}}\varepsilon/(dH)\right)$ is (ε,δ)-PAC for BPI with demonstration in finite / linear MDPs and has sample complexity of order*

$$\mathcal{C}(\varepsilon,N^{\mathrm{E}},\delta) = \widetilde{\mathcal{O}}\left(\frac{H^6 S^3 A^2}{N^{\mathrm{E}}\varepsilon^2}\right) \text{ (finite)} \qquad \mathcal{C}(\varepsilon,N^{\mathrm{E}},\delta) = \widetilde{\mathcal{O}}\left(\frac{H^6 d^3}{N^{\mathrm{E}}\varepsilon^2}\right) \text{ (linear)}.$$

In the finite setting, `UCBVI-Ent+` improves the previous fast-rate sample complexity result of order $\widetilde{\mathcal{O}}(H^8 S^4 A/(\lambda\varepsilon))$ by Tiapkin et al. (2023a). For the linear setting, we would like to acknowledge that `LSVI-UCB-Ent` is the first algorithm that achieves fast rates for exploration in regularized linear MDPs.

3.5 Demonstration-regularized RLHF

In this section, we consider the problem of reinforcement learning with human feedback. We assume that the MDP is finite, i.e., $|\mathcal{S}| < +\infty$ to simplify the manipulations with the trajectory space. However, the state space could be arbitrarily large. In this setting, we do not observe the true reward function $r^\star$ but have access to an oracle that provides a preference feedback between two trajectories. We assume that the preference is a random variable with parameters that depend on the cumulative rewards of the trajectories as detailed in Assumption 4. Given a reward function $r = \{r_h\}_{h=1}^H$, we define the reward of a trajectory $\tau \in (\mathcal{S}\times\mathcal{A})^H$ as the sum of rewards collected over this trajectory $r(\tau) \triangleq \sum_{h=1}^H r_h(s_h,a_h)$.

Assumption 4 (Preference-based model)**.** Let τ_0,τ_1 be two trajectories. The preference for τ_1 over τ_0 is a Bernoulli random variable o with a parameter $q_\star(\tau_0,\tau_1) = \sigma(r^\star(\tau_1) - r^\star(\tau_0))$, where $\sigma\colon \mathbb{R}\to[0,1]$ is a monotone increasing link function that satisfies $\inf_{x\in[-H,H]}\sigma'(x) = 1/\zeta$ for $\zeta > 0$.

The main example of the link function is a sigmoid function $\sigma(x) = 1/(1+\exp(-x))$ that leads to the Bradley-Terry-Luce (BTL) model (Bradley and Terry 1952) widely used in the literature (Wirth et al. 2017; Saha et al. 2023). We now introduce the learning framework.

Preference-based BPI with demonstration. We assume, as in Section 3.3, that the agent observes N^{E} independent trajectories $\mathcal{D}_{\mathrm{E}}$ sampled from an expert policy π^{E}. Then the learning is divided in two phases:

1) *Preference collection.* Based on the observed expert trajectories $\mathcal{D}_{\mathrm{E}}$, the agent selects a sampling policy π^{S} to generate a *dataset of preferences* $\mathcal{D}_{\mathrm{RM}} = \{(\tau_0^k, \tau_1^k, o^k)\}_{k=1}^{N^{\mathrm{RM}}}$ consisting of pairs of trajectories and the sampled preferences. Specifically, both trajectories of the pair (τ_0^k, τ_1^k) are sampled with the policy π^{S} and the associated preference o^k is obtained according to the preference-based model described in Assumption 4.

2) *Reward-free interaction.* Next, the agent interacts with the reward-free MDP as follows: at episode t, the agent selects a policy π^t based on *the collected transitions up to time t, demonstrations and preferences*. Then a new trajectory (reward-free) is sampled following the policy π^t and is observed by the agent. At the end of each episode, the agent can decide to stop according to a stopping rule $\mathfrak{t}$ and outputs a policy π^{RLHF}.

Definition 3.14 (PAC algorithm for preference-based BPI with demonstration)**.** An algorithm $((\pi^t)_{t\in\mathbb{N}}, \pi^{\mathrm{S}}, \mathfrak{t}, \pi^{\mathrm{RLHF}})$ is (ε,δ)-PAC for preference-based BPI with demonstrations and sample complexity $\mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta)$ if $\mathbb{P}\Big(V_1^\star(s_1) - V_1^{\pi^{\mathrm{RLHF}}}(s_1) \leqslant \varepsilon,\, \mathfrak{t} \leqslant \mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta)\Big) \geqslant 1-\delta$, where the unknown true reward function $r^\star$ is used in the value-function $V^\star$.

For the above setting, we provide a natural approach that combines demonstration-regularized RL with the maximum likelihood estimation of the reward given preferences dataset.

Demonstration-regularized RLHF. During the preference collection phase, the agent generates a dataset comprising trajectories and observed preferences, denoted as $\mathcal{D}_{\mathrm{RM}} = \{(\tau_0^k, \tau_1^k, o^k)\}_{k=1}^{N^{\mathrm{RM}}}$ by executing the previously computed policy π^{BC}. Using this dataset, the agent can infer the reward via maximum likelihood estimation (MLE).

Algorithm 4 Demonstration-regularized RLHF

1: **Input:** Precision parameter $\varepsilon_{\mathrm{RLHF}}$, probability parameter δ_{RLHF}, demonstrations $\mathcal{D}_{\mathrm{E}}$, preferences budget N^{RM}, regularization parameter λ.
2: Compute behavior cloning policy $\pi^{\mathrm{BC}} = \texttt{BehaviorCloning}(\mathcal{D}_{\mathrm{E}})$;
3: Select sampling policy $\pi^{\mathrm{S}} = \pi^{\mathrm{BC}}$ and collect preference dataset $\mathcal{D}_{\mathrm{RM}}$;
4: Compute reward estimate $\widehat{r} = \texttt{RewardMLE}(\mathcal{G}_r, \mathcal{D}_{\mathrm{RM}})$;
5: Perform regularized BPI using $\widehat{r}$ as reward: $\pi^{\mathrm{RLHF}} = \texttt{RegBPI}(\pi^{\mathrm{BC}}, \lambda, \varepsilon_{\mathrm{RLHF}}, \delta_{\mathrm{RLHF}}; \widehat{r})$
6: **Output:** policy π^{RLHF}.

The core idea behind this approach is to simplify the problem by transforming it into a regularized BPI problem. The agent starts with behavior cloning applied to the expert dataset, resulting in the policy π^{BC}. During the preference collection phase, the agent generates a dataset comprising trajectories and observed preferences, denoted as $\mathcal{D}_{\mathrm{RM}} = \{(\tau_0^k, \tau_1^k, o^k)\}_{k=1}^{N^{\mathrm{RM}}}$ by executing the previously computed policy π^{BC}. Using this dataset, the agent can infer the reward via MLE:

$$\widehat{r} \triangleq \arg\max_{r\in\mathcal{G}} \sum_{k=1}^{N^{\mathrm{RM}}} o^k \log\left(\sigma\Big(r(\tau_1^k) - r(\tau_0^k)\Big)\right) + (1-o^k)\log\left(1-\sigma\Big(r(\tau_1^k) - r(\tau_0^k)\Big)\right),$$

where $\mathcal{G}$ is a function class for trajectory reward functions[3]. Finally, the agent computes π^{RL} by performing regularized BPI with policy π^{BC}, a properly chosen regularization parameter λ and the estimated reward $\widehat{r}$. The complete procedure is outlined in Algorithm 4.

For this algorithm, we use the behavior cloning policy π^{BC} for two purposes. First, it allows efficient offline collection of the preference dataset $\mathcal{D}_{\mathrm{RM}}$, from which a high-quality estimate of the reward can be derived. Second, a regularization towards the behavior cloning policy π^{BC} enables the injection of information obtained from the demonstrations, while also avoiding the direct introduction of pessimism in the estimated reward as in the previous works that handle offline datasets (Zhu et al. 2023; Zhan et al. 2023).

Remark 3.15. Zhan et al. (2024) propose a similar two-stage setting of preference collection and reward-free interaction without prior demonstrations and propose an algorithm for this setup. However, as compared to their result, our pipeline is adapted to any parametric function approximation of rewards and does not require solving any (non-convex) optimization problem during the preference collection phase.

Remark 3.16. Our approach to solve BPI with demonstration within the preference-based model framework draws inspiration from well-established methods for large language model RL fine-tuning (Stiennon et al. 2020; Ouyang et al. 2022; Lee et al. 2024). Specifically, our algorithm's policy learning phase is similar to solving an RL problem with policy-dependent rewards

$$r_h^{\mathrm{RLHF}}(s,a) = \widehat{r}_h(s,a) - \lambda \log\Big(\pi_h^{\mathrm{RLHF}}(a|s)/\pi_h^{\mathrm{BC}}(a|s)\Big)\,.$$

This formulation, coupled with our prior stages of behavior cloning, akin to supervised fine-tuning (SFT), and reward estimation through MLE based on trajectories generated by the SFT policy, mirrors a simplified version of the three-phase RLHF pipeline.

The following sample complexity bounds for tabular and linear MDPs is a simple corollary of Theorem B.39 and Theorem 3.13 and its proof is postponed to Appendix B.6.3.

Corollary 3.17 (Demonstration-regularized RLHF)**.** *Let Assumption 4 hold. For $\varepsilon > 0$ and $\delta \in (0,1)$, assume that an expert policy ε_{E} is $\varepsilon/15$-optimal and satisfies Assumption 3 in the linear case. Let π^{BC} be the behavioral cloning policy obtained using function sets described in Section 3.3 and let the set $\mathcal{G}$ be defined in Lemma B.33 for finite and in Lemma B.34 for linear setting, respectively.*

If the following two conditions hold

$$(1)\, N^{\mathrm{RM}} \geqslant \widetilde{\Omega}\Big(\zeta \widetilde{D}/\varepsilon\Big); \qquad (2)\, N^{\mathrm{E}} \geqslant \widetilde{\Omega}\Big(H^2 \widetilde{D}/\varepsilon\Big)$$

for $\widetilde{D} = SA$ / d in finite / linear MDPs, then demonstration-regularized RLHF based on `UCBVI-Ent+` */* `LSVI-UCB-Ent` *with parameters $\varepsilon_{\mathrm{RL}} = \varepsilon/15$, $\delta_{\mathrm{RL}} = \delta/3$ and $\lambda = \lambda^\star = \widetilde{\mathcal{O}}\Big(N^{\mathrm{E}}\varepsilon/(SAH)\Big)$ / $\widetilde{\mathcal{O}}\Big(N^{\mathrm{E}}\varepsilon/(dH)\Big)$ is (ε,δ)-PAC for BPI with demonstration in finite / linear MDPs with sample complexity*

$$\mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta) = \widetilde{\mathcal{O}}\left(\frac{H^6 S^3 A^2}{N^{\mathrm{E}}\varepsilon^2}\right) \textit{ (finite)} \qquad \mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta) = \widetilde{\mathcal{O}}\left(\frac{H^6 d^3}{N^{\mathrm{E}}\varepsilon^2}\right) \textit{ (linear)},\,.$$

The conditions (1) and (2) control two different terms in the reward estimation error presented in Theorem B.39. In particular, to avoid direct pessimism injection, one needs to simultaneously keep both datasets (expert and reward modeling) large enough.

Additionally, one should notice that additional expert information and the bound on the minimal size of the expert dataset allows to avoid the notion of concentrability coefficients by directly keeping

[3]For the theoretical guarantees on MLE estimate of rewards $\widehat{r}$ we refer to Appendix B.6.1.

the RL policy very close to the data-generation policy. Thus, the lower bound in Theorem 3 by Zhan et al. 2023 is non-applicable in our setting.

The presented bound is non-trivial in the sense that, under this assumption on the expert dataset, imitation learning over a *stochastic and suboptimal expert* allows to obtain only a $\mathcal{O}(\sqrt{\varepsilon})$-optimal policy (see Appendix B.2.6 and Lemma B.13). At the same time, Lemma B.12 shows that in the case of a deterministic expert, this amount of demonstrations is enough to ensure that π^{BC} is $\mathcal{O}(\varepsilon)$-optimal, so additional fine-tuning with reward modeling would not improve upon the behavior cloning policy. The same situation holds under the assumption of an optimal expert ($\varepsilon_{\mathrm{E}} = 0$) by using the algorithm by Rajaraman et al. (2020) to perform imitation learning. However, in practical situations, we cannot assume either the expert's deterministic nature or its optimality.

3.6 Conclusion

In this study, we introduced the BPI with demonstration framework and showed that demonstration-regularized RL, a widely employed technique, is not just practical but also theoretically efficient for this problem. Additionally, we proposed a novel preference-based BPI with demonstration approach, where the agent gathers demonstrations offline. Notably, we proved that a demonstration-regularized RL method can also solve this problem efficiently without explicit pessimism injection. A compelling direction for future research could involve expanding the feedback mechanism in the preference-based setting, transitioning from pairwise comparison to preference ranking (Zhu et al. 2023). Additionally, it would be interesting to explore scenarios where the assumption of a white-box preference-based model, as proposed by Wang et al. (2023), is relaxed.

Chapter 4

Narrowing the Gap between Adversarial and Stochastic MDPs via Policy Optimization

Contents

We consider learning in adversarial Markov decision processes [MDPs] with an oblivious adversary in a full-information setting. The agent interacts with an environment during T episodes, each of which consists of H stages; each episode is evaluated with respect to a reward function revealed only at its end. We propose an algorithm, called APO-MVP, achieving a regret bound of order $\widetilde{\mathcal{O}}(\mathrm{poly}(H)\sqrt{SAT})$, where S and A are sizes of the state and action spaces, respectively. This result improves upon the best-known regret bound by a factor of $\sqrt{S}$, bridging the gap between adversarial and stochastic MDPs, and matching the minimax lower bound $\Omega(\sqrt{H^3SAT})$ as far as the dependencies in S, A, T are concerned. The proposed algorithm and analysis avoid the typical tool of occupancy measures, commonly used in adversarial tabular MDPs, and perform instead policy optimization based only on dynamic programming and on a black-box online linear optimization strategy run over estimated advantage functions, making it easy to implement. The analysis leverages policy optimization based on online linear optimization strategies (Jonckheere et al. 2025) and a refined martingale analysis of the impact on values of estimating transitions kernels by Zhang et al. (2024).

4.1 Introduction

We study adversarial Markov decision processes [MDPs], introduced by Even-Dar et al. (2009) and Yu et al. (2009), in an episodic setup with full monitoring. Unlike the standard setup, the reward function is not known to the learner beforehand and is revealed sequentially at the end of each episode.

To deal with this problem, many earlier works relied on online linear optimization [OLO] strategies (see the monograph by Cesa-Bianchi and Lugosi 2006 for a survey) in the space of so-called occupancy measures (Zimin and Neu 2013). These occupancy measures concern the state-action pairs within an episode induced by a given policy and transition kernel. This family of algorithms, known as `O-REPS`, has been extended to handle unknown transition kernels and bandit feedback by several studies (Rosenberg and Mansour 2019a; Rosenberg and Mansour 2019b; Jin et al. 2020a; Jin et al. 2021), using an exploration mechanism similar to `UCRL2` (Auer et al. 2008). However, this type of exploration leads to an additional $\sqrt{S}$ factor in the regret, where S is the number of states, compared to the state of the art in the non-adversarial case (Azar et al. 2017; Dann et al. 2017; Jin et al. 2018). Furthermore, `O-REPS`-based approaches require solving a high-dimensional convex program at each episode, resulting in a non-explicit policy update.

Another line of research has focused on policy-optimization-based approaches for adversarial MDPs, started by works of Cai et al. (2020) and Shani et al. (2020). These algorithms use a more practical approach combining dynamic programming with optimization directly in the policy space, instead of working with occupancy measures. Moreover, this approach is known to be highly practical due to its connection to the well-known `TRPO` (Schulman et al. 2015) and `PPO` (Schulman et al. 2017) algorithms. However, to the best of our knowledge, policy-optimization-based approaches also suffer from an additional $\sqrt{S}$ factor in the regret bound when specialized to finite MDP settings, compared to the state-of-the-art in the stochastic case (Azar et al. 2017).

To date, the question of whether dependency on the number of states can be matched between adversarial and stochastic cases remains open. We take the first step towards unifying these rates.

In this chapter, we use a black-box policy optimization approach, departing from the current state-of-the-art algorithms based on occupancy measures. This approach of policy optimization based on running online linear optimization strategies in a black-box way on estimated advantage functions was recently introduced by Jonckheere et al. (2025). The dynamic programming counterpart of our algorithm, as well as a part of the analysis, relies on the Monotonic Value Propagation [`MVP`] algorithm of Zhang et al. (2021b) and Zhang et al. (2024), which allowed to achieve optimal regret bounds up to second-order terms. However, since we do not yet target the lower-order terms, we significantly simplify their approach and provide an arguably more transparent exposition thereof. All in all, our policy-optimization-based algorithm achieves a $\widetilde{\mathcal{O}}(\mathrm{poly}(H)\sqrt{SAT})$ regret, where A is the number of actions and T is the number of episodes, and where we recall that H is the length of an episode and S is the number of states. This result improves on the previous regret bound of Rosenberg and Mansour (2019a) by a factor of $\sqrt{S}$, although it introduces an additional $\mathrm{poly}(H)$ factor. It also matches the minimax lower bound derived for the stochastic case (Jin et al. 2018; Domingues et al. 2021a) in all parameters except H.

Therefore, we demonstrate that while *policy optimization is already known to be practical, it is also more sample-efficient in large state spaces compared to existing `O-REPS`-based methods.*

Contributions. This chapter puts forward the following contributions, in the setting of adversarial episodic MDPs with full information: i) we introduce a algorithm called Adversarial Policy Optimization based on Monotonic Value Propagation (`APO-MVP`) that relies on a black-box online linear optimization solver and on dynamic programming, making it easier to implement in practice; ii) we demonstrate that the proposed algorithm is able to achieve a $\widetilde{\mathcal{O}}(\mathrm{poly}(H)\sqrt{TSA})$ regret, improving on the previously best-known dependency on the number of states S and achieving the

minimax lower bound $\Omega(\sqrt{H^3SAT})$ in all parameters, except H; iii) our analysis is modular and rather general, providing high flexibility and providing new tools for the study of adversarial MDPs with policy optimization.

More details on the reasons for the success in shaving off the $\sqrt{S}$ factor present in earlier analysis in this context may be found at the beginning of Section 4.4.

Notation. For any positive integer N, we denote by $[N] \triangleq \{1, \dots, N\}$ and $[N]^* \triangleq \{0\} \cup [N]$ the sets of the first positive and non-negative integers not greater than N, respectively. For $a, b \in \mathbb{R}$, we denote by $a \vee b$ and $a \wedge b$ the maximum and the minimum between a and b, respectively. For a finite set $\mathcal{E}$, we denote by $\Delta(\mathcal{E})$ the set of probability distributions over $\mathcal{E}$. We refer to natural logarithms by ln and to logarithms in base 2 by $\log_2$. When we write $\widetilde{\mathcal{O}}(\,\cdot\,)$, we hide all absolute constants and polylog multiplicative terms. We also adopt an agreement that $\max\{\emptyset\} = 0$.

4.1.1 Related work

Adversarial MDPs. The concept of adversarial (or online) MDPs was introduced by seminal works of Even-Dar et al. (2009) and Yu et al. (2009). In the case of MDPs with known transition kernel, Neu et al. (2010) and Zimin and Neu (2013) provided solutions, but dealing with the case of unknown transition kernels remained open till the work of Rosenberg and Mansour (2019a), who extended the algorithm of Zimin and Neu (2013) to this setting. Subsequent developments considered the extension to bandit feedback (Jin et al. 2020a), linear mixture MDPs (Cai et al. 2020; He et al. 2022), delayed feedback (Lancewicki et al. 2022), and linear MDPs (Luo et al. 2021). Despite these results, the setting of unknown transition kernel in finite MDPs with full-information feedback was not yet fully understood. Specifically, the upper bound established by Rosenberg and Mansour (2019a) does not match the best-known lower bound, leaving a gap in the theoretical understanding of this problem.

Policy optimization in adversarial MDPs. Policy optimization methods for adversarial MDPs with unknown transitions were first studied by Cai et al. (2020) and Shani et al. (2020). The first work focused on the case of linear mixture MDPs with full-information feedback, later improved by He et al. (2022), while the second one considered the case of tabular MDPs with bandit feedback, with Luo et al. (2021) refining the approach to achieve a state-of-the-art regret bound. More recent work has primarily addressed policy optimization under function approximation settings, including linear mixture MDPs (He et al. 2022) and linear MDPs (Sherman et al. 2023; Liu et al. 2024; Cassel and Rosenberg 2024). However, none of the methods was able to achieve improvements upon the occupancy-measure-based methods in the case of tabular MDPs in terms of a regret bound.

4.2 Problem Formulation

An H–episodic (obliviously) adversarial Markov decision process (MDP), where $H \geqslant 1$, is determined by a finite set of states $\mathcal{S}$, with cardinality S, a finite set of actions $\mathcal{A}$, with cardinality A, a sequence $\boldsymbol{P} = (P_h)_{h\in[H-1]}$ of Markov transition kernels $P_h \colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, and by a (potentially adversarially chosen) fixed-in-advance sequence $(\boldsymbol{r}_t)_{t\geqslant 1}$ of bounded time-inhomogeneous H–episodic reward functions. Each reward function is of the form $\boldsymbol{r}_t = (r_{t,h})_{h\in[H]}$, where $r_{t,h} \colon \mathcal{S} \times \mathcal{A} \to [0,1]$. For simplicity (and with no loss of generality, up to resorting to some doubling trick), we assume that the number T of episodes is fixed and known. We set some initial state s_1 for each episode. At each episode t and at each stage h, based on past observations, the learner picks a stage policy $\pi_{t,h} \colon \mathcal{S} \to \Delta(\mathcal{A})$ to draw the action. The interaction with the environment is therefore governed by the following protocol. For each episode $t = 1, \dots, T$:

1. Reset state $s_{t,1} = s_1$;
2. Start new episode — for each stage $h = 1, \dots, H$:

- Pick a policy $\pi_{t,h}$ and sample an action $a_{t,h} \sim \pi_{t,h}(\cdot \mid s_{t,h})$;
- If $h \leqslant H-1$, move to the next state $s_{t,h+1} \sim P_h(\cdot \mid s_{t,h}, a_{t,h})$;

3. Observe the reward function $\boldsymbol{r}_t = (r_{t,h})_{h\in[H]}$.

We compare the performance of the policies $\boldsymbol{\pi}_t = (\pi_{t,h})_{h\in[H]}$ picked to the one achieved by resorting to a static policy $\boldsymbol{\pi} = (\pi_h)_{h\in[H]}$ in each episode, in terms of value functions. We define the value function of a policy $\boldsymbol{\pi}$, at episode $t \in [T]$, and started from step $h \in [H]$, as

$$V_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s) \triangleq \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{P}}\left[\sum_{j=h}^{H} r_{t,j}(s_{t,j}, a_{t,j}) \;\middle|\; s_{t,h} = s\right];$$

we recall the environment (reward functions, transition kernels) in the notation for value functions and expectations, as environments will vary in the algorithm and analysis. The regret of the learner is defined as the difference between the accumulated value of the best static policy in hindsight and the gained value of the learner, that is,

$$R_T \triangleq \max_{\boldsymbol{\pi}} \sum_{t=1}^{T} \Big(V_1^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s_1) - V_1^{\boldsymbol{\pi}_t,\boldsymbol{r}_t,\boldsymbol{P}}(s_1)\Big). \tag{4.1}$$

The goal of the learner is to design policies $(\boldsymbol{\pi}_t)_{t\in[T]}$ minimizing the above-defined regret.

Additional notation. For the analysis, we define Q–value functions and remind Bellman's equations. For any policy $\boldsymbol{\pi}$, we define the Q–value function at episode $t \in [T]$, and started from step $h \in [H]$, as

$$Q_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s,a) \triangleq \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{P}}\left[\sum_{j=h}^{H} r_{t,j}(s_{t,j}, a_{t,j}) \;\middle|\; s_{t,h} = s,\ a_{t,h} = a\right].$$

The advantage function is in turn defined as $A_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s,a) \triangleq Q_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s,a) - V_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s)$.

We use the usual convention that for a transition kernel $K\colon \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$, a policy $\pi\colon \mathcal{S} \to \Delta(\mathcal{A})$, and two functions $f\colon \mathcal{S} \to \mathbb{R}$ and $g\colon \mathcal{S}\times\mathcal{A} \to \mathbb{R}$,

$$Kf(s,a) \triangleq \sum_{s'\in\mathcal{S}} K(s' \mid s,a)\, f(s') \quad \text{and} \quad \pi g(s) \triangleq \sum_{a\in\mathcal{A}} \pi(a \mid s)\, g(s,a)\,.$$

Then, Bellman's equations read for all episodes $t \in [T]$, steps $h \in [H-1]$, and policies $\boldsymbol{\pi}$, as

$$Q_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s,a) = r_{t,h}(s,a) + P_h V_{h+1}^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s,a) \quad \text{and} \quad V_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s) = \pi_{t,h} Q_h^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s)\,,$$

while for $h = H$, one has $Q_H^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s,a) = r_{t,H}(s,a)$ as well as $V_H^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s) = \pi_{t,H} Q_H^{\boldsymbol{\pi},\boldsymbol{r}_t,\boldsymbol{P}}(s)$.

4.3 Algorithm and Main Result

In this section, we first describe our algorithm, `APO-MVP`, which stands for Adversarial Policy Optimization based on Monotonic Value Propagation, and then state the performance bound obtained.

4.3.1 Algorithm `APO-MVP`

Let us start with a high-level description of the proposed algorithm, and details will be provided below. Similarly to Rosenberg and Mansour (2019a), our algorithm proceeds in random epochs $\mathcal{E}_e \subseteq [T]$ indexed by $e = 1, 2, \ldots, m(T)$ of random lengths denoted by $E_1, \ldots, E_{m(T)} \in [T]$, i.e., $E_e \triangleq |\mathcal{E}_e|$. At the beginning of each epoch e,

$$\text{estimates } \ \widehat{\boldsymbol{P}}^{(e)} = \left(\widehat{P}_h^{(e)}\right)_{h\in[H-1]}$$

and bonus functions $\boldsymbol{b}^{(e)} = \left(b_h^{(e)}\right)_{h\in[H]}$, where $b_h^{(e)} \colon \mathcal{S}\times\mathcal{A} \to [0,H]$, are computed and will be used during the entire epoch e, as detailed in (4.4)–(4.5) and in Fact 4.3. Actually, $b_H^{(e)}$ will be identically null, but we consider it so that the bonus functions $\boldsymbol{b}^{(e)}$ may be added to reward functions $\boldsymbol{r}_t$.

Within-epoch statement. We now explain the updates and choices made at each episode $t \in [T]$. First, the policies $\boldsymbol{\pi}_t$ are picked, as indicated below. Then, denoting by e_t the epoch such that $t \in \mathcal{E}_{e_t}$, at the end of episode t, i.e., once $\boldsymbol{r}_t$ is revealed, we build optimistic estimates of the Q–value and value functions in a backward fashion, based on Bellman's equations: $\widehat{Q}_{t,H}(s,a) \triangleq r_{t,H}(s,a)$ and $\widehat{V}_{t,H}(s) \triangleq \pi_{t,H}\widehat{Q}_{t,H}(s)$ and for $h \in [H-1]$,

$$\begin{aligned}\widehat{Q}_{t,h}(s,a) &\triangleq r_{t,h}(s,a) + b_h^{(e_t)}(s,a) + \widehat{P}_h^{(e_t)}\widehat{V}_{t,h+1}(s,a)\,,\\ \widehat{V}_{t,h}(s) &\triangleq \pi_{t,h}\widehat{Q}_{t,h}(s)\,.\end{aligned} \tag{4.2}$$

For all $h \in [H]$, estimated advantage functions are defined by $\widehat{A}_{t,h}(s,a) \triangleq \widehat{Q}_{t,h}(s,a) - \widehat{V}_{t,h}(s)$, and we denote $\widehat{A}_{t,h}(s,\cdot) \triangleq \left(\widehat{A}_{t,h}(s,a)\right)_{a\in\mathcal{A}}$.

The policies $\boldsymbol{\pi}_t = (\pi_{t,h})_{h\in[H]}$ are picked based on an online linear optimization [OLO] strategy $\varphi = (\varphi_t)_{t\geqslant 1}$, which is a sequence of functions $\varphi_t \colon (\mathbb{R}^{\mathcal{A}})^{t-1} \to \Delta(\mathcal{A})$ satisfying some performance guarantee stated in Definition 4.6. (The function φ_1 is constant.) We run SH such strategies in parallel as follows: $\forall (s,h) \in \mathcal{S}\times[H]$,

$$\pi_{t,h}(\cdot \mid s) = \varphi_t\left(\left(\widehat{A}_{\tau,h}(s,\cdot)\right)_{\tau\in\mathcal{E}_{e_t}\cap[t-1]}\right). \tag{4.3}$$

Note that these choices indeed exploit information available at the beginning of episode t (at the end of episode $t-1$), and rely only on the estimated advantage functions of the current epoch. One may see φ as an adaptive version of `PPO`- or `TRPO`-like updates (Schulman et al. 2015; Schulman et al. 2017). We will consider, for the sake of concreteness, the polynomial-potential- and exponential-potential-based strategies (see Examples 4.7 and 4.8 and references therein), but many other OLO strategies would work. Appendix C.1 states closed-form expressions of the policies constructed with these strategies as well as their computational complexities.

Remark 4.1 (Two technical remarks). The kernel estimate and the bonus functions are fixed within a given epoch, which is the main reason why we are able to provide a black-box treatment of the problem relying on any OLO strategy satisfying Definition 4.6.

As the reward function takes values in $[0,1]$, the Q–value functions are bounded by H, and it is a common practice in the case of non-adversarial reward functions to clip the estimates to $[0,H]$ (see, e.g., Azar et al. 2017), which only helps. Unfortunately, our adversarial analysis related to the OLO part of the proof heavily relies on the so-called performance-difference lemma (Kakade and Langford 2002), which does not hold once clipping is involved. Thus, we opt out from clipping, paying an additional H factor at the eventual regret bound of Theorem 4.4 but still improving the dependency on S. Successful incorporation of clipping could improve the regret by an H multiplicative factor. A possible alternative to the clipping mechanism, connected to a notion of contracted MDPs, was recently proposed by Cassel and Rosenberg (2024) and might form a viable option to improve the dependency of our regret bound in H.

Epoch switching. The epoch-switching conditions below were also considered and analyzed by Zhang et al. (2024). We introduce the following empirical counts, for all episodes $t \in [T]$, stages $h \in [H-1]$, state–action pairs $(s,a) \in \mathcal{S}\times\mathcal{A}$, and states $s' \in \mathcal{S}$:

$$n_{t,h}(s,a,s') \triangleq \sum_{\tau=1}^{t} \mathbb{1}\left\{(s_{\tau,h},a_{\tau,h},s_{\tau,h+1}) = (s,a,s')\right\}, \qquad n_{t,h}(s,a) \triangleq \sum_{s'\in\mathcal{S}} n_{t,h}(s,a,s')\,.$$

We start at epoch $e = 1$. When for some $(t, h, s, a) \in [T] \times [H-1] \times \mathcal{S} \times \mathcal{A}$, the count $n_{t,h}(s,a)$ equals $2^{\ell-1}$ for some integer $\ell \geqslant 1$, the next epoch is started at episode $t+1$.

Now, for each episode $t \in [T]$, stage $h \in [H-1]$, and state–action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, we denote by

$$\mathrm{sw}_{t,h}(s,a) \triangleq \max\Big\{\tau \in [t] : \exists \ell \geqslant 1,\, n_{\tau,h}(s,a) = 2^{\ell-1}\Big\}\,.$$

In words, $\mathrm{sw}_{t,h}(s,a)$ is the last episode when an epoch switch took place because, among others, of (s,a). We refer to the largest value of ℓ in the maximum defining $\mathrm{sw}_{t,h}(s,a)$ by

$$\ell_{t,h}(s,a) \triangleq \max\Big\{\ell \geqslant 1 : n_{t,h}(s,a) \geqslant 2^{\ell-1}\Big\}\,.$$

The values $\ell_{t,h}(s,a)$ index local epochs for a given state–action pair (s,a), while the global epochs e_t are defined based on all local epochs. (More details may be found in Section 4.4.2.)

Fact 4.2. By design, the functions $\mathrm{sw}_{t-1,h}$ and $\ell_{t-1,h}$ defined above (note the subscripts $t-1$ here) are identical for all episodes $t \in \mathcal{E}_e$ of a given epoch e.

We may now define the estimated transition kernels $\widehat{\boldsymbol{P}}_t$ and bonus functions $\boldsymbol{b}_t$: for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $s' \in \mathcal{S}$, first for all $h \in [H-1]$,

$$\widehat{P}_{t,h}(s' \mid s,a) \triangleq \begin{cases} 1/S & \text{if } n_{\tau,h}(s,a)=0\,, \\ \dfrac{n_{\tau,h}(s,a,s')}{n_{\tau,h}(s,a)} & \text{if } n_{\tau,h}(s,a)\geqslant 1\,, \end{cases} \tag{4.4}$$

with $\tau = \mathrm{sw}_{t-1,h}(s,a)$, and

$$b_{t,h}(s,a) \triangleq \begin{cases} H & \text{if } \ell = 0\,, \\ \sqrt{\dfrac{2H^2 \ln(J)}{2^{\ell-1}}} \wedge H & \text{if } \ell \geqslant 1\,, \end{cases} \tag{4.5}$$

with $J = 2SATH \log_2(2T)/\delta$ and $\ell = \ell_{t-1,h}(s,a)$. We also set, by convention, $b_{t,H}(s,a) = 0$. In particular, $\widehat{P}_{t,h}(\,\cdot \mid s,a)$ corresponds to an empirical frequency vector based on $2^{\ell_{t-1,h}(s,a)-1}$ values when $\ell_{t-1,h}(s,a) \geqslant 1$.

Fact 4.3. By Fact 4.2, the above-defined $\widehat{\boldsymbol{P}}_t$ and $\boldsymbol{b}_t$ are indeed identical over all episodes $t \in \mathcal{E}_e$ of a given epoch e.

Summary. The strategy outlined above is summarized in the algorithm sketch titled Algorithm 5; for particular choices of function φ, see Examples 4.7 and 4.8. We refer to Appendix C.1 for a more detailed algorithmic description, titled Algorithm 10, including the details of the computation of $\widehat{Q}_{t,h}$, $\widehat{V}_{t,h}$, and $\widehat{A}_{t,h}$.

It is important to note that the computational complexity of the algorithm is equivalent to that of standard dynamic programming methods, as the policy optimization phase introduces no additional computational overhead beyond the requirements of transition kernel estimation and value function updates.

4.3.2 Main Result

We may now state our main result and discuss its relation to previously known bounds.

Theorem 4.4 (Main theorem)**.** *Algorithm* `APO-MVP`*, used, for instance, with the OLO strategies based on polynomial or exponential potential (see Examples 4.7 and 4.8), satisfies, with probability*

Algorithm 5 Adversarial Policy Optimization based on Monotonic Value Propagation (`APO-MVP`)

Input: Confidence level δ, numbers T, S, A, H, online linear optimization strategy $\varphi = (\varphi_t)_{t\geqslant 1}$.

Define $J = 2SATH\log_2(2T)/\delta$;
Initialize $\widehat{P}_h \equiv 1/S$, $n_h \equiv 0$, $\mathcal{H}_{h,s} \equiv \emptyset$ for all (h,s);
Initialize $\pi_{1,h}(\,\cdot \mid s) = \varphi_1(\emptyset)$ for all (h,s);
Set initial state $s_1 \in \mathcal{S}$;
for rounds $t = 1, \dots, T$ **do**
  Set $s_{t,1} = s_1$;
  for $h = 1, \dots, H$ **do**
    Play action $a_{t,h} \sim \pi_{t,h}(\,\cdot \mid s_{t,h})$;
    Receive $s_{t,h+1} \sim P_h(\,\cdot \mid s_{t,h}, a_{t,h})$;
    Update counters $n_h(s_{t,h}, a_{t,h})$ += 1 and $n_h(s_{t,h}, a_{t,h}, s_{t,h+1})$ += 1;
    if $\exists\, \ell \geqslant 1 : n_h(s_{t,h}, a_{t,h}) = 2^{\ell-1}$ **then**
      $\forall s', \quad \widehat{P}_h(s' | s_{t,h}, a_{t,h}) = \frac{n_h(s_{t,h}, a_{t,h}, s')}{2^{\ell-1}}$;
      $b_h(s_{t,h}, a_{t,h}) = \sqrt{\frac{2H^2 \ln(J)}{2^{\ell-1}}} \wedge H$;
      Activate `trigger`;
    end if
  end for
  Receive reward function $\boldsymbol{r}_t = (r_{t,h})_{h\in[H]}$;
  if `trigger` **then**
    Set $\mathcal{H}_{h,s} = \emptyset$ for all (h,s);
    Deactivate `trigger`;
  else
    Compute $\widehat{Q}_{t,h}$, $\widehat{V}_{t,h}$, and $\widehat{A}_{t,h}$ via (4.2);
    Add $\widehat{A}_{t,h}(s,\cdot)$ to $\mathcal{H}_{h,s}$ for all (h,s);
  end if
  Let $\pi_{t+1,h}(\,\cdot \mid s) = \varphi_t(\mathcal{H}_{h,s})$ for all (h,s);
end for
Output: sequence of policies $(\boldsymbol{\pi}_t)_{t=1}^T$.

at least $1 - 3\delta$,

$$\begin{aligned} R_T \leqslant\ & \sqrt{H^7SAT\log_2(2T)}\left(2\log_2(2T) + 16\sqrt{\ln(A)}\right) \\ & + 7\sqrt{H^4SAT\ln\left(2SATH\log_2(2T)/\delta\right)} \\ & + 2\sqrt{2H^6T\log_2(2T)\ln(2/\delta)} + 2H^3SA\,. \end{aligned}$$

Proof. The result follows the decomposition stated in the introduction of Section 4.4 together with Lemmas 4.5–4.13–4.14–4.15 located therein. □

Theorem 4.4 shows that the regret bound is of order $\widetilde{\mathcal{O}}(\sqrt{H^7SAT})$, matching the minimax lower bound $\Omega(\sqrt{H^3SAT})$ for stochastic MDPs in terms of dependencies on S, A, and T, up to logarithmic factors (Jin et al. 2018; Domingues et al. 2021a). To the best of our knowledge, it is the first result that achieves the minimax optimal dependency on the number of states S in the adversarial setting.

Comparison to Rosenberg and Mansour (2019a). Algorithm `UC-O-REPS` by Rosenberg and Mansour (2019a) achieves $\widetilde{\mathcal{O}}(\sqrt{H^4S^2AT})$ regret bound in our setting and with our notation (taking $L = H$ and $|\mathcal{X}| = HS$ since a state-space layer $\mathcal{X}$ may be represented as H independent copies of $\mathcal{S}$). In particular, our result improves upon the previous best known bound in the regime of large state spaces $S \geqslant H^3$. We suspect that—perhaps through successful incorporation of clipping, see Remark 4.1—the regret bound could be improved to $\widetilde{\mathcal{O}}(\sqrt{H^5SAT})$. We plan to investigate this in future works, and it remains an open problem to fully match the minimax lower bound $\Omega(\sqrt{H^3SAT})$.

Finally, our analysis in Section 4.4.3 relies significantly on the fact that the adversary is oblivious, while `UC-O-REPS` can handle fully adversarial setups. However, due to the exploration mechanism used, this algorithm is not able to take advantage of the oblivious adversary and would still pay the same $\sqrt{S}$ factor.

Comparison to Cai et al. (2020). Our algorithm shares similarities with the online proximal policy optimization [`OPPO`] approach of Cai et al. (2020) and Shani et al. (2020), which also uses dynamic programming and policy optimization through online mirror descent. However, our approach incorporates the doubling trick to stabilize value updates, enabling us to: i) employ any online linear optimization strategy in a black-box manner without unnecessary adaptations; and ii) improve the dependency on S by a multiplicative factor $\sqrt{S}$, by leveraging the analysis of Zhang et al. (2024).

4.4 Proof Sketch for Theorem 4.4

We decompose the regret into four terms that are treated separately. Denoting by $\boldsymbol{\pi}^\star$ as the policy that achieves the maximum in (4.1), we decompose the regret R_T, following ideas of Auer et al. (2008) and Azar et al. (2017), as: $R_T = (\mathbf{A}) + (\mathbf{B}) + (\mathbf{C}) + (\mathbf{D})$, where

$$
\begin{aligned}
(\mathbf{A}) &\triangleq \sum_{t=1}^{T} \left(V_1^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s_1) - V_1^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_1) \right), \\
(\mathbf{B}) &\triangleq \sum_{t=1}^{T} \left(V_1^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_1) - V_1^{\boldsymbol{\pi}_t, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_1) \right), \\
(\mathbf{C}) &\triangleq \sum_{t=1}^{T} \left(V_1^{\boldsymbol{\pi}_t, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_1) - V_1^{\boldsymbol{\pi}_t, \boldsymbol{r}_t + \boldsymbol{b}_t, \boldsymbol{P}}(s_1) \right), \\
(\mathbf{D}) &\triangleq \sum_{t=1}^{T} V_1^{\boldsymbol{\pi}_t, \boldsymbol{b}_t, \boldsymbol{P}}(s_1) \,,
\end{aligned}
$$

where we used the linearity of the value functions: $V_1^{\boldsymbol{\pi}, \boldsymbol{g}+\boldsymbol{g}', \boldsymbol{Q}} \equiv V_1^{\boldsymbol{\pi}, \boldsymbol{g}, \boldsymbol{Q}} + V_1^{\boldsymbol{\pi}, \boldsymbol{g}', \boldsymbol{Q}}$.

The keys to success / Challenges overcome. We now provide a high-level overview of the techniques used for each term. We do not claim novelty in the very control of any of the terms involved; novelty lies in the decomposition above of the regret, which somehow separates adversarial and stochastic parts. We then apply or adapt the right tools, some of them being developed recently. One could argue that prior works on adversarial tabular MDPs have struggled to eliminate the $\sqrt{S}$ factor precisely because they adopted an overly adversarial approach to the problem. In contrast, our approach stems from deliberately blending adversarial and stochastic elements—a balance that enables us to derive the bound.

The treatment of $(\mathbf{B})$ definitely belongs to the adversarial parts of the proof scheme. We crucially use that within each epoch the considered transition kernels are constant (see Fact 4.3), so that

we may resort to the adversarial-learning techniques reviewed by Jonckheere et al. (2025), which consists of running SH independent OLO strategies on advantage functions or on Q–value functions. (Advantage functions are preferred in practice, though.) The OLO strategies that are the most popular in the literature are based on exponential weights (sometimes called multiplicative weights; see the detailed discussion in Jonckheere et al. 2025, Section 1.1), but other strategies are suitable. We deal with the underlying OLO strategies as black boxes, to get a modular and transparent proof.

Dealing with (**C**) forms a stochastic part of the proof scheme, and is actually the most involved part of the analysis from the probabilistic standpoint: we resort to the machinery developed by Zhang et al. (2024), which relies greatly on a doubling trick that we mimicked in the definition of the `APO-MVP` algorithm. We however make the argument of Zhang et al. (2024) more modular and extract the essential elements of the proof.

Finally, for (**A**) we leverage the careful choice of the bonuses $\boldsymbol{b}_t$ to show that on a properly chosen high-probability event, (**A**) is non-positive, while (**D**) is the least involved term: it can be controlled by some lines of elementary calculations.

In what follows, each section provides additional details of the analysis per term, in the order: (**B**) – additional technical concepts – (**A**) – (**D**) – (**C**).

4.4.1 Term (B): OLO Analysis

The goal of this section is to prove the following result, which also holds for other OLO strategies satisfying the performance guarantee of Definition 4.6.

Lemma 4.5. *Among others, the OLO strategies based on polynomial or exponential potentials (see Examples 4.7 and 4.8) satisfy*

$$(\mathbf{B}) \leqslant 16\sqrt{H^7 SAT \log_2(2T) \ln(A)}\,.$$

Before proving this result, let us briefly recall what online linear optimization [OLO] consists of; see the monograph by Cesa-Bianchi and Lugosi (2006) for a more detailed exposition. We take some generic notation for now but will later connect OLO to constructions of policies; in particular, we consider for now reward vectors of length $K \geqslant 2$, but will later replace $[K]$ by the action space $\mathcal{A}$.

Online linear optimization. At each round $t \geqslant 1$ and based on the past, a learning strategy $\varphi = (\varphi_t)_{t\geqslant 1}$ picks a convex combination $\boldsymbol{w}_t = (w_{t,1}, \dots, w_{t,K}) \in \Delta\Big([K]\Big)$ while an opponent player picks, possibly at random, a vector $\boldsymbol{g}_t = (g_{t,1}, \dots, g_{t,K})$ of signed rewards. Both $\boldsymbol{w}_t$ and $\boldsymbol{g}_t$ are revealed at the end of the round. By "based on the past", we mean, for the learning strategy, that $\boldsymbol{w}_t = \varphi_t\Big((\boldsymbol{g}_\tau)_{\tau\leqslant t-1}\Big)$. The initial vector $\boldsymbol{w}_1$ is constant.

Definition 4.6. A learning strategy φ controls the regret in the adversarial setting with rewards bounded by $M > 0$ if there exists a sequence $(B_{T,K})_{T\geqslant 1}$ with $B_{T,K}/T \to 0$ such that, against all opponent players sequentially picking reward vectors $\boldsymbol{g}_t$ in $[-M, M]^K$, for all $T \geqslant 1$,

$$\max_{k\in[K]} \sum_{t=1}^{T} g_{t,k} - \sum_{t=1}^{T} \sum_{j\in[K]} w_{t,j}\, g_{t,j} \leqslant 2M\, B_{T,K}\,.$$

The optimal orders of magnitude of $B_{T,K}$ are $\sqrt{T\ln(K)}$. In Definition 4.6, the strategy may know M and rely on its value. Also, the strategy should work for any optimization horizon T (see the final "for all $T \geqslant 1$" in the definition above): this is because the lengths E_e of the global epochs $\mathcal{E}_e$ are not known in advance. There exist several strategies meeting the requirements of Definition 4.6; we provide two examples.

Example 4.7. The potential-based strategies by Cesa-Bianchi and Lugosi (2003) are defined based on a non-decreasing function $\Phi\colon \mathbb{R} \to [0, +\infty)$. They resort to $w_{1,k} = \frac{1}{K}$ and for $t \geqslant 2$, to $w_{t,k} = \frac{v_{t,k}}{\sum_{j\in[K]} v_{t,j}}$, where

$$v_{t,k} = \Phi\left(\sum_{\tau=1}^{t-1} g_{\tau,k} - \sum_{\tau=1}^{t-1}\sum_{j\in[K]} w_{\tau,j} g_{\tau,j}\right). \tag{4.6}$$

For $\Phi\colon x \mapsto \Big(\max\{x, 0\}\Big)^{2\ln(K)}$, the polynomial potential, Cesa-Bianchi and Lugosi (2003, Section 2) show that the strategy satisfies the performance guarantee of Definition 4.6 with $B_{T,K} = \sqrt{6T\ln(K)}$.

Example 4.8. Auer et al. (2002) studied the use of exponential potential with time-varying learning rates $\eta_t = (1/M)\sqrt{\ln(K)/t}$, i.e., using $\Phi_t(x) = \exp(\eta_t x)$ in (4.6) to define the weights at round t. This strategy satisfies the performance guarantee of Definition 4.6 with $B_{T,K} = \sqrt{T\ln(K)}$.

There exist adaptive versions of the previous two strategies: `ML-Poly` in Gaillard et al. (2014), `AdaHedge` in Erven et al. (2011), Rooij et al. (2014), Orabona and Pál (2015).

Appendix C.1 states closed-form expressions of the policies (4.3) constructed with the strategies of Examples 4.7 and 4.8, as well as `AdaHedge`.

Connection between OLO and the construction of policies. Jonckheere et al. (2025) prove the following. Let $r'_{t,h}\colon \mathcal{S}\times\mathcal{A} \to [0, M]$ be a sequence of reward functions. Define a sequence of policies $(\boldsymbol{\pi}'_t)_{t\geqslant 1}$ as: for each $t \geqslant 1$, for each $s \in \mathcal{S}$, for each $h \in [H]$,

$$\pi'_{t,h}(\,\cdot\,|s) = \varphi_t\left(\left(A_h^{\boldsymbol{\pi}'_\tau, \boldsymbol{r}'_\tau, \boldsymbol{P}'}(s,\,\cdot\,)\right)_{\tau\leqslant t-1}\right)$$
$$\text{where}\quad A_h^{\boldsymbol{\pi}'_\tau, \boldsymbol{r}'_\tau, \boldsymbol{P}'}(s,\,\cdot\,) = \left(A_h^{\boldsymbol{\pi}'_\tau, \boldsymbol{r}'_\tau, \boldsymbol{P}'}(s,a)\right)_{a\in\mathcal{A}}$$

Lemma 4.9 (Jonckheere et al. 2025)**.** *If the learning strategy satisfies the conditions of Definition 4.6, then the sequence of policies defined right above is such that, for all fixed policies* $\boldsymbol{\pi} = (\pi_h)_{h\in[H]}$*, for all* $T \geqslant 1$*,*

$$\sum_{t=1}^{T}\left(V_1^{\boldsymbol{\pi}, \boldsymbol{r}'_t, \boldsymbol{P}'}(s_1) - V_1^{\boldsymbol{\pi}'_t, \boldsymbol{r}'_t, \boldsymbol{P}'}(s_1)\right) \leqslant 2MH^2\, B_{T,A}\,.$$

For completeness, the proof of Lemma 4.9 is provided in Appendix C.2. We are ready to prove Lemma 4.5.

Proof of Lemma 4.5. We apply Lemma 4.9 in each global epoch $\mathcal{E}_e$, with $\boldsymbol{P}' = \widehat{\boldsymbol{P}}_t$ (see Fact 4.3) and $\boldsymbol{r}'_t = \boldsymbol{r}_t + \boldsymbol{b}_t$ for all $t \in \mathcal{E}_e$. Since $r_{t,h} \in [0, 1]$ and $b_{t,h} \in [0, H]$, we can pick $M = 1 + H \leqslant 2H$. Decomposing term (**B**) into a summation over the global epochs and using the bound of Lemma 4.9 for each of them, we deduce that, for both strategies of Examples 4.7 and 4.8,

$$(\mathbf{B}) \leqslant 4H^3 \sum_{e=1}^{m(T)} B_{E_e,A} \leqslant 4H^3 \sum_{e=1}^{m(T)} \sqrt{6E_e\ln(A)} \leqslant 16H^3\sqrt{Tm(T)\ln(A)}\,, \tag{4.7}$$

where we applied Jensen's inequality to the root. Lemma 4.10 below then yields the claimed result. □

4.4.2 Additional Technical Concepts

To deal with the remaining terms (**A**)–(**D**)–(**C**), we will not need anymore to pay attention to global epochs $\mathcal{E}_{e_t}$, only local epochs $\ell_{t,h}(s,a)$ are of interest.

We review two concepts which have been successfully used by Zhang et al. (2024) to derive minimax optimal regret bounds in the case of stochastic MDPs.

The first concept: epoch-switching conditions and profiles. The functions indicating local epochs $\ell_{t,h}(s,a)$ were called a profile by Zhang et al. (2024); they take bounded values: $\ell_{t,h}\colon \mathcal{S}\times\mathcal{A}\to\Big[\lceil\log_2(T)\rceil\Big]^*$, and let

$$\boldsymbol{\ell}_t = (\ell_{t,h})_{h\in[H-1]} \tag{4.8}$$

with the agreement $\ell_{0,h}(\,\cdot\,,\,\cdot\,)\equiv 0$ for $h\in[H-1]$. We also introduce $\boldsymbol{\ell}_{<t}=(\boldsymbol{\ell}_\tau)_{0\leqslant\tau\leqslant t-1}$. Using the above-defined profiles, we note that the global epoch e_t of a given episode $t\in[T]$ may be obtained as a function of $\boldsymbol{\ell}_{<t}$, namely,

$$e_t = \sum_{\tau=1}^{t-1}\Bigg(\sum_{(s,a,h)}\Big(\ell_{\tau,h}(s,a)-\ell_{\tau-1,h}(s,a)\Big)\Bigg)\wedge 1. \tag{4.9}$$

Indeed, if the counter of no triplet (s,a,h) has reached a value of the form 2^r for some integer r by passing from episode $\tau-1$ to τ, then the summation in the minimum is zero, meaning that the episodes τ and $\tau+1$ belong to the same global epoch. On the contrary, if the counter of at least one (s,a,h) reached such a value, then this sum is at least 1 (there can be more than one triplets satisfying this), meaning that τ and $\tau+1$ belong to different global epochs. Thus, thanks to the minimum, the above quantity counts the number of (global) epoch switches from $\tau=1$ to $\tau=t$. In other words, the global epoch e_t is uniquely determined by the preceding profiles.

Since there are $SA(H-1)$ different triplets (s,a,h) and each such triplet is associated with at most $\lceil\log_2(T)\rceil$ doubling conditions, we obtain the following bound.

Lemma 4.10. *There are $m(T)\leqslant SAH\log_2(2T)$ global epochs.*

The second concept: optional skipping for estimated transition kernels. The trick detailed here is standard in the bandit and reinforcement-learning literature. The original reference is Theorem 5.2 of Doob (1953, Chapter III, p. 145); one can also check Chow and Teicher (1988, Section 5.3) for a more recent reference. A pedagogical exposition of the trick and of its uses in the bandit literature may be found in Garivier et al. (2022, Section 4.1), which we adapt now to the setting of reinforcement learning.

For each triplet $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$ and each integer $j\geqslant 1$, we denote by

$$N_{h,s,a,j}\triangleq\inf\Big\{t\geqslant 1 : n_{t,h}(s,a)=j\Big\}$$

(with the convention that the infimum of an empty set equals $+\infty$) the predictable stopping time whether and when (s,a) occurs for the j–th time. We are interested in the distribution of the states $s_{t,h+1}$ drawn at rounds t when $(s_{t,h},a_{t,h})=(s,a)$; these rounds are given by the stopping times $N_{h,s,a,j}$ introduced above. It turns out that these states are i.i.d. with distribution $P_h(\,\cdot\mid s,a)$. We also have independence across sequences of states. All these results are formally stated in the following lemma: to do so, one needs to set the values of the number of times $n_{t,h}(s,a)$ each triplet (h,s,a) was encountered till a given round.

Lemma 4.11 (Doob's optional skipping)**.** *Fix $t\geqslant 1$ and consider sequences of integers $J_{h,s,a}\geqslant 1$ and the intersection of events*

$$\mathcal{C}=\bigcap_{h\in[H-1]}\bigcap_{(s,a)\in\mathcal{S}\times\mathcal{A}}\Big\{n_{t,h}(s,a)=J_{h,s,a}\Big\}.$$

It holds that on $\mathcal{C}$, each of the sequences

$$(\widetilde{s}_{h,s,a,j})_{j\in[J_{h,s,a}]} \triangleq \left(s_{N_{h,s,a,j},h+1}\right)_{j\in[J_{h,s,a}]}$$

is formed by i.i.d. variables, with common distribution $P_h(\cdot \mid s,a)$. In addition, these sequences are independent from each other as (h,s,a) vary in $[H-1]\times\mathcal{S}\times\mathcal{A}$.

One of our applications of Lemma 4.11 will be the following, to handle term (**A**). The proof consists of noting first that on $\left\{\ell_{t-1,h}(s,a)=\ell\right\}$, the distribution $\widehat{P}_{t,h}(\cdot \mid s,a)$ corresponds to the empirical measure of the i.i.d. variables $\widetilde{s}_{h,s,a,j}$ with $1\leqslant j\leqslant 2^{\ell-1}$, and second, by dropping the indicator function.

Notation-wise, we will be using $\widetilde{s}_{h,s,a,j}$ (as in Lemma 4.11) for random variables generated by the MDP interactions and $\sigma_{h,s,a,j}$ (as in Corollary 4.12) for random variables independent from everything else and that are representations of the former.

Corollary 4.12. *Fix $h\in[H-1]$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$ and let $(\sigma_{h,s,a,j})_{j\geqslant 1}$ be a sequence of i.i.d. variables with distribution $P_h(\cdot \mid s,a)$. For all functions $\psi\colon\mathbb{R}\to[0,+\infty)$, all functions $g\colon\mathcal{S}\to\mathbb{R}$, all integers $\ell\geqslant 1$,*

$$\mathbb{E}\left[\psi\left(\widehat{P}_{t,h}g(s,a)\right)\mathbb{1}\left\{\ell_{t-1,h}(s,a)=\ell\right\}\right]\leqslant\mathbb{E}\left[\psi\left(\frac{1}{2^{\ell-1}}\sum_{j=1}^{2^{\ell-1}}g(\sigma_{h,s,a,j})\right)\right].$$

4.4.3 Term (A): Optimism

Term (**A**) is handled thanks to a result already present in the analysis of the `UCBVI` algorithm (Azar et al. 2017, Lemma 18), relying on an induction, and thanks to applications of Hoeffding's inequalities together with optional skipping. Appendix C.3 provides the (straightforward) details of the proof of the following lemma.

Lemma 4.13. *With probability at least $1-\delta$, for all $t\in[T]$ and all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$,*

$$Q_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t,\boldsymbol{P}}(s,a)\leqslant Q_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s,a)\,,\qquad V_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t,\boldsymbol{P}}(s)\leqslant V_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s)\,.$$

*Specifically, with probability $\geqslant 1-\delta$, we have (**A**) $\leqslant 0$.*

As in the original proof of Azar et al. (2017, Lemma 18), the result relies only on the concentration of the estimated transition kernel in the direction of $V_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t,\boldsymbol{P}}$ for any $(t,h)\in[T]\times[H]$. Crucially, it assumes an oblivious adversary, allowing $V_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t,\boldsymbol{P}}$ to be treated as deterministic. If the adversary depended on past actions and states, this assumption would break. The issue could then be addressed by concentrating $\widehat{\boldsymbol{P}}_t$ around $\boldsymbol{P}$ in ℓ_1–norm, incurring an additional $\sqrt{S}$ factor in $\boldsymbol{b}_t$ and propagating to the final regret via (**D**).

4.4.4 Term (D): Bonus Summation

Without the doubling trick, the exploration bonuses summed up along the trajectory can be classically bounded by a $O(\sqrt{T})$ term. The doubling trick introduces only minor changes to this classical step. Appendix C.4 provides the (straightforward) details of the proof of the following lemma, based on the Hoeffding–Azuma inequality together with simple controls of the form, for all $h\in[H-1]$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\sum_{t=1}^{T}\frac{1}{\sqrt{n_{t,h}(s,a)}}\mathbb{1}\left\{(s_{t,h},,a_{t,h})=(s,a)\right\}\mathbb{1}\left\{n_{t,h}(s,a)\geqslant 2\right\}=\sum_{n=2}^{n_{T,h}(s,a)}\frac{1}{\sqrt{n}}\leqslant 2\sqrt{n_{T,h}(s,a)}\,.$$

Lemma 4.14. *With probability at least* $1-\delta$*, we have*

$$(\mathbf{D}) \leqslant 7\sqrt{H^4SAT\ln(2SATH\log_2(2T)/\delta)} + H^2SA\,.$$

4.4.5 Term (C): Concentration

Let us start by formally stating the result, whose detailed proof may be found in Appendix C.5; below, we only sketch that proof. The analysis is essentially borrowed from Zhang et al. (2024) with minor technical modifications but a much simplified exposition (as we do not target optimized bounds yet). Also, we explain in Remark C.3 of Appendix C.5 that the dependency in H of the leading term in the upper bound of Lemma 4.15 could be improved to $\sqrt{H^6}$ with more efforts, but that there is no point in doing so, given the bound of Lemma 4.5, which scales with H as $\sqrt{H^7}$.

Lemma 4.15. *With probability at least* $1-\delta$*,*

$$(\mathbf{C}) \leqslant 2\sqrt{H^7SAT\Big(\log_2(2T)\Big)^3} + 2\sqrt{2H^6\,T\log_2(2T)\ln(4H/\delta)} + SAH^3\,.$$

Proof sketch. An application of the performance-difference lemma in case of different transition kernels (see, e.g., Russo 2019, Lemma 3) together with the Hoeffding–Azuma inequality first shows that with probability at least $1-\delta/2$, term (**C**) equals

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\sum_{h=1}^{H-1}\Big(\widehat{P}_{t,h}-P_h\Big)V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s_{t,h},a_{t,h})\,\middle|\,\boldsymbol{\pi}_t,\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t\right]\\
&\qquad\leqslant \sum_{h=1}^{H-1}\underbrace{\sum_{t=1}^{T}\Big(\widehat{P}_{t,h}-P_h\Big)V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s_{t,h},a_{t,h})}_{=\xi_{T,h}} + \sqrt{2H^5T\ln(2/\delta)}\,.
\end{aligned}$$

We bound the quantities $\xi_{T,h}$ for each fixed $h\in[H-1]$. We apply optional skipping in a careful way on the event $\mathcal{C}_{\ell,j,h,s,a,t}$ when $(s,a)\in\mathcal{S}\times\mathcal{A}$ is played for the $(2^{\ell-1}+j)$–th time in stage h at episode t: on $\mathcal{C}_{\ell,j,h,s,a,t}$,

$$\begin{aligned}
&\Big(\widehat{P}_{t,h}-P_h\Big)V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s_{t,h},a_{t,h}) \quad \text{behaves like}\\
&\frac{1}{2^{\ell-1}}\sum_{j\in[2^{\ell-1}]}\Big(\widetilde{V}_{s,a,h+1}(\sigma_{h,s,a,j})-P_h\widetilde{V}_{s,a,h+1}(s,a)\Big)\,,
\end{aligned}$$

for some random variable $\widetilde{V}_{s,a,h+1}$, where the $\sigma_{h,s,a,j}$ are i.i.d. according to $P_h(\cdot\mid s,a)$ and are independent from $\widetilde{V}_{s,a,h+1}$. The argument also extends between pairs (s,a) so that a careful application of the Hoeffding-Azuma inequality (this is the delicate part of the proof), together with the consideration of all values for ℓ and j, then shows that, with probability at least $1-\delta/2$, we have $\xi_{T,h}\leqslant$

$$\sum_{\ell=1}^{\lceil\log_2 T\rceil}\sum_{j=1}^{2^{\ell-1}}\sqrt{\frac{2H^4}{2^{\ell-1}}\sum_{(s,a)}\mathbb{1}\Big\{n_{T,h}(s,a)\geqslant 2^{\ell-1}+j\Big\}\ln\frac{1}{\delta'}}\,,$$

where δ' equals $\delta/2$ divided by the number of times we applied the union bound over ℓ, j, H, and in the course of optional skipping; we bound this number of times by $4H(T+1)^{1+SAH\lceil\log_2(T)\rceil}$.

We conclude the proof by two consecutive applications of Jensen's inequality for the root:

$$\sum_{\ell=1}^{\lceil \log_2 T \rceil} \sum_{j=1}^{2^{\ell-1}} \sqrt{\frac{2H^4}{2^{\ell-1}} \sum_{(s,a)} \mathbb{1}\Big\{n_{T,h}(s,a) \geqslant 2^{\ell-1}+j\Big\} \ln \frac{1}{\delta'}} \leqslant \sqrt{2H^4 \lceil \log_2 T \rceil \underbrace{\sum_{(s,a)} \sum_{\ell=1}^{\lceil \log_2 T \rceil} \sum_{j=1}^{2^{\ell-1}} \mathbb{1}\Big\{n_{T,h}(s,a) \geqslant 2^{\ell-1}+j\Big\}}_{\leqslant \sum_{(s,a)} n_{T,h}(s,a) = T} \ln \frac{1}{\delta'}}$$

together with some algebra. □

4.5 Conclusion and Limitations

In this chapter, we proposed an algorithm called `APO-MVP` that extends algorithm `MVP` of Zhang et al. (2024) and its analysis to the case of adversarial reward functions, thanks, in particular, to a black-box adversarial aggregation mechanism due to Jonckheere et al. (2025) that takes care of the adversarial nature of reward functions. Algorithm `APO-MVP` is easy to implement in practice as it relies on OLO learning strategies in the policy space combined with dynamic programming; it does not at all rely on so-called occupancy measures. Furthermore, it achieves a better regret bound compared to previous approaches based on the occupancy measures, reducing their regret bounds by a $\sqrt{S}$ multiplicative factor and narrowing the gap between the adversarial and stochastic regret bounds, which are both shown to be of order $\sqrt{SAT}$ up to logarithmic factors, as far as dependencies on S, A, and T are concerned.

We believe that this work opens many interesting follow-up questions. The two main open questions are inevitably linked with the main limitations of this chapter and are discussed below.

Limitations. The main limitation is rather a high dependency of our regret bound on the length H of the episodes, of order $\sqrt{H^7}$. Improving this dependency while maintaining a regret of order $\sqrt{SAT}$ up to logarithmic terms is one of the remaining open questions. We have only considered full monitoring; extending our approach to bandit monitoring seems to be non-trivial and it is still unknown if a $\sqrt{SAT}$ regret is possible in the adversarial case with bandit feedback.

Appendix A

Complements on Chapter 2

Contents

A.1 Notation

For any $(s,a) \in \mathcal{S}$, transition kernel $p(s,a) \in \mathcal{P}(\mathcal{S})$ and $f\colon \mathcal{S} \to \mathbb{R}$ define $pf(s,a) = \mathbb{E}_{p(s,a)}[f]$ and $\mathrm{Var}_p[f](s,a) = \mathrm{Var}_{p(s,a)}[f]$. For any $s \in \mathcal{S}$, policy $\pi(s) \in \mathcal{P}(\mathcal{S})$ and $f\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ define $\pi f(s) = \mathbb{E}_{a\sim\pi(s)}[f(s,a)]$ and $\mathrm{Var}_\pi f(s) = \mathrm{Var}_{a\sim\pi(s)}[f(s,a)]$.

Table A.1: Table of notation use throughout the chapter

Notation	**Meaning**
$\mathcal{S}$	state space of size S
$\mathcal{A}$	action space of size A
H	length of one episode
s_1	initial state
$\mathfrak{t}$	stopping time
$\mathcal{T}$	trajectory space, $\mathcal{T} \triangleq (\mathcal{S} \times \mathcal{A})^H$
ε	desired accuracy of solving the problem
δ	desired upper bound on failure probability
λ	regularization parameter in regularized MDPs (see Appendix A.4).
κ	weight of entropy in reward in regualrized MDPs (see Appendix A.4)
$p_h(s'\|s,a)$	probability transition
$r_h(s,a)$	reward function
$d_h^\pi(s,a)$	state-action visitation distribution at step h for the policy π
$q^\pi(m)$	visitation probability of trajectory $m \in \mathcal{T}$ by policy π
$\mathcal{K}_p$	polytope of all admissible state-action visitation distributions
$\mathcal{K}$	polytope of all admissible distributions over state-actions, $\mathcal{K} \triangleq (\Delta_{SA})^H$
$\mathcal{H}_{\text{visit}}(d^\pi)$	visitation entropy $\mathcal{H}_{\text{visit}}(d^\pi) \triangleq \sum_{h=1}^H \mathcal{H}(d_h^\pi)$ for $d^\pi \in \mathcal{K}_p$,
$\pi^{\star,\text{VE}}$	policy that maximizes $\mathcal{H}_{\text{visit}}(d^\pi)$, a solution to the MVEE problem
$\mathcal{H}_{\text{traj}}(q^\pi)$	trajectory entropy $\mathcal{H}_{\text{traj}}(q^\pi) \triangleq \mathcal{H}(q^\pi)$
$\pi^{\star,\text{TE}}$	policy that maximizes $\mathcal{H}_{\text{traj}}(q^\pi)$, a solution to the MTEE problem
n_0	number of prior visits for the forecaster-player in `EntGame`
t_0	total number of prior visits
s_h^t	state that was visited at h step during t episode
a_h^t	action that was picked at h step during t episode
$n_h^t(s,a)$	number of visits of state-action $n_h^t(s,a) = \sum_{k=1}^t \mathbb{1}\left\{(s_h^k, a_h^k) = (s,a)\right\}$
$n_h^t(s'\|s,a)$	number of transition to s' from state-action.
$\overline{n}_h^t(s,a)$	pseudo number of visits of state-action $\overline{n}_h^t(s,a) = n_h^t(s,a) + n_0$
$\widehat{p}_h^t(s'\|s,a)$	empirical probability transition $\widehat{p}_h^t(s'\|s,a) = n_h^t(s'\|s,a)/n_h^t(s,a)$
$\overline{d}_h^t(s,a)$	predicted distribution by the forecaster-player in `EntGame`
$\overline{Q}_h^t(s,a)$,$\overline{V}_h^t(s,a)$	for `EntGame`: upper bound on the optimal Q/V-functions
$Q_h^\pi(s,a)$, $V_h^\pi(s,a)$	Q- and V-functions for the MTEE problem
$Q_h^\star(s,a)$, $V_h^\star(s,a)$	optimal Q- and V-function for the MTEE problem
$\overline{Q}_h^t(s,a)$, $\overline{V}_h^t(s,a)$	for `UCBVI-Ent`: the upper bound on the optimal Q/V-functions
$\underline{Q}_h^t(s,a)$, $\underline{V}_h^t(s,a)$	for `UCBVI-Ent`: the lower bound on the optimal Q/V-functions
$Q_{\lambda,h}^\pi(s,a)$, $V_{\lambda,h}^\pi(s,a)$	Q- and V-functions in a regularized MDP
$Q_{\lambda,h}^\star(s,a)$, $V_{\lambda,h}^\star(s,a)$	optimal Q- and V-function in a regularized MDP
$\widehat{Q}_{\lambda,h}^\pi(s,a)$, $\widehat{V}_{\lambda,h}^\pi(s,a)$	for `RL-Explore-Ent`: the empirical Q-/V-functions in a regularized MDP

For a MDP $\mathcal{M}$,a policy π and a sequence of function f_h define $\mathbb{E}_\pi[\sum_{h'=h}^H f(s_{h'}, a_{h'})|s_h]$ as a conditional expectation of $\sum_{h'=h}^H f(s_{h'}, a_{h'})$ with respect to the sigma-algebra $\mathcal{F}_h = \sigma\{(s_{h'}, a_{h'})|h' \leqslant h\}$, where for any $h \in [H]$ we have $a_h \sim \pi(s_h)$, $s_{h+1} \sim p_h(s_h, a_h)$.

We write $f(S,A,H,\varepsilon) = \mathcal{O}(g(S,A,H,\varepsilon,\delta))$ if there exist $S_0, A_0, H_0, \varepsilon_0, \delta_0$ and constant $C_{f,g}$ such that for any $S \geqslant S_0, A \geqslant A_0, H \geqslant H_0, \varepsilon < \varepsilon_0, \delta < \delta_0, f(S,A,H,T,\delta) \leqslant C_{f,g} \cdot g(S,A,H,T,\delta)$.

We write $f(S,A,H,\varepsilon,\delta) = \widetilde{\mathcal{O}}(g(S,A,H,\varepsilon,\delta))$ if $C_{f,g}$ in the previous definition is poly-logarithmic in $S, A, H, 1/\varepsilon, 1/\delta$.

A.2 Proofs for Visitation Entropy

We first define the regrets of each players obtained by playing T times the games. For the forecaster-player, for any $\bar{d} \in \mathcal{K}$ we define

$$\mathfrak{R}^T_{\mathrm{Fore}}(\bar{d}) \triangleq \sum_{t=1}^{T} \sum_{h,s,a} \widetilde{d}^t_h(s,a) \left(\log \frac{1}{\bar{d}^t_h(s,a)} - \log \frac{1}{\bar{d}_h(s,a)} \right)$$

where $\widetilde{d}^t_h(s,a) \triangleq \mathbb{1}\Big\{(s^t_h, a^t_h) = (s,a)\Big\}$ is a sample from $d^{\pi^t}_h(s,a)$. Similarly for the sampler-player, for any $d \in \mathcal{K}_p$ we define

$$\mathfrak{R}^T_{\mathrm{Samp}}(d) \triangleq \sum_{t=1}^{T} \sum_{h,s,a} \Big(d_h(s,a) - d^{\pi^t}_h(s,a) \Big) \log \frac{1}{\bar{d}^t_h(s,a)} \,.$$

Recall that the visitation distribution of the policy π returned by `EntGame` is the average of the visitation distributions of the sampler-player $d^{\widehat{\pi}}_h(s,a) = \widehat{d}^T_h(s,a) \triangleq (1/T)\sum_{t=1}^T d^{\pi^t}_h(s,a)$. We also denote by $\mathring{d}^T_h(s,a) \triangleq (1/T)\sum_{t=1}^T \widetilde{d}^t(s,a)$ the average of the 'sample' visitation distributions.

We now relate the difference between the optimal visitation entropy and the visitation entropy of the outputted policy $\widehat{\pi}$ with the regrets of the two players. Indeed, using $\mathcal{H}(p) = \sum_{i\in[n]} p_i \log(1/q_i) - \mathrm{KL}(p,q) \leqslant \sum_{i\in[n]} p_i \log(1/q_i)$ for all $(p,q) \in (\Delta_n)^2$, we obtain

$$\begin{aligned}
T\Big(\mathcal{H}_{\mathrm{visit}}(d^{\pi^\star}) - \mathcal{H}_{\mathrm{visit}}(d^{\widehat{\pi}})\Big) &\leqslant \sum_{t=1}^{T} \sum_{h,a,s} d^{\pi^\star}_h(s,a) \log \frac{1}{\bar{d}^t_h(s,a)} - \widetilde{d}^t_h(s,a) \log \frac{1}{\mathring{d}^T_h(s,a)} \\
&\quad + T\Big(\mathcal{H}_{\mathrm{visit}}(\mathring{d}^T) - \mathcal{H}_{\mathrm{visit}}(\widehat{d}^T)\Big) \\
&= \mathfrak{R}^T_{\mathrm{Samp}}(d^{\pi^\star}) + \underbrace{\sum_{t=1}^{T} \sum_{h,s,a} \Big(d^{\pi^t}_h(s,a) - \widetilde{d}^t_h(s,a) \Big) \log \frac{1}{\bar{d}^t_h(s,a)}}_{\mathrm{Bias}_1} \\
&\quad + \mathfrak{R}^T_{\mathrm{Fore}}(\mathring{d}^T) + \underbrace{T\Big(\mathcal{H}_{\mathrm{visit}}(\mathring{d}^T) - \mathcal{H}_{\mathrm{visit}}(\widehat{d}^T)\Big)}_{\mathrm{Bias}_2} \,.
\end{aligned}$$

It remains to upper bound each terms separately in order to obtain a bound on the gap. We first bound the two regrets terms. The first bias term is martingale and can easily be bounded with a deviation inequality, whereas for the second one we introduce just instrumentally smoothing of the entropy.

A.2.1 Regret of the Forecaster-Player

We prove in this section a regret-bound for the mixture forecaster.

Lemma A.1. *For $n_0 = 1$, for any $\bar{d} \in \mathcal{K}$ it holds almost surely*

$$\mathfrak{R}^T_{\mathrm{Fore}}(\bar{d}) \leqslant HSA \log\Big(\mathrm{e}(T+1)\Big) - T\sum_{h=1}^{H} \mathrm{KL}(\mathring{d}^T_h, \bar{d}_h) \,.$$

Proof. We will bound the regret at step h,

$$\mathfrak{R}^T_{\mathrm{Fore},h}(\bar{d}) \triangleq \sum_{t=1}^{T}\sum_{s,a} \widetilde{d}^t_h(s,a)\left(\log\frac{1}{\bar{d}^t_h(s,a)} - \log\frac{1}{\bar{d}_h(s,a)}\right)$$

and then sum the upper bounds. Recall

$$\bar{d}^t_h(s,a) = \frac{n^{t-1}_h(s,a)+1}{t-1+SA},$$

and for $(s,a) = (s^t_h, a^t_h)$ and any $t \in [T], h \in [H]$ we have $n^{t-1}_h(s,a)+1 = n^t_h(s,a)$. Since $n_0 = 1$, we have $\bar{d}^t_h(s^t_h,a^t_h) = n^t_h(s^t_h,a^t_h)/(SA+t-1)$. Armed with this observation we can rewrite the regret as follows

$$\begin{aligned}\mathfrak{R}^T_{\mathrm{Fore},h}(\bar{d}) &= -T\,\mathrm{KL}(\mathring{d}^T_h, \bar{d}_h) - T\mathcal{H}(\mathring{d}^T_h) - \sum_{t=1}^{T}\log\left(\bar{d}^t_h(s^t_h,a^t_h)\right)\\ &= -T\,\mathrm{KL}(\mathring{d}^T_h, \bar{d}_h) - T\mathcal{H}(\mathring{d}^T_h) - \log\left(\prod_{t=1}^{T}\bar{d}^t_h(s^t_h,a^t_h)\right).\end{aligned}$$

Then we have an explicit formula for the product of $\bar{d}^t_h$

$$\begin{aligned}\prod_{t=1}^{T}\bar{d}^t_h(s^t_h,a^t_h) &= \prod_{t=1}^{T}\frac{n^t_h(s^t_h,a^t_h)}{SA+t-1} = \frac{(SA-1)!}{(SA+T-1)!}\prod_{(s,a)\in\mathcal{S}\times\mathcal{A}}[n^T_h(s,a)]!\\ &= \frac{1}{\binom{T}{(n^T_h(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}}}\frac{1}{\binom{T+SA-1}{SA-1}}\\ &\geqslant \exp\left(-T\mathcal{H}(\mathring{d}^T_h) - (T+SA-1)\mathcal{H}\left(\frac{SA-1}{T+SA-1}\right)\right)\end{aligned}$$

where in the last inequality we used Theorem 11.1.3 by Cover and Thomas (2006) and overload the entropy notation $\mathcal{H}(p) = -p\log(p) - (1-p)\log(1-p)$ for $p \in [0,1]$. Putting all together we get

$$\mathfrak{R}^T_{\mathrm{Fore},h}(\bar{d}) \leqslant (T+SA-1)\mathcal{H}\left(\frac{A-1}{T+A-1}\right) - T\,\mathrm{KL}(\mathring{d}^T_h, \bar{d}_h)\,.$$

Bounding the entropic term

$$\begin{aligned}(T+SA-1)\mathcal{H}\left(\frac{SA-1}{T+SA-1}\right) &= (SA-1)\log\frac{T+SA-1}{SA-1} + T\log\frac{T+SA-1}{T}\\ &\leqslant (SA-1)\log\frac{T+SA-1}{SA-1} + T\log\left(1+\frac{SA-1}{T}\right)\\ &\leqslant (SA-1)\log\frac{\mathrm{e}(T+SA-1)}{SA-1}\\ &\leqslant SA\log\left(\mathrm{e}(T+1)\right),\end{aligned}$$

and summing over h allows us to conclude. □

A.2.2 Regret of the Sampler-Player

We start from introducing new notation. Let $\mathcal{M}_t = (\mathcal{S}, \mathcal{A}, \{p_h\}_{h\in[H]}, \{r^t_h\}_{h\in[H]}, s_1)$ be a sequence of MDPs where reward defined as follows $r^t_h(s,a) = \log(1/\bar{d}^t_h(s,a))$. Define $Q^{\pi,t}_h(s,a)$ and $V^{\pi,t}_h(s,a)$

as a action-value and value functions of a policy π on a MDP $\mathcal{M}_t$. Notice that the value-function of initial state in this case could be written as follows

$$V_1^{\pi,t}(s_1) = \sum_{h,s,a} d_h^{\pi}(s,a) \log\left(\frac{1}{\bar{d}_h^t(s,a)}\right)$$

therefore, the regret for the sampler-player could be rewritten in the terms of the regret for this sequence of MDPs

$$\mathfrak{R}_{\mathrm{Samp}}^T(d^\pi) = \sum_{t=1}^{T} V_1^{\pi,t}(s_1) - V_1^{\pi^t,t}(s_1).$$

Since the rewards are changing in each episode and depending on the full history on interaction during previous episodes, we have to handle more uniform approach as in usual `UCBVI` proofs Azar et al. 2017.

Concentration. Let $\alpha^{\mathrm{KL}}, \alpha^{\mathrm{cnt}} : (0,1)\times\mathbb{R}_+ \to \mathbb{R}_+$ be some functions defined later on in Lemma A.2. We define the following favorable events

$$\mathcal{E}^{\mathrm{KL}}(\delta) \triangleq \left\{\forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S}\times\mathcal{A} : \quad \mathrm{KL}(\widehat{p}_h^t(s,a), p_h(s,a)) \leqslant \frac{\alpha^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}\right\},$$

$$\mathcal{E}^{\mathrm{cnt}}(\delta) \triangleq \left\{\forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S}\times\mathcal{A} : \quad n_h^t(s,a) \geqslant \frac{1}{2}\overline{n}_h^t(s,a) - \alpha^{\mathrm{cnt}}(\delta)\right\},$$

Lemma A.2. *For any $\delta \in (0,1)$ and for the following choices of functions α,*

$$\alpha^{\mathrm{KL}}(\delta, n) \triangleq \log(2SAH/\delta) + S\log(\mathrm{e}(1+n)), \quad \alpha^{\mathrm{cnt}}(\delta) \triangleq \log(2SAH/\delta),$$

it holds that

$$\mathbb{P}[\mathcal{E}^{\mathrm{KL}}(\delta)] \geqslant 1-\delta/2, \qquad \mathbb{P}[\mathcal{E}^{\mathrm{cnt}}(\delta)] \geqslant 1-\delta/2$$

In particular, $\mathbb{P}[\mathcal{G}(\delta)] \geqslant 1-\delta$.

Proof. Applying Theorem D.13 and the union bound over $h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$ we get $\mathbb{P}[\mathcal{E}^{\mathrm{KL}}(\delta)] \geqslant 1-\delta/2$. By Theorem D.14 and union bound, $\mathbb{P}[\mathcal{E}^{\mathrm{cnt}}(\delta)] \geqslant 1-\delta/2$. The union bound yields $\mathbb{P}[\mathcal{G}(\delta)] \geqslant 1-\delta$. □

Optimism. Next we define the exploration bonuses $b_h^t(s,a)$ for the sampler-player for $n_0 = 1$

$$b_h^t(s,a) = \sqrt{\frac{2H^2\log^2(t+SA)\cdot\alpha^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \tag{A.1}$$

Lemma A.3. *For any $t \in [T]$ and any policy π, the following holds on event $\mathcal{G}(\delta)$*

$$\overline{Q}_h^t(s,a) \geqslant Q_h^{\pi,t+1}(s,a), \qquad \overline{V}_h^t(s) \geqslant V_h^{\pi,t+1}(s).$$

Proof. Proceed by backward induction over h. For $h = H+1$ the statement trivially holds. Next we assume that the statement holds for any $h' > h$. Then we have by induction hypothesis and Hölder's inequality

$$\begin{aligned}\overline{Q}_h^t(s,a) - Q_h^{\pi,t+1}(s,a) &= \widehat{p}_h^t\overline{V}_{h+1}^t(s,a) - p_h V_h^{\pi,t+1}(s,a) + b_h^t(s,a)\\ &\geqslant [\widehat{p}_h^t - p_h]V_h^{\pi,t+1}(s,a) + b_h^t(s,a) \geqslant -\|V_h^{\pi,t+1}\|_\infty \|\widehat{p}_h^t - p_h\|_1 + b_h^t(s,a).\end{aligned}$$

The fact that $\|V_h^{\pi,t+1}\|_\infty \leqslant H\log(t+SA)$, Pinsker's inequality and the definition of the event $\mathcal{E}^{\mathrm{KL}}(\delta)$ yields

$$\|V_h^{\pi,t+1}\|_\infty \|\widehat{p}_h^t - p_h\|_1 \leqslant H\log(t+SA)\sqrt{\frac{2\alpha^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} = b_h^t(s,a)$$

that shows $\overline{Q}_h^t(s,a) - Q_h^{\pi,t+1}(s,a) \geqslant 0$. The inequality on V-functions could be derived as follows

$$\overline{V}_h^t(s) \geqslant \pi\overline{Q}_h^t(s) \geqslant \pi Q_h^{\pi,t+1}(s) = V_h^{\pi,t+1}(s).$$

□

Regret Bound.

Lemma A.4. *Let π be any fixed policy. Then for any $\delta \in (0,1)$ with probability at least $1-\delta$ the following holds*

$$\mathfrak{R}_{\mathrm{Samp}}^T(d^\pi) \leqslant 10\log(T+SA)\sqrt{2H^4SAT\cdot(\log(2SAH/\delta) + S\log(\mathrm{e}T))\log(T)}.$$

Proof. Assume that the event $\mathcal{G}(\delta)$ holds. By Lemma A.3 for any $t \in [T]$ and $h \in [H]$ we have

$$V_t^{\pi,t}(s_h) - V_h^{\pi^t,t}(s_h^t) \leqslant \overline{V}_h^{t-1}(s_h^t) - V_h^{\pi^t,t}(s_h^t) = \pi_h^t(\overline{Q}_h^{t-1} - Q_h^{\pi^t,t})(s),$$

thus we can define $\delta_h^t(s,a) = \overline{Q}_h^{t-1}(s,a) - Q_h^{\pi^t,t}(s,a)$ and upper bound the regret as follows

$$\mathfrak{R}_{\mathrm{Samp}}^T(d^\pi) \leqslant \sum_{t=1}^T \pi_1^t\delta_1^t(s_1).$$

Next we analyze $\delta_h^t(s_h^t)$. By the same argument as in Lemma A.3

$$\begin{aligned}\delta_h^t(s,a) &= [\widehat{p}_h^{t-1} - p_h]\overline{V}_{h+1}^{t-1}(s,a) + b_h^t(s,a) + p_h[\overline{V}_{h+1}^{t-1} - V_{h+1}^{\pi^t,t}](s,a)\\ &\leqslant 2b_h^{t-1}(s,a) + p_h\pi_{h+1}^t[\overline{Q}_{h+1}^{t-1} - Q_{h+1}^{\pi^t,t}](s,a)\end{aligned}$$

that could be rewritten as follows

$$\delta_h^t(s,a) \leqslant \mathbb{E}_{\pi^t}\Big[2b_h^{t-1}(s,a) + \delta_{h+1}^t(s_{h+1},a_{h+1})|(s_h,a_h) = (s,a)\Big],$$

thus, rolling out the initial bound on regret we have

$$\mathfrak{R}_{\mathrm{Samp}}^T(d^\pi) \leqslant H\log(T+SA)\sum_{t=1}^{T-1}\mathbb{E}_{\pi^t}\left[\sum_{h=1}^H 2\sqrt{\frac{2\alpha^{\mathrm{KL}}(\delta, n_h^t(s_h,a_h))}{n_h^t(s_h,a_h)}} \wedge 1\middle| s_1\right] + H\log(T+SA).$$

By Lemma D.5 and Jensen's inequality we have

$$\mathfrak{R}_{\mathrm{Samp}}^T(d^\pi) \leqslant 5H^{3/2}\log(T+SA)\sqrt{2T}\sqrt{\sum_{h,s,a}\sum_{t=1}^{T-1} d_h^{\pi^t}(s,a)\frac{\alpha^{\mathrm{KL}}(\delta, \overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a)\vee 1}}.$$

Notice that $d_h^{\pi^t}(s,a) = \overline{n}_h^{t+1}(s,a) - \overline{n}_h^t(s,a)$ and $\alpha^{\mathrm{KL}}(\delta, \overline{n}_h^t(s,a)) \leqslant \alpha^{\mathrm{KL}}(\delta, T-1)$. Combined with Lemma D.6 it implies

$$\mathfrak{R}_{\mathrm{Samp}}^T(d^\pi) \leqslant 10\log(T+SA)\sqrt{2H^4SAT\cdot(\log(2SAH/) + S\log(\mathrm{e}T))\log(T)}.$$

Finally, the fact that $\mathbb{P}[\mathcal{G}(\delta)] \geqslant 1-\delta$ concludes the statement of theorem. □

Remark A.5. It is possible to improve the H-dependence by introducing Bernstein-type bonuses, however, we are focused on improvement in a dependence in ε^{-1} and leave this regret bound as simple as possible.

A.2.3 Bias Terms

Lemma A.6. *Let $\delta \in (0,1)$ and $n_0 = 1$. Then with probability at least $1-\delta$ the following two bounds hold*

$$\mathrm{Bias}_1 \triangleq \sum_{t=1}^{T} \sum_{h,s,a} \left(d_h^{\pi^t}(s,a) - \widetilde{d}_h^t(s,a)\right) \log \frac{1}{\bar{d}_h^t(s,a)} \leqslant \log(T+SA)\sqrt{2TH\log(2/\delta)}$$

$$\mathrm{Bias}_2 \triangleq T(\mathcal{H}_{visit}(\mathring{d}^T) - \mathcal{H}_{visit}(\widehat{d}^T)) \leqslant \log(SAT)\left(\sqrt{2TH\log(2/\delta)} + 3H\sqrt{SAT\log(3T)}\right).$$

Proof. Let us define the lexicographic order on the set $[T] \times [H]$ with an additional convention $(t,0) = (t-1,H)$.

Then we can define a filtration $\mathcal{F}_{t,h} = \sigma\Big\{(s_{h'}^{t'}, a_{h'}^{t'})\ \forall t \leqslant t, \forall h' \in [H]\} \cup \{(s_{h'}^t, a_{h'}^t)\ \forall h' \leqslant h\}\Big\}$ that consists of the all history of interactions of the `EntGame` algorithm with an environment up to the h-th step of the episode t. The most important fact is that π^t and $\bar{d}_h^t(s,a)$ are $\mathcal{F}_{t,h-1}$-measurable for $h > 1$ and $\mathcal{F}_{t-1,H}$-measurable for $h = 1$.

Therefore, for any $t \in [T], h \in [H]$

$$\mathbb{E}\left[\sum_{s,a}(d_h^{\pi^t}(s,a) - \widetilde{d}_h^t(s,a)) \log \frac{1}{\bar{d}_h^t(s,a)} \middle| \mathcal{F}_{t,h-1}\right] = 0.$$

Therefore $X_{t,h} = \sum_{s,a}(d_h^{\pi^t}(s,a) - \widetilde{d}_h^t(s,a)) \log \frac{1}{\bar{d}_h^t(s,a)}$ is a martingale-difference sequence adapted to the filtration $\mathcal{F}_{t,h}$. Also we notice that a.s. the following bound holds

$$|X_{t,h}| \leqslant \log(T+SA)$$

All these facts combined with Azuma-Hoeffding inequality implies that with probability at least $1-\delta/2$

$$\mathrm{Bias}_1 = \sum_{t=1}^{T}\sum_{h=1}^{H} X_{t,h} \leqslant \log(T+SA)\sqrt{2TH\log(2/\delta)}.$$

To show the second part of the statement we notice that

$$\mathrm{Bias}_2 = T\sum_{h=1}^{H}(\mathcal{H}(\mathring{d}_h^T) - \mathcal{H}(\widehat{d}_h^T)).$$

Let us introduce the smoothed entropy as it was done by Hazan et al. (2019).

$$\forall d \in \Delta_{SA} : \mathcal{H}_\sigma(d) = \sum_{s,a} d(s,a) \log \frac{1}{d(s,a)+\sigma}.$$

The key difference with our approach and approach of Hazan et al. (2019) that we need the smoothing only instrumentally to provide a bound on Bias_2.

It is easy to see that $\mathcal{H}_\sigma$ is concave and, moreover $d \in \Delta_{SA}$

$$|\mathcal{H}(d) - \mathcal{H}_\sigma(d)| \leqslant \sum_{s,a} d(s,a) \log \frac{d(s,a)+\sigma}{d(s,a)} \leqslant \sigma SA,$$

where we used inequality $\log(1+x) \leqslant x$ for all $x \geqslant 0$, and also for $\sigma < \mathrm{e}^{-1}$

$$\|\nabla \mathcal{H}_\sigma(d)\|_\infty = \sup_{x \in (0,1)} \left|\log(x+\sigma) + \frac{x}{x+\sigma}\right| \leqslant \log(\sigma^{-1}).$$

By replacing an entropy with a smoothed entropy

$$\text{Bias}_2 \leqslant T \sum_{h \in H} (\mathcal{H}_\sigma(\mathring{d}_h^T) - \mathcal{H}_\sigma(\widehat{d}_h^T)) + 2\sigma \cdot TSAH.$$

To analyze the first term we use that $\mathcal{H}_\sigma$ is concave, therefore

$$\sum_{h=1}^H \mathcal{H}_\sigma(\mathring{d}_h^T) - \mathcal{H}_\sigma(\widehat{d}_h^T) \leqslant \sum_{h=1}^H \langle \nabla \mathcal{H}_\sigma(\widehat{d}_h^T), \mathring{d}_h^T - \widehat{d}_h^T \rangle = \frac{1}{T} \sum_{s,a} \sum_{t=1}^T \sum_{h=1}^H (\widetilde{d}_h^t(s,a) - d_h^{\pi^t}(s,a)) \cdot \nabla \mathcal{H}_\sigma(\widehat{d}_h^T)(s,a)$$

For this term situation is more involved than for Bias_1 because $\widehat{d}_h^T$ is dependent on all generated policies. Therefore we have to preform uniform bounds. Define $\mathcal{W} = \{w \in \mathbb{R}^{HSA} \mid |w_h(s,a)| \leqslant 1\}$ as a unit ℓ_∞-ball in $\mathbb{R}^{HSA}$. Then we have

$$T \sum_{h=1}^H \mathcal{H}_\sigma(\mathring{d}_h^T) - \mathcal{H}_\sigma(\widehat{d}_h^T) \leqslant \log(\sigma^{-1}) \cdot \sup_{w \in \mathcal{W}} \sum_{t=1}^T \sum_{h=1}^H \left(\sum_{s,a} (\widetilde{d}_h^t(s,a) - d_h^{\pi^t}(s,a)) \cdot w_h(s,a) \right).$$

Define $N(\varepsilon, \|\cdot\|_\infty, \mathcal{W})$ as ε-covering number for a set $\mathcal{W}$ with ℓ_∞-norm as a distance, and $\mathcal{W}_\varepsilon$ as a minimal ε-net. Next we can use the well-known result on upper bound on the covering number (e.g. see Exercise 5.5 by Handel (2016))

$$N(\varepsilon, \|\cdot\|_\infty, \mathcal{W}) \leqslant \left(\frac{3}{\varepsilon}\right)^{SAH},$$

and replace our maximization problem with the maximization over ε-net

$$\sup_{w \in \mathcal{W}} \sum_{t=1}^T \sum_{h=1}^H \left(\sum_{s,a} (\widetilde{d}_h^t(s,a) - d_h^{\pi^t}(s,a)) \cdot w_h(s,a) \right) \leqslant \sup_{\widehat{w} \in \mathcal{W}_\varepsilon} \sum_{t=1}^T \sum_{h=1}^H \left(\sum_{s,a} (\widetilde{d}_h^t(s,a) - d_h^{\pi^t}(s,a)) \cdot \widehat{w}_h(s,a) \right) + \varepsilon TH.$$

For any fixed $\widehat{w} \in \mathcal{W}_\varepsilon$ we apply Azuma-Hoeffding inequality exactly in the same manner as in the bound for Bias_1-term. We have that with probability at least $1 - \delta/(2N_\varepsilon)$ for $N = N(\varepsilon, \|\cdot\|_\infty, \mathcal{W})$ we have

$$\sum_{t=1}^T \sum_{h=1}^H \left(\sum_{s,a} (\widetilde{d}_h^t(s,a) - d_h^{\pi^t}(s,a)) \cdot \widehat{w}_h(s,a) \right) \leqslant \sqrt{2TH \log(2N_\varepsilon/\delta)}.$$

Thus, by union bound we have with probability at least $1 - \delta/2$

$$\sup_{w \in \mathcal{W}} \sum_{t=1}^T \sum_{h=1}^H \left(\sum_{s,a} (\widetilde{d}_h^t(s,a) - d_h^{\pi^t}(s,a)) \cdot w_h(s,a) \right) \leqslant \sqrt{2TH(\log(2/\delta) + SAH \log(3/\varepsilon))} + \varepsilon TH.$$

Taking $\varepsilon = 1/T$ we have

$$\text{Bias}_2 \leqslant \log(\sigma^{-1})(\sqrt{2TH(\log(2/\delta) + SAH \log(3T))} + H) + \sigma SATH.$$

Next we choose $\sigma = 1/SAT$ and by inequality $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ obtain

$$\text{Bias}_2 \leqslant \log(SAT)\left(\sqrt{2TH \log(2/\delta)} + 3H\sqrt{SAT \log(3T)}\right).$$

□

A.2.4 Proof of Theorem 2.2

We state the version of this theorem with all prescribed dependencies factors.

Theorem A.7. *For all $\varepsilon > 0$ and $\delta \in (0,1)$. For $n_0 = 1$ and*

$$T \geqslant 1 + \frac{648(\log(SA) + L)H^4SA \cdot (\log(4SAH/\delta) + S + L) \cdot L}{\varepsilon^2} + \frac{2HSA(2+L)}{\varepsilon}$$

for $L = 9\log\Big(1010\sqrt{H^4S^{8/3}A^{8/3}\log(4SAH/\delta)}/\varepsilon\Big)$ the algorithm EntGame *is (ε,δ)-PAC.*

Proof. We start from writing down the decomposition defined in the beginning of the appendix

$$T(\mathcal{H}_{\text{visit}}(d^{\pi^\star,\text{VE}}) - \mathcal{H}_{\text{visit}}(d^{\widehat{\pi}})) \leqslant \mathfrak{R}^T_{\text{Samp}}(d^{\pi^\star,\text{VE}}) + \mathfrak{R}^T_{\text{Fore}}(\mathring{d}^T) + \text{Bias}_1 + \text{Bias}_2.$$

By Lemma A.4 with probability at least $1-\delta/2$ it holds

$$\mathfrak{R}^T_{\text{Samp}}(d^{\pi^\star,\text{VE}}) = 10\log(T+SA)\sqrt{2H^4SAT \cdot (\log(4SAH/\delta) + S\log(\text{e}T))\log(T)}$$

By Lemma A.1

$$\mathfrak{R}^T_{\text{Fore}}(\mathring{d}^T) \leqslant HSA\log\Big(\text{e}(T+1)\Big).$$

By Lemma A.6 with probability at least $1-\delta/2$

$$\text{Bias}_1 + \text{Bias}_2 \leqslant 3\log(SAT)\Big(\sqrt{TH\log(4/\delta)} + H\sqrt{SAT\log(3T)}\Big).$$

By union bound all these inequalities hold simultaneously with probability at least $1-\delta$. Combining all these bounds we get

$$\begin{aligned} T(\mathcal{H}_{\text{visit}}(d^{\pi^\star,\text{VE}}) - \mathcal{H}_{\text{visit}}(d^{\widehat{\pi}})) \leqslant{}& 18\log(SAT)\sqrt{H^4SAT(\log(4SAH/\delta) + S\log(\text{e}T))\log(T)} \\ &+ HSA\log(\text{e}(T+1)). \end{aligned}$$

Therefore, it is enough to choose T such that $\mathcal{H}_{\text{visit}}(d^{\pi^\star,\text{VE}}) - \mathcal{H}_{\text{visit}}(d^{\widehat{\pi}})$ is guaranteed to be less than ε. In this case EntGame become automatically (ε,δ)-PAC.

It is equivalent to find a maximal T such that

$$\varepsilon T \leqslant 18\log(SAT)\sqrt{H^4SAT(\log(4SAH/\delta) + S\log(\text{e}T))\log(T)} + HSA\log(\text{e}(T+1))$$

and add 1 to it.

We start from obtaining a loose bound to eliminate logarithmic factors in T.

First, we assume that $T \geqslant 1$, thus $T+1 \leqslant 2T$. Additionally, let us use inequality $\log(x) \leqslant x^\beta/\beta$ for any $x > 0$ and $\beta > 0$. We obtain

$$\begin{aligned} \varepsilon T &\leqslant 18\frac{(SAT)^{1/3}}{1/3}\sqrt{H^4S^2AT\log(4SAH/\delta)\frac{(\text{e}T)^{1/18}}{1/18}\frac{T^{1/18}}{1/18}} + HSA\frac{(2\text{e}T)^{8/9}}{8/9} \\ &\leqslant T^{8/9}\Big(1010\sqrt{H^4S^2A^2\log(2SAH/\delta)}\Big), \end{aligned}$$

thus we can define $L = 9\log\Big(1010\sqrt{H^4S^{8/3}A^{8/3}\log(4SAH/\delta)}/\varepsilon\Big)$ for which $\log(T) \leqslant L$. Thus we have

$$\varepsilon T \leqslant 18(\log(SA) + L)\sqrt{H^4SAT(\log(4SAH/\delta) + S + L)L} + HSA(2+L).$$

Solving this quadratic inequality, we obtain the minimal required T to guarantee $\mathcal{H}_{\text{visit}}(d^{\pi^\star,\text{VE}}) - \mathcal{H}_{\text{visit}}(d^{\widehat{\pi}}) \leqslant \varepsilon$. In particular,

$$T \geqslant 1 + \frac{648(\log(SA) + L)H^4SA \cdot (\log(4SAH/\delta) + S + L) \cdot L}{\varepsilon^2} + \frac{2HSA(2+L)}{\varepsilon}.$$

□

A.3 Regularized Bellman Equations

In this section we provide complete proofs for regularized Bellman Equations in the general setting. Let $\Phi\colon \Delta_{\mathcal{A}} \to \mathbb{R}$ be a strictly convex function.

Then we can define the regularized value function as follows

$$V^{\pi}_{\lambda,h}(s) \triangleq \mathbb{E}_{\pi}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) - \lambda\Phi(\pi_{h'}(s_{h'})) \mid s_h = s\right]. \tag{A.2}$$

Notably, for a specific choice of rewards $r_h(s,a) = \mathcal{H}(p_h(s,a))$, the regularizer is equal to the negative entropy $\Phi(\pi) = -\mathcal{H}(\pi)$, and $\lambda = 1$ we have $V^{\pi}_{\lambda,1}(s_1) = \mathcal{H}_{\text{traj}}(q^{\pi})$, see Lemma D.2 In more general setting let $r_h(s,a)$ be equal to the sum of deterministic reward and $\lambda\mathcal{H}(p_h(s,a))$. In this case we have $V^{\pi}_{\lambda,1}(s_1) = V^{\pi}_1(s_1) + \lambda\mathcal{H}_{\text{traj}}(q^{\pi})$ in terms of a usual non-regularized value function.

Let us define an entropy action-value function as follows

$$Q^{\pi}_{\lambda,h}(s,a) \triangleq \mathbb{E}_{\pi}\left[r_h(s_h,a_h) + \sum_{h'=h+1}^{H} [r_{h'}(s_{h'},a_{h'}) - \lambda\Phi(\pi_h(s_{h'}))] \mid (s_h,a_h) = (s,a)\right]. \tag{A.3}$$

Additionally, we define an optimal entropy-regularized value functions a follows

$$V^{\star}_{\lambda,h}(s) \triangleq \max_{\pi} V^{\pi}_{\lambda,h}(s), \quad Q^{\star}_{\lambda,h}(s,a) \triangleq \max_{\pi} Q^{\pi}_{\lambda,h}(s,a) \quad \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H].$$

A.3.1 Proof of Entropy-Regularized Bellman Equations

Theorem A.8 (Regularized Bellman Equations)**.** *For any stochastic policy π the following decomposition of the entropy-regularized value function holds*

$$\begin{aligned} Q^{\pi}_{\lambda,h}(s,a) &= r_h(s,a) + p_h V^{\pi}_{\lambda,h+1}(s,a), \\ V^{\pi}_{\lambda,h}(s) &= \pi_h Q^{\pi}_{\lambda,h}(s) - \lambda\Phi(\pi_h(s)). \end{aligned} \tag{A.4}$$

Moreover, for optimal Q- and V-functions we have

$$\begin{aligned} Q^{\star}_{\lambda,h}(s,a) &= r_h(s,a) + p_h V^{\star}_{\lambda,h+1}(s,a), \\ V^{\star}_{\lambda,h}(s) &= \max_{\pi\in\Delta_{\mathcal{A}}}\{\pi Q^{\star}_h(s) - \lambda\Phi(\pi)\}. \end{aligned} \tag{A.5}$$

Remark A.9. For the case of interest $\Phi(\pi) = -\mathcal{H}(\pi)$ the expression for the V-function allows the closed-form formula by a well-known LogSumExp smooth maximum approximation

$$V^{\star}_{\lambda,h}(s) = \lambda\log\left(\sum_{a\in\mathcal{A}} \exp\left(\frac{1}{\lambda}Q^{\star}_{\lambda,h}(s,a)\right)\right),$$

and as $\lambda \to 0$ we see that entropy-regularized value function tends to a usual value function without regularization.

Proof. We proceed by induction. For $h = H+1$ the equation is trivial. By definition and tower property of conditional expectation

$$\begin{aligned} Q^{\pi}_{\lambda,h}(s,a) &= r_h(s,a) + \mathbb{E}\left[\sum_{t=h+1}^{H} r_t(s_t,a_t) - \lambda\Phi(\pi_t(s_t))\middle| s_h = s, a_h = a\right] \\ &= r(s,a) + \mathbb{E}\left[\mathbb{E}\left[\sum_{t=h+1}^{H} r_t(s_t,a_t) - \lambda\Phi(\pi_t(s_t))\middle| s_{h+1}\right]\middle| s_h = s, a_h = a\right] \\ &= r_h(s,a) + p_h V^{\pi}_{\lambda,h+1}(s,a). \end{aligned}$$

Next we provide the second Bellman equation by tower property and the definition of the regularized Q-function

$$\begin{aligned} V^{\pi}_{\lambda,h}(s) &= -\lambda\Phi(\pi_h(s)) + \mathbb{E}\left[r_h(s_h,a_h) + \sum_{t=h+1}^{H} r_t(s_t,a_t) - \lambda\Phi(\pi_t(s_t)) \middle| s_h = s \right] \\ &= \pi_h Q^{\pi}_{\lambda,h}(s) - \lambda\Phi(\pi_h(s)). \end{aligned}$$

For optimal Bellman equation we proceed by induction. For $h = H+1$ the equation is also trivial. By Bellman equations

$$Q^{\star}_{\lambda,h}(s,a) = \max_{\pi}\Big\{ r_h(s,a) + p_h V^{\pi}_{\lambda,h+1}(s,a) \Big\} = r_h(s,a) + p_h V^{\star}_{\lambda,h+1}(s,a),$$

and, finally

$$V^{\star}_{\lambda,h}(s) = \max_{\pi_1,\dots,\pi_H \in \Delta(\mathcal{A})} \Big\{ \pi_h Q^{\star}_{\lambda,h}(s) - \lambda\Phi(\pi_h(s)) \Big\} = \max_{\pi\in\Delta_{\mathcal{A}}} \Big\{ \pi Q^{\star}_{\lambda,h}(s) - \lambda\Phi(\pi) \Big\}.$$

□

A.3.2 A Bellman-type Equations for Variance

For a stochastic policy π we define Bellman-type equations for the variances as follows

$$\begin{aligned} \sigma Q^{\pi}_{\lambda,h}(s,a) &\triangleq \mathrm{Var}_{p_h} V^{\pi}_{\lambda,h+1}(s,a) + p_h \sigma V^{\pi}_{\lambda,h+1}(s,a) \\ \sigma V^{\pi}_{\lambda,h}(s) &\triangleq \mathrm{Var}_{\pi_h} Q^{\pi}_{\lambda,h}(s) + \pi_h \sigma Q^{\pi}_{\lambda,h}(s) \\ \sigma V^{\pi}_{\lambda,H+1}(s) &\triangleq 0, \end{aligned}$$

where $\mathrm{Var}_{p_h}(f)(s,a) \triangleq \mathbb{E}_{s'\sim p_h(\cdot|s,a)}\Big[(f(s') - p_h f(s,a))^2\Big]$ denotes the *variance operator over transitions* and $\mathrm{Var}_{\pi_h}(f)(s) \triangleq \mathbb{E}_{a'\sim\pi_h(s)}\Big[(f(s,a') - \pi_h f(s))^2\Big]$ denoted the *variance operator over the policy*. In particular, the function $s \mapsto \sigma V^{\pi}_{\lambda,1}(s)$ represents the average sum of the local variances $\mathrm{Var}_{p_h} V^{\pi}_{\lambda,h+1}(s,a)$ and $\mathrm{Var}_{\pi_h} Q^{\pi}_{\lambda,h}(s)$ over a trajectory following the policy π, starting from (s,a). Indeed, the definition above implies that

$$\sigma V^{\pi}_{\lambda,1}(s_1) = \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d^{\pi}_h(s)\mathrm{Var}_{\pi_h} Q^{\pi}_{\lambda,h}(s) + \sum_{h=1}^{H}\sum_{s,a} d^{\pi}_h(s,a)\mathrm{Var}_{p_h}(V^{\pi}_{\lambda,h+1})(s,a).$$

The lemma below shows that we can relate the global variance of the cumulative reward over a trajectory to the average sum of local variances.

Lemma A.10 (Law of total variance)**.** *Let us define* $\widetilde{r}_h(s,a,\pi) = r_h(s,a) - \lambda\Phi(\pi(s))$*. Then, for any stochastic policy* π *and for all* $h \in [H]$*,*

$$\begin{aligned} \sigma Q^{\pi}_{\lambda,h}(s,a) &= \mathbb{E}_{\pi}\left[\left(\sum_{h'=h}^{H} \widetilde{r}_{h'}(s_{h'},a_{h'},\pi_{h'}) - \Big(Q^{\pi}_{\lambda,h}(s_h,a_h) - \lambda\Phi(\pi_h(s_h)) \Big) \right)^2 \middle| (s_h,a_h) = (s,a) \right], \\ \sigma V^{\pi}_{\lambda,h}(s) &= \mathbb{E}_{\pi}\left[\left(\sum_{h'=h}^{H} \widetilde{r}_{h'}(s_{h'},a_{h'},\pi_{h'}) - V^{\pi}_{\lambda,h}(s_h) \right)^2 \middle| s_h = s \right]. \end{aligned}$$

Proof. We proceed by induction. The statement in Lemma A.10 is trivial for $h = H + 1$. We now assume that it holds for $h + 1$ and prove that it also holds for h. For this purpose, we compute

$$\begin{aligned}
&\mathbb{E}_\pi\left[\left(\sum_{h'=h}^{H} \widetilde{r}_{h'}(s_{h'}, a_{h'}, \pi_{h'}) - \left(Q^\pi_{\lambda,h}(s_h, a_h) - \lambda\Phi(\pi_h(s_h))\right)\right)^2 \middle| (s_h, a_h)\right] \\
&= \mathbb{E}_\pi\left[\left(V^\pi_{\lambda,h+1}(s_{h+1}) - p_h V^\pi_{\lambda,h+1}(s_h, a_h) + \sum_{h'=h+1}^{H} \widetilde{r}_{h'}(s_{h'}, a_{h'}, \pi_{h'}) - V^\pi_{\lambda,h+1}(s_{h+1})\right)^2 \middle| (s_h, a_h)\right] \\
&= \mathbb{E}_\pi\left[\left(V^\pi_{\lambda,h+1}(s_{h+1}) - p_h V^\pi_{\lambda,h+1}(s_h, a_h)\right)^2 \middle| (s_h, a_h)\right] \\
&+ \mathbb{E}_\pi\left[\left(\sum_{h'=h+1}^{H} \widetilde{r}_{h'}(s_{h'}, a_{h'}, \pi_{h'}) - V^\pi_{\lambda,h+1}(s_{h+1})\right)^2 \middle| (s_h, a_h)\right] \\
&+ 2\mathbb{E}_\pi\left[\left(\sum_{h'=h+1}^{H} \widetilde{r}_{h'}(s_{h'}, a_{h'}, \pi_{h'}) - V^\pi_{\lambda,h+1}(s_{h+1})\right)\left(V^\pi_{\lambda,h+1}(s_{h+1}) - p_h V^\pi_{\lambda,h+1}(s_h, a_h)\right) \middle| (s_h, a_h)\right].
\end{aligned}$$

The definition of $V^\pi_{\lambda,h+1}(s_{h+1})$ implies that

$$\mathbb{E}_\pi\left[\sum_{h'=h+1}^{H} (r_{h'}(s_{h'}, a_{h'}) - \lambda\Phi(\pi_{h'}(s_{h'}))) - V^\pi_{\lambda,h+1}(s_{h+1}) \middle| s_{h+1}\right] = 0.$$

Therefore, the tower property of conditional expectation gives us

$$\begin{aligned}
&\mathbb{E}_\pi\left[\left(\sum_{h'=h}^{H} \widetilde{r}_{h'}(s_{h'}, a_{h'}, \pi_{h'}) - \left(Q^\pi_{\lambda,h}(s_h, a_h) - \lambda\Phi(\pi_h(s_h))\right)\right)^2 \middle| (s_h, a_h)\right] \\
&= \mathbb{E}_\pi\left[\left(V^\pi_{\lambda,h+1}(s_{h+1}) - p_h V^\pi_{\lambda,h+1}(s_h, a_h)\right)^2 \middle| (s_h, a_h)\right] \\
&+ \mathbb{E}\left[\mathbb{E}_\pi\left[\left(\sum_{h'=h+1}^{H} \widetilde{r}_{h'}(s_{h'}, a_{h'}, \pi_{h'}) - V^\pi_{\lambda,h+1}(s_{h+1})\right)^2 \middle| s_{h+1}\right] \middle| (s_h, a_h)\right] \\
&= \mathrm{Var}_{p_h} V^\pi_{\lambda,h+1}(s_h, a_h) + p_h \sigma V^\pi_{\lambda,h+1}(s_h, a_h) = \sigma Q^\pi_{\lambda,h}(s_h, a_h)
\end{aligned}$$

where in the third equality we used the inductive hypothesis and the definition of σV^π_{h+1}. To prove the second equation we use the entropy-regularized Bellman equations

$$\begin{aligned}
&\mathbb{E}_\pi\left[\left(\sum_{h'=h}^{H} \widetilde{r}_{h'}(s_{h'}, a_{h'}, \pi_{h'}) - V^\pi_{\lambda,h}(s_h)\right)^2 \middle| s_h = s\right] \\
&= \mathbb{E}_\pi\left[\left(\sum_{h'=h}^{H} \widetilde{r}_{h'}(s_{h'}, a_{h'}, \pi_{h'}) - (Q^\pi_{\lambda,h}(s_h, a_h) - \lambda\Phi(\pi_h(s_h)))\right)^2 \middle| s_h = s\right] \\
&\quad + 2\mathbb{E}_\pi\left[\left(\sum_{h'=h}^{H} \widetilde{r}_{h'}(s_{h'}, a_{h'}, \pi_{h'}) - (Q^\pi_{\lambda,h}(s_h, a_h) - \lambda\Phi(\pi_h(s_h)))\right) \times \right. \\
&\qquad\qquad \left. \times \left(\pi_h Q^\pi_{\lambda,h}(s_h) - Q^\pi_{\lambda,h}(s_h, a_h)\right) \middle| s_h = s\right] \\
&\quad + \mathbb{E}_\pi\left[\left(\pi_h Q^\pi_{\lambda,h}(s_h) - Q^\pi_{\lambda,h}(s_h, a_h)\right)^2 \middle| s_h = s\right].
\end{aligned}$$

By definition of $Q^{\pi}_{\lambda,h}$ we have

$$\mathbb{E}_{\pi}\left[\left(\sum_{h'=h}^{H} \widetilde{r}_{h'}(s_{h'}, a_{h'}, \pi_{h'}) - (Q^{\pi}_{\lambda,h}(s_h, a_h) - \lambda\Phi(\pi_h(s_h)))\right)\middle|(s_h, a_h) = (s, a)\right] = 0,$$

thus, by the tower property

$$\mathbb{E}_{\pi}\left[\left(\sum_{h'=h}^{H} \widetilde{r}_{h'}(s_{h'}, a_{h'}, \pi_{h'}) - V^{\pi}_{\lambda,h}(s_h)\right)^2\middle|s_h = s\right] = \pi_h\sigma Q^{\pi}_{\lambda,h}(s_h) + \mathrm{Var}_{\pi_h} Q^{\pi}_{\lambda,h}(s_h) = \sigma V^{\pi}_{\lambda,h}(s_h).$$

□

A.3.3 Performance-Difference Lemma

In this section we provide a version of performance-difference lemma (see e.g. Lemma E.15 by Dann et al. (2017)) for regularized Bellman equations.

Lemma A.11 (Performance-Difference Lemma)**.** *Let $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, H, \{p'_h\}_{h\in[H]}, \{r'_h\}_{h\in[H]}, s_1)$ and $\mathcal{M}'' = (\mathcal{S}, \mathcal{A}, H, \{p''_h\}_{h\in[H]}, \{r''_h\}_{h\in[H]}, s_1)$ be two MDPs and let $Q^{\mathcal{M},\pi}_{\lambda,h}(s, a)$ be a Q-value of policy π in the MDP $\mathcal{M}$ with regularization. Then for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$,*

$$\begin{aligned} Q^{\mathcal{M}',\pi}_{\lambda,h}(s, a) - Q^{\mathcal{M}'',\pi}_{\lambda,h}(s, a) = \mathbb{E}_{\mathcal{M}'',\pi}&\left[\sum_{h'=h}^{H} r'_{h'}(s_{h'}, a_{h'}) - r''_{h'}(s_{h'}, a_{h'}) \;\middle|\; (s_h, a_h) = (s, a)\right] \\ &+ \mathbb{E}_{\mathcal{M}'',\pi}\left[\sum_{h'=h}^{H} [p'_{h'} - p''_{h'}]V^{\mathcal{M}',\pi}_{\lambda,h'+1}(s_{h'}, a_{h'}) \;\middle|\; (s_h, a_h) = (s, a)\right]. \end{aligned}$$

Proof. Let us proceed by induction over h. For $h = H + 1$ this statement is trivially true. Next we assume that it holds for any $h' > h$. By regularized Bellman equations

$$\begin{aligned} Q^{\mathcal{M}',\pi}_{\lambda,h}(s, a) - Q^{\mathcal{M}'',\pi}_{\lambda,h}(s, a) &= [r'_h(s, a) - r''_h(s, a)] + p'_h V^{\mathcal{M}',\pi}_{\lambda,h}(s, a) - p''_h V^{\mathcal{M}'',\pi}_{\lambda,h}(s, a) \\ &= [r'_h(s, a) - r''_h(s, a)] + [p'_h - p''_h]V^{\mathcal{M}',\pi}_{\lambda,h+1}(s, a) - p''_h\left[V^{\mathcal{M}'',\pi}_{\lambda,h+1} - V^{\mathcal{M}',\pi}_{\lambda,h+1}\right](s, a) \\ &= [r'_h(s, a) - r''_h(s, a)] + [p'_h - p''_h]V^{\mathcal{M}',\pi}_{\lambda,h+1}(s, a) + p''_h\left[V^{\mathcal{M}',\pi}_{\lambda,h+1} - V^{\mathcal{M}'',\pi}_{\lambda,h+1}\right](s, a). \end{aligned}$$

Next we notice that

$$V^{\mathcal{M}',\pi}_{\lambda,h+1}(s) - V^{\mathcal{M}'',\pi}_{\lambda,h+1}(s) = \pi Q^{\mathcal{M}',\pi}_{\lambda,h+1}(s) - \lambda\Phi(\pi) - \pi Q^{\mathcal{M}'',\pi}_{\lambda,h+1}(s) + \lambda\Phi(\pi) = \pi\left[Q^{\mathcal{M}',\pi}_{\lambda,h+1} - Q^{\mathcal{M}'',\pi}_{\lambda,h+1}\right](s)$$

since the regularizer is cancelled out. Thus, we can rewrite this difference as follows

$$\begin{aligned} Q^{\mathcal{M}',\pi}_{\lambda,h}(s, a) - Q^{\mathcal{M}'',\pi}_{\lambda,h}(s, a) = \mathbb{E}_{\pi,\mathcal{M}''}&\Big[r'_h(s_h, a_h) - r''_h(s_h, a_h) + [p'_h - p''_h]V^{\mathcal{M}',\pi}_{\lambda,h+1}(s_h, a_h) \\ &+ \left[Q^{\mathcal{M}',\pi}_{\lambda,h+1}(s_{h+1}, a_{h+1}) - Q^{\mathcal{M}'',\pi}_{\lambda,h+1}(s_{h+1}, a_{h+1})\right] \;\Big|\; (s_h, a_h) = (s, a)\Big]. \end{aligned}$$

By induction hypothesis we conclude the statement. □

A.4 Sample Complexity for MTEE and Regularized MDPs

In this section we describe the general setting of regularized MDPs, not only entropy-regularized.

A.4.1 Preliminaries

First we define class of regularizers we are interested in. For more exposition on this definition, see Bubeck (2015).

Definition A.12. Let $\Phi\colon \Delta(\mathcal{A}) \to \mathbb{R}$ be a proper closed strongly-convex function. We will call Φ a mirror-map if the following holds

- Φ is 1-strongly convex with respect to norm $\|\cdot\|$;
- $\nabla\Phi$ takes all possible values in $\mathbb{R}^{\mathcal{A}}$;
- $\nabla\Phi$ diverges on the boundary of $\Delta(\mathcal{A})$: $\lim_{x\in\partial\Delta(\mathcal{A})}\|\nabla\Phi(x)\| = +\infty$;

We explain three main examples of a such regularizers.

- The negative Shannon entropy $\Phi(\pi) = -\mathcal{H}(\pi)$ for $\mathcal{H}(\pi) = \sum_{a\in\mathcal{A}} \pi_a \log\Big(\frac{1}{\pi_a}\Big)$ satisfies the Definition A.12 for ℓ_1-norm;
- The negative Tsallis entropy $\Phi(\pi) = -\frac{1}{q}\mathcal{T}_q(\pi)$ for $\mathcal{T}_q(\pi) = \frac{1}{q-1}(1 - \sum_{a\in\mathcal{A}} \pi_a^q)$ satisfied the Definition A.12 for ℓ_2 norm for every $q \in (0, 1)$. In particular, $q = 0.5$ corresponds to the choice by Grill et al. (2019) in Appendix E that is tightly connected to the UCB algorithm;
- For any other fixed policy $\pi' \in \Delta(\mathcal{A})$ we can choose $\Phi(\pi) = \mathrm{KL}(\pi, \pi') = \sum_{a\in\mathcal{A}} \pi_a \log\Big(\frac{\pi_a}{\pi'_a}\Big)$ that inherits all the properties from the choice of the negative entropy.

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \{p_h\}_{h\in[H]}, \{r_h\}_{h\in[H]}, s_1)$ be a finite-horizon MDP, where $r_h(s, a)$ is a deterministic reward function. For simplicity we assume that $0 \leqslant r_h(s, a) \leqslant r_{\max}$ for any $(h, s, a) \in [H]\times\mathcal{S}\times\mathcal{A}$.

Then we can define entropy-augmented rewards as follows

$$r_{\kappa,h}(s, a) = r_h(s, a) + \kappa\mathcal{H}(p_h(s, a)).$$

This definition is required to cover the following case of practical interest. Let $\kappa = \lambda$ and $\Phi(\pi) = -\mathcal{H}(\pi)$, then we obtain the following representation for the λ-regularized value function

$$V^{\pi}_{\lambda,1}(s_1) = V^{\pi}_1(s_1) + \lambda\mathcal{H}_{\mathrm{traj}}(q^{\pi}),$$

where $V^{\pi}_1(s_1)$ is a usual value function for a MDP $\mathcal{M}$. For $r_{\max} = 0$ and $\kappa = \lambda = 1$ we recover just a trajectory entropy $V^{\pi}_{\lambda,1}(s_1) = \mathcal{H}_{\mathrm{traj}}(q^{\pi})$.

Next we define a convex conjugate to $\lambda\Phi$ as $F_\lambda\colon \mathbb{R}^{\mathcal{A}} \to \mathbb{R}$

$$F_\lambda(x) = \max_{\pi\in\Delta(\mathcal{A})} \{\langle\pi, x\rangle - \lambda\Phi(\pi)\}$$

and, with a sight abuse of notation extend the action of this function to the Q-function as follows

$$V^{\star}_{\lambda,h}(s) = F_\lambda(Q^{\star}_{\lambda,h})(s) = \max_{\pi\in\Delta_{\mathcal{A}}}\Big\{\pi Q^{\star}_{\lambda,h}(s) - \lambda\Phi(\pi)\Big\}.$$

Thanks to the fact that Φ satisfies Definition A.12, we have exact formula for the optimal policy by Fenchel-Legendre transform

$$\pi^{\star}_h(s) = \arg\max_{\pi\in\Delta_{\mathcal{A}}}\Big\{\pi Q^{\star}_{\lambda,h}(s) - \lambda\Phi(\pi)\Big\} = \nabla F_\lambda(Q^{\star}_{\lambda,h}(s, \cdot)).$$

Notice that we have $\nabla F_\lambda(Q^\star_{\lambda,h}(s,\cdot)) \in \Delta(\mathcal{A})$ since the gradient of Φ diverges on the boundary of $\Delta(\mathcal{A})$. For entropy regularization this formula become the softmax function.

Finally, it is known that the smoothness property of F_λ plays a key role in reduced sample complexity for planning in regularized MDPs (Grill et al. 2019). For our general setting we have that since $\lambda\Phi$ is λ-strongly convex with respect to $\|\cdot\|$, then F_λ is $1/\lambda$-strongly smooth with respect to a dual norm $\|\cdot\|_*$

$$F_\lambda(x) \leqslant F_\lambda(x') + \langle \nabla F_\lambda(x'), x - x' \rangle + \frac{1}{2\lambda}\|x - x'\|_*^2.$$

Let us define R_Φ as a maximal possible value of $|\Phi|$. Without loss of generality assume that $\Phi \leqslant 0$. In this case we define $\mathrm{R}_{\max} = r_{\max} + \kappa \log(S) + \lambda \mathrm{R}_\Phi$ as an upper bound of an about of reward obtain at the one step. By this definition we have $0 \leqslant V^\pi_{\lambda,h}(s) \leqslant H\mathrm{R}_{\max}$ for any $h \in [H], s \in \mathcal{S}$ and any policy π.

Also, since all norms in $\mathbb{R}^{\mathcal{A}}$ are equivalent, we define a constant r_A that defined for a dual norm $\|\cdot\|_*$ as follows

$$\|\cdot\|_* \leqslant r_A \cdot \|\cdot\|_\infty. \tag{A.6}$$

For example, for ℓ_2-norm $r_A = \sqrt{A}$ and for ℓ_1-norm $r_A = A$. In the case $\Phi = -\mathcal{H}$ we have $r_A = 1$ since the entropy is 1-strongly convex with respect to a ℓ_1-norm, thus the dual norm is exactly a ℓ_∞-norm.

The rest of this section is devoted to obtain the sample complexity guarantees for `UCBVI-Ent` algorithm with *a regularization-agnostic stopping rule* (A.12) with the gap notion (A.11). In this case Theorem A.20 gives us $\widetilde{\mathcal{O}}\Big(\frac{H^3SA}{\varepsilon^2} + \frac{H^3S^2A}{\varepsilon}\Big)$ sample complexity guarantee ignoring $\mathrm{R}_{\max}$ and poly-logarithmic factors. Notably, this sample complexity result does depend directly on $1/\lambda$, so small regularization does not affect the complexity.

In Section A.5 we present another algorithm `RL-Explore-Ent`, based on ideas of reward-free exploration, that achieves sample complexity of order $\widetilde{\mathcal{O}}(\mathrm{poly}(S,A,H)/(\lambda\varepsilon))$. As a particular example, it yields an algorithm for the MTEE problem with sample complexity $\widetilde{\mathcal{O}}(\mathrm{poly}(S,A,H)/\varepsilon)$ by taking as a regularizer negative entropy, $\lambda = \kappa = 1$ and $r_{\max} = 0$. Moreover, in Section A.6 we apply this algorithm to achieve sample complexity of order $\widetilde{\mathcal{O}}(H^2SA/\varepsilon^2)$ for the maximum visitation entropy exploration problem.

A.4.2 `UCBVI-Ent` Algorithm

In this section we describe `UCBVI-Ent` algorithm to solve MTEE problem, however as we show in the proofs, it is capable to work with general regularized MDPs.

We now describe our algorithm `UCBVI-Ent` for MTEE. Since one only needs to solve regularized Bellman equations to obtain a maximum trajectory entropy policy, we can use an algorithm of the same flavor as the ones proposed for best policy identification (Tirinzoni et al. 2023; Ménard et al. 2021; Kaufmann et al. 2021; Dann et al. 2019). In particular, `UCBVI-Ent` is close to the `BPI-UCRL` algorithm by Kaufmann et al. (2021) and can be characterized by the following rules.

Sampling rule. As sampling rule we use an optimistic policy π^{t+1} obtained by optimistic planning in the regularized MDP

$$\begin{aligned}
\overline{Q}^t_h(s,a) &= \mathrm{clip}\Big(\mathcal{H}\big(\widehat{p}_h(s,a)\big) + b^{\mathcal{H},t}_h(s,a) + \widehat{p}^t_h \overline{V}^t_{h+1}(s,a) + b^{p,t}_h(s,a), 0, \log(SA)H\Big)\,, \\
\overline{V}^t_h(s) &= \max_{\pi \in \Delta_A} \pi \overline{Q}^t_h(s) + \mathcal{H}(\pi)\,, \\
\pi^{t+1}_h(s) &= \arg\max_{\pi \in \Delta_A} \pi \overline{Q}^t_h(s) + \mathcal{H}(\pi)\,,
\end{aligned} \tag{A.7}$$

with $\overline{V}^t_{H+1} = 0$ by convention where $b^{\mathcal{H},t}$, $b^{p,t}$ are bonuses for the entropy and the transition probabilities, respectively. Precisely, we use bonuses of the form

$$
\begin{aligned}
b_h^{\mathcal{H},t}(s,a) &= \sqrt{\frac{2\beta^{\mathcal{H}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + \min\left(\frac{\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}, \log(S)\right), \\
b_h^{p,t}(s,a) &\triangleq b_h^{B,t}(s,a) + b_h^{\mathrm{corr},t}(s,a), \\
b_h^{B,t}(s,a) &\triangleq 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\overline{V}^t_{h+1})(s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + \frac{9H^2\log(SA)\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}, \\
b_h^{\mathrm{corr},t}(s,a) &\triangleq \frac{1}{H}\widehat{p}_h^t(\overline{V}^t_{h+1} - \underline{V}^t_{h+1})(s,a).
\end{aligned}
$$

for some functions $\beta^{\mathcal{H}}, \beta^{\mathrm{KL}}$ and β^{conc} and $\underline{V}^t$ is a lower confidence bound on the optimal value function defined in Appendix A.4.4.

Stopping and decision rule. To define the stopping rule, we first recursively build an upper-bound on the difference between the value of the optimal policy and the value of the current policy π^{t+1},

$$
\begin{aligned}
G_h^t(s,a) \triangleq \mathrm{clip}\Bigg(2b_h^{B,t}(s,a) + 2b_h^{\mathcal{H},t}(s,a) + \frac{4H^2\log(SA)\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \\
+ \left(1 + \frac{3}{H}\right)\widehat{p}_h^t\left[\pi_{h+1}^{t+1}G_{h+1}^t\right](s,a), 0, H\mathrm{R}_{\max}\Bigg),
\end{aligned}
\tag{A.8}
$$

where $\underline{V}^t$ is a lower-bound on the optimal value function defined in Appendix A.4.4 and $G^t_{H+1}(s,a) = 0$ by convention. Then the stopping time $\mathfrak{t} = \inf\{t \in \mathbb{N} : \pi^{t+1}G_1^t(s_1) \leqslant \varepsilon\}$ corresponds to the first episode when this upper-bound is smaller than ε. At this episode we return the Markovian policy $\widehat{\pi} = \pi^{\mathfrak{t}+1}$.

Algorithm 6 `UCBVI-Ent`

Input: Target precision ε, target probability δ, threshold functions $\beta^{\mathcal{H}}, \beta^{p}$, β^{conc}.
while true **do**
  Compute π^t by optimistic planning with (A.7).
  Compute bound on the gap $G_1^{t-1}(s,a)$ with (A.8).
  if $\pi^t G_1^{t-1}(s_1) \leqslant \varepsilon$ **then break**
  for $h \in [H]$ **do**
    Play $a_h^t \sim \pi_h^t(s_h^t)$
    Observe $s_{h+1}^t \sim p_h(s_h^t, a_h^t)$
  end for
  Update counts n^t, transition estimates $\widehat{p}^t$ and episode number $t \leftarrow t+1$.
end while
Output policy $\widehat{\pi} = \pi^t$.

The complete procedure is described in Algorithm 6. We now state our main theoretical result for `UCBVI-Ent`. We prove that for the well calibrated threshold functions $\beta^{\mathcal{H}}, \beta^{\mathrm{KL}}$ and β^{conc}, the `UCBVI-Ent` is (ε, δ)-PAC for MTEE and provide a high-probability upper bound on its sample complexity.

Theorem A.13. *Let $\beta^{\mathrm{KL}}, \beta^{\mathrm{conc}}$ and $\beta^{\mathcal{H}}$ be defined in Lemma A.14 of Appendix A.4. Fix $\varepsilon > 0$ and $\delta \in (0,1)$, then the* `UCBVI-Ent` *algorithm is (ε, δ)-PAC for MTEE. Moreover, the candidate to*

be an ε-optimal policy is given by $\widehat{\pi} = \pi^{\mathfrak{t}+1}$ where

$$\mathfrak{t} = \widetilde{\mathcal{O}}\left(\frac{H^3SA}{\varepsilon^2} + \frac{H^3S^2A}{\varepsilon}\right).$$

with probability at least $1-\delta$. Here $\widetilde{\mathcal{O}}$ hides poly-logarithmic factors in $\varepsilon, \delta, H, S, A$.

See Theorem A.21 in Appendix A.4 for a precise bound and a proof. Basically, this result is a simple corollary of the general result on regularized MDPs.

Space and time complexity. Since `UCBVI-Ent` is a model based algorithm, its space-complexity is of order $\mathcal{O}(HS^2A)$ whereas its time-complexity for one episode is of order $\mathcal{O}(HS^2A)$ because of the optimistic planning.

A.4.3 Concentration Events

Following the ideas of Ménard et al. (2021), we define the following concentration events.

Let $\beta^{\mathrm{KL}}, \beta^{\mathrm{conc}}, \beta^{\mathrm{cnt}}, \beta^{\mathcal{H}} : (0,1) \times \mathbb{N} \to \mathbb{R}_+$ be some functions defined later on in Lemma A.14. We define the following favorable events

$$\mathcal{E}^{\mathrm{KL}}(\delta) \triangleq \left\{ \forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \quad \mathrm{KL}(\widehat{p}_h^t(s,a), p_h(s,a)) \leqslant \frac{\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \right\},$$

$$\mathcal{E}^{\mathrm{conc}}(\delta) \triangleq \Bigg\{ \forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} :$$
$$|(\widehat{p}_h^t - p_h) V_{\lambda,h+1}^\star(s,a)| \leqslant \sqrt{2\mathrm{Var}_{p_h}(V_{\lambda,h+1}^\star)(s,a) \frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}}$$
$$+ 3H\mathrm{R}_{\max} \frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \Bigg\},$$

$$\mathcal{E}^{\mathrm{cnt}}(\delta) \triangleq \left\{ \forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \quad n_h^t(s,a) \geqslant \frac{1}{2}\overline{n}_h^t(s,a) - \beta^{\mathrm{cnt}}(\delta) \right\},$$

$$\mathcal{E}^{\mathcal{H}}(\delta) \triangleq \Bigg\{ \forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} :$$
$$|\mathcal{H}(\widehat{p}_h^t(s,a)) - \mathcal{H}(p_h(s,a))| \leqslant \sqrt{\frac{2\beta^{\mathcal{H}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + \left(\frac{\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \wedge \log(S) \right) \Bigg\}.$$

We also introduce two intersections of these events of interest, $\mathcal{G}(\delta) \triangleq \mathcal{E}^{\mathrm{KL}}(\delta) \cap \mathcal{E}_B^{\mathrm{conc}}(\delta) \cap \mathcal{E}^{\mathrm{cnt}}(\delta) \cap \mathcal{E}^{\mathcal{H}}(\delta)$. We prove that for the right choice of the functions $\beta^{\mathrm{KL}}, \beta^{\mathrm{conc}}, \beta^{\mathrm{cnt}}, \beta^{\mathcal{H}}$ the above events hold with high probability.

Lemma A.14. *For any $\delta \in (0,1)$ and for the following choices of functions β,*

$$\beta^{\mathrm{KL}}(\delta, n) \triangleq \log(4SAH/\delta) + S\log(\mathrm{e}(1+n)),$$
$$\beta^{\mathrm{conc}}(\delta, n) \triangleq \log(4SAH/\delta) + \log(4\mathrm{e}n(2n+1)),$$
$$\beta^{\mathrm{cnt}}(\delta) \triangleq \log(4SAH/\delta),$$
$$\beta^{\mathcal{H}}(\delta, n) \triangleq \log^2(n)(\log(4SAH/\delta) + \log(n(n+1))),$$

it holds that

$$\mathbb{P}[\mathcal{E}^{\mathrm{KL}}(\delta)] \geqslant 1-\delta/4, \qquad \mathbb{P}[\mathcal{E}^{\mathrm{conc}}(\delta)] \geqslant 1-\delta/4,$$
$$\mathbb{P}[\mathcal{E}^{\mathrm{cnt}}(\delta)] \geqslant 1-\delta/4, \qquad \mathbb{P}[\mathcal{E}^{\mathcal{H}}(\delta)] \geqslant 1-\delta/4.$$

In particular, $\mathbb{P}[\mathcal{G}(\delta)] \geqslant 1-\delta$.

Proof. Applying Theorem D.13 and the union bound over $h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$ we get $\mathbb{P}[\mathcal{E}^{\mathrm{KL}}(\delta)] \geqslant 1-\delta/4$. Next, Theorem D.15 and the union bound over $h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$ yield $\mathbb{P}[\mathcal{E}^{\mathrm{conc}}(\delta)] \geqslant 1-\delta/4$. By Theorem D.14 and union bound, $\mathbb{P}[\mathcal{E}^{\mathrm{cnt}}(\delta)] \geqslant 1-\delta/4$. Finally, by Theorem D.21 and union bound over $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ $\mathbb{P}[\mathcal{E}^{\mathcal{H}}(\delta)] \geqslant 1-\delta/4$. The union bound over four prescribed events concludes $\mathbb{P}[\mathcal{G}(\delta)] \geqslant 1-\delta$. □

Lemma A.15. *Assume conditions of Lemma A.14. Then on event* $\mathcal{E}^{\mathrm{KL}}(\delta)$*, for any* $f\colon \mathcal{S} \to [0, H\mathrm{R}_{\max}]$*,* $t \in \mathbb{N}, h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$*,*

$$[p_h - \widehat{p}_h^t]f(s,a) \leqslant \frac{1}{H}\widehat{p}_h^t f(s,a) + H\mathrm{R}_{\max}\left(\frac{5H\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \wedge 1\right),$$
$$[\widehat{p}_h^t - p_h]f(s,a) \leqslant \frac{1}{H}p_h f(s,a) + H\mathrm{R}_{\max}\left(\frac{5H\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \wedge 1\right).$$

Proof. Let us start from the first statement. We apply Lemma D.9 and Lemma D.10 to obtain

$$\begin{aligned}[p_h - \widehat{p}_h^t]f(s,a) &\leqslant \sqrt{2\mathrm{Var}_{p_h}[f](s,a)\cdot \mathrm{KL}(\widehat{p}_h^t, p_h)} + \frac{H}{3}\mathrm{KL}(\widehat{p}_h^t \| p_h) \\ &\leqslant 2\sqrt{\mathrm{Var}_{\widehat{p}_h^t}[f](s,a)\cdot \mathrm{KL}(\widehat{p}_h^t, p_h)} + 4H\mathrm{R}_{\max}\mathrm{KL}(\widehat{p}_h^t, p_h).\end{aligned}$$

Since $0 \leqslant f(s) \leqslant H\mathrm{R}_{\max}$ we get

$$\mathrm{Var}_{\widehat{p}_h^t}[f](s,a) \leqslant \widehat{p}_h^t[f^2](s,a) \leqslant H\mathrm{R}_{\max}\cdot\widehat{p}_h^t f(s,a).$$

Finally, applying $2\sqrt{ab} \leqslant a+b, a, b \geqslant 0$, we obtain the following inequality

$$(\widehat{p}_h^t - p_h)f(s,a) \leqslant \frac{1}{H}\widehat{p}_h^t f(s,a) + 5H^2\mathrm{R}_{\max}\mathrm{KL}(\widehat{p}_h^t, p_h).$$

Definition of $\mathcal{E}^{\mathrm{KL}}(\delta)$ implies the part of the statement. At the same time we have a trivial bound since $f(s) \in [0, H\mathrm{R}_{\max}]$

$$[p_h - \widehat{p}_h^t]f(s,a) \leqslant H\mathrm{R}_{\max} \leqslant \frac{1}{H}\widehat{p}_h^t f(s,a) + H\mathrm{R}_{\max}.$$

To prove the second statement, apply the first inequality of Lemma D.9 and proceed similarly. □

Lemma A.16. *Assume conditions of Lemma A.14 and assume that* $\beta^{\mathrm{conc}}(\delta) \leqslant \beta^{\mathrm{KL}}(\delta)$*. Then conditioned on event* $\mathcal{G}(\delta)$*, for any* $U\colon \mathcal{S} \to [0, H\mathrm{R}_{\max}]$*,* $t \in \mathbb{N}, h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$*,*

$$\begin{aligned}|(\widehat{p}_h^t - p_h)V_{\lambda,h+1}^\star(s,a)| \leqslant 3\sqrt{\mathrm{Var}_{\widehat{p}_{h+1}^t}(U)(s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + \frac{9H^2\mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \\ + \frac{1}{H}\widehat{p}_h^t|U - V_{\lambda,h+1}^\star|(s,a).\end{aligned}$$

Proof. First, we apply the definition of event $\mathcal{E}^{\mathrm{conc}}(\delta)$

$$|(\widehat{p}_h^t - p_h)V_{\lambda,h+1}^\star(s,a)| \leqslant \sqrt{2\mathrm{Var}_{p_h}(V_{\lambda,h+1}^\star)(s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + 3H\mathrm{R}_{\max}\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}.$$

Next we apply Lemma D.10 and Lemma D.11 and obtain

$$\begin{aligned}\mathrm{Var}_{p_h}(V_{\lambda,h+1}^\star)(s,a) &\leqslant 2\mathrm{Var}_{\widehat{p}_h^t}(V_{\lambda,h+1}^\star)(s,a) + 4H^2\mathrm{R}_{\max}^2\,\mathrm{KL}(\widehat{p}_h^t(s,a), p_h(s,a)) \\ &\leqslant 4\mathrm{Var}_{\widehat{p}_{h+1}^t}(U)(s,a) + 4H\mathrm{R}_{\max}\widehat{p}_h^t|U - V_{\lambda,h+1}^\star|(s,a) \\ &\quad + 4H^2\mathrm{R}_{\max}^2\,\mathrm{KL}(\widehat{p}_h^t(s,a), p_h(s,a)).\end{aligned}$$

Thus, by inequality $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$.

$$\begin{aligned}|(\widehat{p}_h^t - p_h)V_{\lambda,h+1}^\star(s,a)| \leqslant\ & 3\sqrt{\mathrm{Var}_{\widehat{p}_{h+1}^t}(U)(s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \\ &+ 3\sqrt{H\mathrm{R}_{\max}\widehat{p}_h^t|U - V_{\lambda,h+1}^\star|(s,a)\cdot\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \\ &+ 3H\mathrm{R}_{\max}\sqrt{\mathrm{KL}(\widehat{p}_h^t(s,a), p_h(s,a))\cdot\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \\ &+ 3H\mathrm{R}_{\max}\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}.\end{aligned}$$

By inequality $2\sqrt{ab} \leqslant a + b$ we have

$$\begin{aligned}3\sqrt{H\mathrm{R}_{\max}\widehat{p}_h^t|U - V_{\lambda,h+1}^\star|(s,a)\cdot\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \leqslant\ & \frac{1}{H}\widehat{p}_h^t|U - V_{\lambda,h+1}^\star|(s,a) \\ &+ \frac{9H^2\mathrm{R}_{\max}\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{4n_h^t(s,a)}.\end{aligned}$$

By the definition of the event $\mathcal{E}^{\mathrm{KL}}(\delta)$ and the fact $\beta^{\mathrm{conc}}(\delta) \leqslant \beta^{\mathrm{KL}}(\delta)$ we have

$$\sqrt{\mathrm{KL}(\widehat{p}_h^t(s,a), p_h(s,a))\cdot\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \leqslant \frac{\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}.$$

□

A.4.4 Confidence Intervals

Similar to Azar et al. (2017), Zanette and Brunskill (2019), and Ménard et al. (2021), we define the upper confidence bound for the optimal regularized Q-function of two types: with Hoeffding bonuses and with Bernstein bonuses.

Let us define empirical estimate of entropy-augmented rewards as follows

$$\widehat{r}_{\kappa,h}^t(s,a) = r_h(s,a) + \kappa\mathcal{H}(\widehat{p}_h^t(s,a)).$$

Then we have the following sequences defined as follows

$$\begin{aligned}\overline{Q}_h^t(s,a) &= \mathrm{clip}\Big(\widehat{r}_{\kappa,h}^t(s,a) + \widehat{p}_h^t\overline{V}_h^t(s,a) + b_h^{p,t}(s,a) + \kappa b_h^{\mathcal{H},t}(s,a), 0, H\mathrm{R}_{\max}\Big) \\ \pi_h^{t+1}(s) &= \max_{\pi\in\Delta_{\mathcal{A}}}\{\pi\overline{Q}_h^t(s) - \lambda\Phi(\pi)\}, \\ \overline{V}_h^t(s) &= \mathcal{H}(\pi_h^{t+1}(s)) + \pi_h^{t+1}\overline{Q}_h^t(s) \\ \overline{V}_{H+1}^t(s) &= 0,\end{aligned}$$

and the lower confidence bound as follows

$$\begin{aligned}\underline{Q}_h^t(s,a) &= \operatorname{clip}\Big(\widehat{r}_{\kappa,h}^t(s,a) + \widehat{p}_h^t \underline{V}_h^t(s,a) - b_h^{p,t}(s,a) - \kappa b_h^{\mathcal{H},t}(s,a), 0, H\mathrm{R}_{\max}\Big)\\ \underline{V}_h^t(s) &= \max_{\pi\in\Delta_{\mathcal{A}}}\{\pi\underline{Q}_h^t(s) - \lambda\Phi(\pi)\}\\ \underline{V}_{H+1}^t(s) &= 0,\end{aligned}$$

where the Bernstein bonuses for transitions are defined as follows

$$\begin{aligned}b_h^{p,t}(s,a) &\triangleq b_h^{B,t}(s,a) + b_h^{\mathrm{corr},t}(s,a),\\ b_h^{B,t}(s,a) &\triangleq 3\sqrt{\operatorname{Var}_{\widehat{p}_h^t}(\overline{V}_{h+1}^t)(s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + \frac{9H^2\mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)},\\ b_h^{\mathrm{corr},t}(s,a) &\triangleq \frac{1}{H}\widehat{p}_h^t(\overline{V}_{h+1}^t - \underline{V}_{h+1}^t)(s,a).\end{aligned} \tag{A.9}$$

The entropy bonuses are defined bellow

$$b_h^{\mathcal{H},t}(s,a) \triangleq \sqrt{\frac{2\beta^{\mathcal{H}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + \left(\frac{\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \wedge \log(S)\right). \tag{A.10}$$

Theorem A.17. *Let $\delta \in (0,1)$. Assume Bernstein bonuses (A.9). Then on event $\mathcal{G}(\delta)$ for any $t \in \mathbb{N}$, $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$ it holds*

$$\underline{Q}_h^t(s,a) \leqslant Q_{\lambda,h}^\star(s,a) \leqslant \overline{Q}_h^t(s,a), \qquad \underline{V}_h^t(s) \leqslant V_{\lambda,h}^\star(s) \leqslant \overline{V}_h^t(s,a).$$

Proof. Proceed by induction over h. For $h = H+1$ the statement is trivial. Now we assume that inequality holds for any $h' > h$ for a fixed $h \in [H]$. Fix a timestamp $t \in \mathbb{N}$ and a state-action pair (s,a) and assume that $\overline{Q}_h^t(s,a) < H\mathrm{R}_{\max}$, i.e. no clipping occurs. Otherwise the inequality $Q_{\lambda,h}^\star(s,a) \leqslant \overline{Q}_h^t(s,a)$ is trivial. In particular, it implies $n_h^t(s,a) > 0$.

In this case, by Bellman equations (A.5) we have

$$\begin{aligned}[\overline{Q}_h^t - Q_{\lambda,h}^\star](s,a) = &\underbrace{r_h(s,a) + \kappa\mathcal{H}(\widehat{p}_h^t(s,a)) - r_h(s,a) - \kappa\mathcal{H}(p_h(s,a)) + \kappa b_h^{\mathcal{H},t}(s,a)}_{T_1}\\ &+ \underbrace{\widehat{p}_h^t\overline{V}_{h+1}^t(s,a) - p_hV_{\lambda,h+1}^\star(s,a) + b_h^{p,t}(s,a)}_{T_2}.\end{aligned}$$

By the definition of event $\mathcal{E}^{\mathcal{H}}(\delta)$ that is subset of $\mathcal{G}(\delta)$ we have $T_1 \geqslant 0$. To show that $T_2 \geqslant 0$, we start from induction hypothesis

$$T_2 \geqslant [\widehat{p}_h^t - p_h]V_{\lambda,h+1}^\star(s,a) + b_h^{p,t}(s,a).$$

Next we apply Lemma A.16 with $U = \overline{V}_{h+1}^t$ and definition of transition bonuses we have

$$T_2 \geqslant -\frac{1}{H}\widehat{p}_h^t|\overline{V}_{h+1}^t - V_{\lambda,h+1}^\star|(s,a) + \frac{1}{H}\widehat{p}_h^t[\overline{V}_{h+1}^t - \underline{V}_{h+1}^t](s,a).$$

By induction hypothesis we have $\overline{V}_{h+1}^t(s) \geqslant V_{\lambda,h+1}^\star(s) \geqslant \underline{V}_{h+1}^t(s)$, thus $T_2 \geqslant 0$.

To prove the second inequality on Q-function, we assume $\underline{Q}_h^t(s,a) > 0$ and, as a consequence, $n_h^t(s,a) > 0$. Thus we have

$$\begin{aligned}[\underline{Q}_h^t - Q_{\lambda,h}^\star](s,a) = &\underbrace{\mathcal{H}(\widehat{p}_h^t(s,a)) - \mathcal{H}(p_h(s,a)) - b_h^{\mathcal{H},t}(s,a)}_{T_1'}\\ &+ \underbrace{\widehat{p}_h^t\underline{V}_{h+1}^t(s,a) - p_hV_{\lambda,h+1}^\star(s,a) - b_h^{p,t}(s,a)}_{T_2'}.\end{aligned}$$

Again, by the definition of event $\mathcal{E}^{\mathcal{H}}(\delta)$ we have $T_1' \leqslant 0$ and, by induction hypothesis

$$T_2' \leqslant [\widehat{p}_h^t - p_h]V_{h+1}^{\mathcal{H},\star}(s,a) - b_h^{p,t}(s,a).$$

We again apply Lemma A.16 with $U = \overline{V}_{h+1}^t$

$$T_2' \leqslant \frac{1}{H}\widehat{p}_h^t|\overline{V}_{h+1}^t - V_{\lambda,h+1}^\star|(s,a) - \frac{1}{H}\widehat{p}_h^t[\overline{V}_{h+1}^t - \underline{V}_{h+1}^t](s,a).$$

We conclude the statement by induction hypothesis for $h' = h+1$.

Finally, we have to show the inequality for V-functions. To do it, we use the fact that V-functions are computed by F_λ applied to Q-functions

$$\underline{V}_h^t(s) = F_\lambda(\underline{Q}_h^t)(s), \quad V_{\lambda,h}^\star(s) = F_\lambda(Q_{\lambda,h}^\star)(s), \quad \overline{V}_h^t(s) = F_\lambda(\overline{Q}_h^t)(s).$$

Notice that ∇F_λ takes values in a probability simplex, thus, all partial derivatives of F_λ are non-negative and therefore F_λ is monotone in each coordinate. Thus, since $\underline{Q}_h^t(s,a) \leqslant Q_{\lambda,h}^\star(s,a) \leqslant \overline{Q}_h^t(s,a)$, we have the same inequality $\underline{V}_h^t(s) \leqslant V_{\lambda,h}^\star(s) \leqslant \overline{V}_h^t(s)$. □

A.4.5 Regularization-Agnostic Stopping Rule

In this section we provide guarantees for the so-called *regularization-agnostic gap*: this notion of gap does not influenced by regularization except the changing of the range of value functions and basically mimics `UCBVI-BPI` algorithm by Ménard et al. (2021) in definition of the similar gap.

Let us define $G_{H+1}^t(s,a) \triangleq 0$ for all s, a and

$$\begin{aligned} G_h^t(s,a) \triangleq \operatorname{clip}\Bigg(& 2b_h^{B,t}(s,a) + \frac{5H^2\mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} + 2\kappa b_h^{\mathcal{H},t}(s,a) \\ & + \left(1 + \frac{3}{H}\right)\widehat{p}_h^t\left[\pi_{h+1}^{t+1}G_{h+1}^t\right](s,a), 0, H\mathrm{R}_{\max}\Bigg), \end{aligned} \tag{A.11}$$

where $b_h^{B,t}(s,a)$ is defined in (A.9). For this notion of the gap we can define the stopping rule as follows

$$\mathfrak{t} = \min\{t \in \mathbb{N} : \pi_1^{t+1}G_1^t(s_1) \leqslant \varepsilon\}. \tag{A.12}$$

The lemma below justifies this choice of the stopping time.

Lemma A.18. *Assume the choice of Bernstein bonuses (A.9) and let the event $\mathcal{G}(\delta)$ defined in Lemma A.14 holds. Then for all $t \in \mathbb{N}$ we have*

$$V_{\lambda,1}^\star(s_1) - V_{\lambda,1}^{\pi^{t+1}}(s_1) \leqslant \pi_1^{t+1}G_1^t(s_1).$$

Proof. Following Ménard et al. (2021), we start by defining the following quantities

$$\begin{aligned} \widetilde{Q}_h^t(s,a) &\triangleq \operatorname{clip}\Big(\widehat{r}_{\kappa,h}^t(s,a) + \widehat{p}_h^t\widetilde{V}_{h+1}^t(s,a) - b_h^{p,t}(s,a) - \kappa b_h^{\mathcal{H},t}(s,a), 0, r_{\kappa,h}(s,a) + p_h\widetilde{V}_{h+1}^t(s,a)\Big), \\ \widetilde{V}_h^t(s) &\triangleq \pi_h^{t+1}\widetilde{Q}_h^t(s) - \lambda\Phi(\pi_h^{t+1}(s)), \\ \widetilde{V}_{H+1}^t(s) &\triangleq 0. \end{aligned}$$

By Theorem A.17 and Lemma A.19 we have

$$\begin{aligned} V_{\lambda,1}^\star(s_1) - V_{\lambda,1}^{\pi^{t+1}}(s_1) &\leqslant \overline{V}_1^t(s_1) - V_{\lambda,1}^{\pi^{t+1}}(s_1) \leqslant \overline{V}_1^t(s_1) - \widetilde{V}_1^t(s_1) \\ &= \pi_1^{t+1}\overline{Q}_1^t(s_1) - \lambda\Phi(\pi_1^{t+1}(s_1)) - \pi_1^{t+1}\widetilde{Q}_1^t(s_1) + \lambda\Phi(\pi_1^{t+1}(s_1)) \\ &= \pi_1^{t+1}[\overline{Q}_1^t - \widetilde{Q}_1^t](s_1)\,. \end{aligned}$$

Therefore, it is enough to show that for any $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$

$$[\overline{Q}_h^t - \widetilde{Q}_h^t](s,a) \leqslant G_h^t(s,a), \qquad [\overline{V}_h^t - \widetilde{V}_h^t](s) \leqslant \pi_h^{t+1} G_h^t(s).$$

Proceed by backward induction over h. The case $h = H+1$ is trivial, thus we may assume that the statement holds for any $h' > h$ for a fixed h. Also fix $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Notice that if $G_h^t(s,a) = H\mathrm{R}_{\max}$, then the inequality is trivially true. Therefore we may assume that $G_h^t(s,a) < H\mathrm{R}_{\max}$ and, consequently, $n_h^t(s,a) > 0$. Now we have to separate cases.

First case. In this case we have $\widetilde{Q}_h^t(s,a) = r_{\kappa,h}(s,a) + p_h \widetilde{V}_{h+1}^t(s,a)$, i.e. maximal clipping occurs. Therefore

$$\begin{aligned}\overline{Q}_h^t(s,a) - \widetilde{Q}_h^t(s,a) = r_h(s,a) + \kappa\mathcal{H}(\widehat{p}_h^t(s,a)) + \kappa b_h^{\mathcal{H},t}(s,a) - r_h(s,a) - \kappa\mathcal{H}(p_h(s,a)) \\ + \widehat{p}_h^t \overline{V}_{h+1}^t(s,a) - p_h \widetilde{V}_{h+1}^t(s,a) + b_h^{p,t}(s,a).\end{aligned}$$

By the definition of the event $\mathcal{E}^{\mathcal{H}}(\delta) \subseteq \mathcal{G}(\delta)$ we have

$$\kappa\mathcal{H}(\widehat{p}_h^t(s,a)) + \kappa b_h^{\mathcal{H},t}(s,a) - \kappa\mathcal{H}(p_h(s,a)) \leqslant 2\kappa b_h^{\mathcal{H},t}(s,a),$$

for the next term we have

$$\widehat{p}_h^t \overline{V}_{h+1}^t - p_h \widetilde{V}_{h+1}^t(s,a) = \widehat{p}_h^t[\overline{V}_{h+1}^t - \widetilde{V}_{h+1}^t(s,a)] + [\widehat{p}_h^t - p_h]V_{\lambda,h+1}^\star(s,a) + [p_h - \widehat{p}_h^t][V_{\lambda,h+1}^\star - \widetilde{V}_{h+1}^t](s,a).$$

By induction hypothesis

$$\widehat{p}_h^t[\overline{V}_{h+1}^t - \widetilde{V}_{h+1}^t(s,a)] \leqslant \widehat{p}_h^t[\pi_{h+1}^{t+1} G_{h+1}^t](s,a).$$

Next we apply Lemma A.16 with $U = \overline{V}_{h+1}^t(s,a)$ and Theorem A.17

$$[\widehat{p}_h^t - p_h]V_{\lambda,h+1}^\star(s,a) \leqslant b_h^{p,t}(s,a).$$

Finally, we apply Lemma A.15 and obtain

$$[p_h - \widehat{p}_h^t][V_{\lambda,h+1}^\star - \widetilde{V}_{h+1}^t](s,a) \leqslant \frac{1}{H}\widehat{p}_h^t[V_{\lambda,h+1}^\star - \widetilde{V}_{h+1}^t](s,a) + \frac{5H^2\mathrm{R}_{\max} \cdot \beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}.$$

Summing all these bounds up, we have

$$\begin{aligned}\overline{Q}_h^t(s,a) - \widetilde{Q}_h^t(s,a) \leqslant \widehat{p}_h^t[\pi_{h+1}^{t+1} G_{h+1}^t](s,a) + 2\kappa b_h^{\mathcal{H},t}(s,a) + 2b_h^{p,t}(s,a) \\ + \frac{1}{H}\widehat{p}_h^t[V_{\lambda,h+1}^\star - \widetilde{V}_{h+1}^t](s,a) + \frac{5H^2\mathrm{R}_{\max} \cdot \beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}.\end{aligned}$$

Notice that by Theorem A.17 and the induction hypothesis

$$\frac{1}{H}\widehat{p}_h^t[V_{\lambda,h+1}^\star - \widetilde{V}_{h+1}^t](s,a) \leqslant \frac{1}{H}\widehat{p}_h^t[\overline{V}_{h+1}^t - \widetilde{V}_{h+1}^t](s,a) \leqslant \frac{1}{H}\widehat{p}_h^t[\pi_{h+1}^{t+1} G_{h+1}^t](s,a),$$

and by decomposing the transition bonus (A.9) to Bernstein bonus and correction term and applying Lemma A.19

$$\begin{aligned}b_h^{p,t}(s,a) = b_h^{B,t}(s,a) + \frac{1}{H}\widehat{p}_h^t[\overline{V}_{h+1}^t - \underline{V}_{h+1}^t](s,a) \leqslant b_h^{B,t}(s,a) + \frac{1}{H}\widehat{p}_h^t[\overline{V}_{h+1}^t - \widetilde{V}_{h+1}^t](s,a) \\ \leqslant b_h^{B,t}(s,a) + \frac{1}{H}\widehat{p}_h^t[\pi_{h+1}^{t+1} G_{h+1}^t](s,a),\end{aligned}$$

thus

$$\begin{aligned}\overline{Q}_h^t(s,a) - \widetilde{Q}_h^t(s,a) \leqslant \left(1 + \frac{3}{H}\right)\widehat{p}_h^t[\pi_{h+1}^{t+1} G_{h+1}^t](s,a) + 2\kappa b_h^{\mathcal{H},t}(s,a) + 2b_h^{B,t}(s,a) \\ + \frac{5H^2\mathrm{R}_{\max} \cdot \beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} = G_h^t(s,a).\end{aligned}$$

Second case. In this case we have $\widetilde{Q}_h^t(s,a) = \widehat{r}_{\lambda,h}^t(s,a) + \widehat{p}_h^t \widetilde{V}_{h+1}^t(s,a) - b_h^{p,t}(s,a) - \kappa b_h^{\mathcal{H},t}(s,a)$. Thus

$$\overline{Q}_h^t(s,a) - \widetilde{Q}_h^t(s,a) \leqslant 2\kappa b_h^{\mathcal{H},t}(s,a) + 2b_h^{B,t}(s,a) + \widehat{p}_h^t[\overline{V}_{h+1}^t - \widetilde{V}_{h+1}^t](s,a) + \frac{2}{H}\widehat{p}_h^t[\overline{V}_{h+1}^t - \underline{V}_{h+1}^t](s,a).$$

By Lemma A.19 and induction hypothesis we have

$$\overline{Q}_h^t(s,a) - \widetilde{Q}_h^t(s,a) \leqslant 2\kappa b_h^{\mathcal{H},t}(s,a) + 2b_h^{B,t}(s,a) + \left(1 + \frac{2}{H}\right)\widehat{p}_h^t[\pi_{h+1}^{t+1} G_{h+1}^t](s,a) \leqslant G_h^t(s,a).$$

Conclusion. From the two cases above we conclude

$$[\overline{Q}_h^t - \widetilde{Q}_h^t](s,a) \leqslant G_h^t(s,a).$$

Moreover, we have

$$\overline{V}_h^t(s) - \widetilde{V}_h^t(s) = \pi_h^{t+1}\overline{Q}_h^t(s) - \lambda\Phi(\pi_h^{t+1}(s))\pi_h^{t+1}\widetilde{Q}_h^t(s) + \lambda\Phi(\pi_h^{t+1}(s)) = \pi_h^{t+1}[\overline{Q}_h^t - \widetilde{Q}_h^t](s) \leqslant \pi_h^{t+1} G_h^t(s).$$

The last inequality concludes the statement of Lemma A.18. □

Lemma A.19. *Under the choice of Bernstein bonuses* (A.9)*, on event* $\mathcal{G}(\delta)$ *for any* $t \in \mathbb{N}$ *and any* $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$

$$\widetilde{Q}_h^t(s,a) \leqslant \min\{Q_{\lambda,h}^{\pi^{t+1}}(s,a), \underline{Q}_h^t(s,a)\}, \qquad \widetilde{V}_h^t(s) \leqslant \min\{V_{\lambda,h}^{\pi^{t+1}}(s), \underline{V}_h^t(s)\}.$$

Proof. Proceed by backward induction over h. The case $h = H+1$ is trivially true, assume that the statement holds for any $h' > h$ for a fixed h. Also let us fix t, s, a. By induction hypothesis we have

$$\widetilde{Q}_h^t(s,a) \leqslant r_{\kappa,h}(s,a) + p_h \widetilde{V}_{h+1}^t \leqslant r_{\kappa,h}(s,a) + p_h V_{\lambda,h+1}^{\pi^{t+1}}(s,a) = Q_{\lambda,h}^{\pi^{t+1}}(s,a).$$

In the same manner

$$\begin{aligned}\widetilde{Q}_h^t(s,a) &\leqslant \widehat{r}_{\kappa,h}^t(s,a) + \widehat{p}_h^t \widetilde{V}_{h+1}^t - b_h^{p,t}(s,a) - \kappa b_h^{\mathcal{H},t}(s,a) \\ &\leqslant \widehat{r}_{\kappa,h}^t(s,a) + \widehat{p}_h^t \underline{V}_{h+1}^t - b_h^{p,t}(s,a) - \kappa b_h^{\mathcal{H},t}(s,a) = \underline{Q}_h^t(s,a).\end{aligned}$$

Next, we prove the same inequalities for V-functions

$$\widetilde{V}_h^t(s) = \pi_h^{t+1}(s)\widetilde{Q}_h^t(s,a) - \lambda\Phi(\pi_h^{t+1}(s)) \leqslant \pi_h^{t+1}(s) Q_{\lambda,h}^{\pi^{t+1}}(s,a) - \lambda\Phi(\pi_h^{t+1}(s)) = V_{\lambda,h}^{\pi^{t+1}}(s),$$

and

$$\begin{aligned}\widetilde{V}_h^t(s) &= \pi_h^{t+1}\widetilde{Q}_h^t(s) - \lambda\Phi(\pi_h^{t+1}(s)) \leqslant \pi_h^{t+1}\underline{Q}_h^t(s) - \lambda\Phi(\pi_h^{t+1}(s)) \\ &\leqslant \max_{\pi \in \Delta(\mathcal{A})}\left\{\pi \underline{Q}_h^t(s) - \lambda\Phi(\pi)\right\} = \underline{V}_h^t(s).\end{aligned}$$

□

After defining a proper quantity for a stopping rule we may proceed with the final proof for sample-complexity of the presented algorithm `UCBVI-Ent`.

Theorem A.20. *Let* $\delta \in (0,1)$*. Then* `UCBVI-Ent` *algorithm with Bernstein bonuses* (A.9) *and a regularization-agnostic stopping rule* $\mathfrak{t}$ *is* (ε,δ)*-PAC for the best policy identification in regularized MDPs.*

Moreover, with probability at least $1-\delta$ *the stopping time* $\mathfrak{t}$ *is bounded as follows*

$$\mathfrak{t} = \mathcal{O}\left(\frac{H^3 SA \mathrm{R}_{\max}^2 \log(SAH/\delta) L^4}{\varepsilon^2} + \frac{H^3 SA(\log(SAH/\delta) + SL)\cdot L}{\varepsilon}\right),$$

where $L = \mathcal{O}((\log(SAH\mathrm{R}_{\max}/\varepsilon)) + \log\log(SAH/\delta))$.

Proof. Notice that if $\mathfrak{t} = 0$, then our sample complexity bound is trivial, thus we assume that $\mathfrak{t} > 0$. Let us start from deriving an upper bound for $G_h^t(s,a)$ for $t < \mathfrak{t}, h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$.

$$\begin{aligned}
G_h^t(s,a) &\leqslant 2b_h^{B,t}(s,a) + 2\kappa b_h^{\mathcal{H},t}(s,a) + \frac{5H^2 \mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} + \left(1 + \frac{3}{H}\right)\widehat{p}_h^t[\pi_{h+1}^{t+1} G_{h+1}^t](s,a) \\
&\leqslant 6\sqrt{\mathrm{Var}_{\widehat{p}_h^t}[\overline{V}_{h+1}^t](s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + 2\kappa\sqrt{\frac{2\beta^{\mathcal{H}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \\
&\quad + \frac{23H^2 \mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} + \left(1 + \frac{3}{H}\right)[\widehat{p}_h^t - p_h][\pi_{h+1}^{t+1} G_{h+1}^t](s,a) \\
&\quad + \left(1 + \frac{3}{H}\right)p_h[\pi_{h+1}^{t+1} G_{h+1}^t](s,a)\,.
\end{aligned}$$

By Lemma A.15

$$[\widehat{p}_h^t - p_h][\pi_{h+1}^{t+1} G_{h+1}^t](s,a) \leqslant \frac{1}{H}p_h[\pi_{h+1}^{t+1} G_{h+1}^t](s,a) + \frac{5H^2 \mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}.$$

Also we have to replace the variance of the empirical model with the real variance of the value function for π^{t+1} in order to apply the law of total variance (Lemma A.10).

Apply Lemma D.11 and Lemma D.10

$$\begin{aligned}
\mathrm{Var}_{\widehat{p}_h^t}[\overline{V}_{h+1}^t](s,a) &\leqslant 2\mathrm{Var}_{p_h}[\overline{V}_{h+1}^t](s,a) + \frac{4H^2 \mathrm{R}_{\max}^2\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \\
&\leqslant 4\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s,a) + 2H\mathrm{R}_{\max}p_h[\overline{V}_{h+1}^t - V_{\lambda,h+1}^{\pi^{t+1}}](s,a) \\
&\quad + \frac{4H^2 \mathrm{R}_{\max}^2\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}.
\end{aligned}$$

In the proof of Lemma A.18 it was proven that

$$[\overline{V}_{h+1}^t - V_{\lambda,h+1}^{\pi^{t+1}}](s) \leqslant \pi_{h+1}^{t+1} G_{h+1}^t(s),$$

thus, combining with an inequality $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$

$$\begin{aligned}
6\sqrt{\mathrm{Var}_{\widehat{p}_h^t}[\overline{V}_{h+1}^t](s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} &\leqslant 12\sqrt{\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \\
&\quad + 6\sqrt{p_h[\pi_{h+1}^{t+1} G_{h+1}^t](s,a)\frac{2H\mathrm{R}_{\max}\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \\
&\quad + \frac{12H\mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s,a)}{n_h^t(s,a)}
\end{aligned}$$

To bound the second term, we use inequality $2\sqrt{ab} \leqslant a + b$

$$\begin{aligned}
6\sqrt{p_h[\pi_{h+1}^{t+1} G_{h+1}^t](s,a)\frac{2H\mathrm{R}_{\max}\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} &\leqslant \frac{3}{H}p_h[\pi_{h+1}^{t+1} G_{h+1}^t](s,a) \\
&\quad + \frac{3H^2 \mathrm{R}_{\max}\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}.
\end{aligned}$$

Finally, we have the following bound on $G_h^t(s,a)$

$$\begin{aligned}
G_h^t(s,a) &\leqslant 12\sqrt{\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + 2\kappa\sqrt{\frac{2\beta^{\mathcal{H}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \\
&\quad + \frac{54H^2 \mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} + \left(1 + \frac{10}{H}\right)p_h[\pi_{h+1}^{t+1} G_{h+1}^t](s,a).
\end{aligned}$$

Notice that his inequality could be rewritten in the following form

$$G_h^t(s,a) \leqslant \mathbb{E}_{\pi^{t+1}}\Bigg[12\sqrt{\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + 2\kappa\sqrt{\frac{2\beta^{\mathcal{H}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \\ + \frac{54H^2\mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} + \Big(1+\frac{10}{H}\Big)G_{h+1}^t(s_{h+1},a_{h+1})\Big|(s_h,a_h)=(s,a)\Bigg],$$

thus by rolling out we have

$$\pi_1^{t+1}G_1^t(s_1) \leqslant \mathbb{E}_{\pi^{t+1}}\Bigg[\underbrace{12\sum_{h=1}^H\Big(1+\frac{10}{H}\Big)^h\sqrt{\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s_h,a_h)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s_h,a_h))}{n_h^t(s_h,a_h)}}}_{(\mathbf{A})} \\ \underbrace{+2\kappa\sum_{h=1}^H\Big(1+\frac{10}{H}\Big)^h\sqrt{\frac{2\beta^{\mathcal{H}}(\delta, n_h^t(s_h,a_h))}{n_h^t(s_h,a_h)}}}_{(\mathbf{B})} \\ \underbrace{+\sum_{h=1}^H\Big(1+\frac{10}{H}\Big)^h\frac{54H^2\mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s_h,a_h))}{n_h^t(s_h,a_h)}}_{(\mathbf{C})}\Big|s_1\Bigg],$$

where $(1+10/H)^h \leqslant \mathrm{e}^{10}$ for any $h \in [H]$. Now we bound each term separately.

Term (A). To bound this term, we apply Cauchy-Schwarz inequality

$$(\mathbf{A}) \leqslant 12\mathrm{e}^{10}\sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} d_h^{\pi^{t+1}}(s,a)\sqrt{\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \\ \leqslant 12\mathrm{e}^{10}\sqrt{\sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} d_h^{\pi^{t+1}}(s,a)\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s,a)}\times \\ \times\sqrt{\sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}}.$$

For the first multiplier we apply the law of total variance (Lemma A.10)

$$\sum_{h,s,a} d_h^{\pi^{t+1}}(s,a)\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s,a) \leqslant \sum_{h,s,a} d_h^{\pi^{t+1}}(s,a)\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s,a) + \sum_{h,s} d_h^{\pi^{t+1}}(s)\mathrm{Var}_{\pi_h^{t+1}}[Q_{\lambda,h}^{\pi^{t+1}}](s) \\ = \sigma V_1^{\mathcal{H},\pi^{t+1}}(s_1) \leqslant H^2\mathrm{R}_{\max}^2.$$

Therefore,

$$(\mathbf{A}) \leqslant 24\mathrm{e}^{10}H\mathrm{R}_{\max}\sqrt{\sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}}.$$

Term (B). For this term we may apply Jensen's inequality

$$(\mathbf{B}) \leqslant 2\kappa H\mathrm{e}^{10}\mathbb{E}_{\pi^{t+1}}\Bigg[\frac{1}{H}\sum_{h=1}^H\sqrt{\frac{2\beta^{\mathcal{H}}(\delta, n_h^t(s_h,a_h))}{n_h^t(s_h,a_h)}}\Big|s_1\Bigg] \leqslant \kappa\sqrt{8H}\mathrm{e}^{10}\sqrt{\sum_{h,s,a} d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathcal{H}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}}.$$

By summing up and replacing counts by pseudo-counts by Lemma D.5 we obtain

$$\begin{aligned}\pi_1^{t+1}G_1^t(s_1) \leqslant{}& 48\mathrm{e}^{10}H\mathrm{R}_{\max}\sqrt{\sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathrm{conc}}(\delta,\overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a)\vee 1}}\\ &+ 4\kappa\mathrm{e}^{10}\sqrt{2H}\sqrt{\sum_{h,s,a} d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathcal{H}}(\delta,\overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a)\vee 1}}\\ &+ 4\mathrm{e}^{10}H^2\mathrm{R}_{\max}\sum_{h,s,a} d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathrm{KL}}(\delta,\overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a)\vee 1}.\end{aligned}$$

The last step is to notice that for $t<\mathfrak{t}$ we have $\pi_1^{t+1}G_1^t(s_1)\geqslant\varepsilon$, thus summing over all $t<\mathfrak{t}$ we have

$$\begin{aligned}(\mathfrak{t}-1)\varepsilon \leqslant{}& 48\mathrm{e}^{10}H\mathrm{R}_{\max}\sum_{t=1}^{\mathfrak{t}-1}\sqrt{\sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathrm{conc}}(\delta,\overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a)\vee 1}}\\ &+ 4\kappa\mathrm{e}^{10}\sqrt{2H}\sum_{t=1}^{\mathfrak{t}-1}\sqrt{\sum_{h,s,a} d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathcal{H}}(\delta,\overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a)\vee 1}}\\ &+ 4\mathrm{e}^{10}H^2\mathrm{R}_{\max}\sum_{t=1}^{\mathfrak{t}-1}\sum_{h,s,a} d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathrm{KL}}(\delta,\overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a)\vee 1}.\end{aligned}$$

Also we notice that $\beta^{\mathrm{KL}}(\delta,\cdot),\beta^{\mathrm{conc}}(\delta,\cdot),\beta^{\mathcal{H}}(\delta,\cdot)$ are monotone, thus we may replace $\overline{n}_h^t(s,a)$ with a stopping time $\mathfrak{t}$. Thus, by Jensen's inequality

$$\begin{aligned}(\mathfrak{t}-1)\varepsilon \leqslant{}& 48\mathrm{e}^{10}H\mathrm{R}_{\max}\sqrt{(\mathfrak{t}-1)\cdot\beta^{\mathrm{conc}}(\delta,\mathfrak{t}-1)}\sqrt{\sum_{t=1}^{\mathfrak{t}-1}\sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} d_h^{\pi^{t+1}}(s,a)\frac{1}{\overline{n}_h^t(s,a)\vee 1}}\\ &+ 4\kappa\mathrm{e}^{10}\sqrt{2H\beta^{\mathcal{H}}(\delta,\mathfrak{t})\cdot(\mathfrak{t}-1)}\sqrt{\sum_{t=1}^{\mathfrak{t}-1}\sum_{h,s,a} d_h^{\pi^{t+1}}(s,a)\frac{1}{\overline{n}_h^t(s,a)\vee 1}}\\ &+ 4\mathrm{e}^{10}H^2\mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta,\mathfrak{t}-1)\sum_{t=1}^{\mathfrak{t}-1}\sum_{h,s,a} d_h^{\pi^{t+1}}(s,a)\frac{1}{\overline{n}_h^t(s,a)\vee 1}.\end{aligned}$$

Furthermore, notice

$$\sum_{t=1}^{\mathfrak{t}-1} d_h^{\pi^{t+1}}(s,a)\frac{1}{\overline{n}_h^t(s,a)\vee 1}=\sum_{t=1}^{\mathfrak{t}-1}\frac{\overline{n}_h^{t+1}(s,a)-\overline{n}_h^t(s,a)}{\overline{n}_h^t(s,a)\vee 1},$$

thus Lemma D.6 is applicable:

$$\begin{aligned}(\mathfrak{t}-1)\varepsilon \leqslant{}& 96\mathrm{e}^{10}\mathrm{R}_{\max}\sqrt{(\mathfrak{t}-1)H^3SA\log(\mathfrak{t})\cdot\beta^{\mathrm{conc}}(\delta,\mathfrak{t}-1)}\\ &+ 2\kappa\mathrm{e}^{10}\sqrt{2H^2SA\log^3(\mathfrak{t})\beta^{\mathcal{H}}(\delta,\mathfrak{t})\cdot(\mathfrak{t}-1)}\\ &+ 8\mathrm{e}^{10}H^3SA\log(\mathfrak{t})\mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta,\mathfrak{t}-1).\end{aligned}$$

By the definitions of $\beta^{\mathrm{KL}},\beta^{\mathrm{conc}},\beta^{\mathcal{H}}$ we have the following inequality

$$\begin{aligned}(\mathfrak{t}-1)\varepsilon \leqslant{}& 96\mathrm{e}^{10}\mathrm{R}_{\max}\sqrt{(\mathfrak{t}-1)H^3SA\log(\mathfrak{t})\cdot(\log(16SAH/\delta)+2\log(\mathrm{e}\mathfrak{t}))}\\ &+ 12\kappa\mathrm{e}^{10}\sqrt{H^2SA(\mathfrak{t}-1)\log^3(\mathfrak{t})\cdot(\log(4SAH/\delta)+2\log(\mathrm{e}\mathfrak{t}))}\\ &+ 16\mathrm{e}^{10}H^3SA\log(\mathfrak{t})\mathrm{R}_{\max}(\log(4SAH/\delta)+S\log(\mathrm{e}\mathfrak{t})).\end{aligned}$$

Under assumption $\mathfrak{t} \geqslant 2$ we can proceed with the further simplifications

$$\begin{aligned}\mathfrak{t}\varepsilon \leqslant 216\mathrm{e}^{10}\mathrm{R}_{\max}\sqrt{\mathfrak{t}H^3SA\log^3(\mathfrak{t})\cdot(\log(16SAH/\delta)+2\log(\mathrm{e}\mathfrak{t}))} \\ + 32\mathrm{e}^{10}H^3SA\log(\mathfrak{t})\mathrm{R}_{\max}(\log(16SAH/\delta)+S\log(\mathrm{e}\mathfrak{t})).\end{aligned} \tag{A.13}$$

Let us define the following constants

$$A = 216\mathrm{e}^{10}\mathrm{R}_{\max}\cdot\sqrt{\frac{H^3SA}{\varepsilon^2}}, \quad B = \log(16SAH/\delta), \quad C = \frac{32\mathrm{e}^{10}\cdot H^3SA\mathrm{R}_{\max}}{\varepsilon}.$$

Then inequality (A.13) has the following form

$$\mathfrak{t} \leqslant A\sqrt{\mathfrak{t}(B+2\log(\mathrm{e}\mathfrak{t}))\cdot\log^3(\mathfrak{t})} + C(B+S\log(\mathrm{e}\mathfrak{t}))\log\mathfrak{t}.$$

First, we obtain a loose inequality on $\mathfrak{t}$. Let us use the inequality $\log(x) \leqslant x^\beta/\beta$ for any $x>0, \beta>0$ with different β for each logarithm

$$\begin{aligned}&\mathfrak{t} \leqslant A\sqrt{216\mathfrak{t}(B+4(\mathrm{e}\mathfrak{t})^{1/4})\mathfrak{t}^{1/2}} + 4C(B+8S/3(\mathrm{e}\mathfrak{t})^{3/8})\mathfrak{t}^{1/4} \\ \Rightarrow\ &\mathfrak{t}^{3/4} \leqslant \mathfrak{t}^{3/8}\left(6A\sqrt{6(B+4\mathrm{e}^{1/4})} + 12CS\mathrm{e}^{3/8}\right) + 4CB.\end{aligned}$$

Notice that the solution to the inequality $x^2 \leqslant ax+b$ could be upper-bounded as follows

$$x \leqslant \frac{a+\sqrt{a^2+4b}}{2} \leqslant a+\sqrt{b},$$

thus

$$\mathfrak{t}^{3/8} \leqslant \left(6A\sqrt{6(B+4\mathrm{e}^{1/4})} + 12CS\mathrm{e}^{3/8}\right) + 2\sqrt{CB}.$$

Define $L = 8/3\log\left(6A\sqrt{6(B+4\mathrm{e}^{1/4})} + 12CS\mathrm{e}^{3/8} + 2\sqrt{CB}\right)$ that is of order $\mathcal{O}(\log(SAH\mathrm{R}_{\max}/\varepsilon)+\log\log(SAH/\delta))$ and we have $\log(\mathfrak{t}) \leqslant L$. Then we have that the solution to (A.13) is a subset of solutions to

$$\mathfrak{t} \leqslant A\sqrt{\mathfrak{t}(B+2(1+L))\cdot L^3} + C(B+S(1+L))L,$$

solving this inequality we obtain the bound

$$\mathfrak{t} \leqslant 2A^2(B+2(1+L))L^3 + 2CB(S(1+L))L.$$

□

After this general result we state the bound for the MTEE problem that was stated in the main text.

Theorem A.21. *For all $\varepsilon>0$ and $\delta\in(0,1)$ the* `UCBVI-Ent` *algorithm is (ε,δ)-PAC for MTEE. Moreover, with probability at least $1-\delta$*

$$\mathfrak{t} \leqslant \mathcal{O}\left(\frac{H^3SA\log^2(SA)\log(SAH/\delta)\cdot L^4}{\varepsilon^2} + \frac{H^3SA(\log(SAH/\delta)+SL)\cdot L}{\varepsilon}\right),$$

where $L = \log(SAH/\varepsilon)+\log\log(SAH/\delta)$.

Proof. Fix $\Phi(\pi) = -\mathcal{H}(\pi), \kappa = \lambda = 1$ and $r_{\max} = 0$. Since $\mathcal{H}(\pi)$ is 1-strongly convex with respect to ℓ_1-norm, we have $r_A = 1$. Also we automatically have $\mathrm{R}_{\max} = \log(SA)$. In this setting, Theorem A.20 yields the desired statement. □

A.5 Fast Rates for MTEE and Regularized MDPs

In this section we describe an algorithm that will achieve $\widetilde{\mathcal{O}}(\text{poly}(S, A, H)/\varepsilon)$ sample complexity for regularized MDPs. Additionally, we show that this algorithm could be used for reward-free exploration under regularization.

A.5.1 RL-Explore-Ent Algorithm

We leverage the reward-free exploration approach by Jin et al. (2020c). Our algorithm is split into two phases: the first phase is devoted to reward-free exploration, and on the second phase the collected samples are used to build estimates of transition probabilities and entropy of transitions. The main idea is that regularization allows us to collect much smaller number of samples to control the policy error.

Exploration phase. We first learn a policy that visit uniformly the MDP. To this aim for each state $s' \in \mathcal{S}$ at each step $h' \in [H]$ we build the reward that put one on this state at step h and zeros everywhere else $r_h(s, a) = \mathbb{1}\{(s, h) = (s', h')\}$. We note that the reward function does not depend on action taken. We then run the EULER algorithm for N_0 episodes in the MDP equipped with the reward r and denote by $\widetilde{\Pi}_{s',h'}$ the set of N_0 policies used by EULER to interact with the MDP. We modify this set of policies by forcing to act uniformly at the goal state s' into the set

$$\Pi_{s',h'} = \left\{ \pi'_h(a|s) = \begin{cases} 1/A & \text{if } s = s', h = h' \\ \pi_h(s, a) & \text{else} \end{cases} : \ \pi \in \widetilde{\Pi}_{s',h'} \right\}.$$

We define the (non-Markovian) policy π^{mix} as the uniform mixture of the policies $\{\pi \in \Pi_{s,h}, (s, h) \in \mathcal{S} \times [H]\}$ we just constructed. As proved by Jin et al. (2020c) the policy π^{mix} is built such that it will visit almost uniformly all the states that can be reached in the MDP from the initial state. Before precising this property we need to introduce the notion of significant state.

Definition A.22. A state s at step h is called ε'-significant if there exists a policy π such that the visitation probability of s under policy π is greater than ε':

$$\max_\pi d_h^\pi(s) \geqslant \varepsilon'.$$

The set of all ε'-significant state-step pairs is called $S_{\varepsilon'}$.

We reproduce here the result by Jin et al. (2020c) that shows that the policy π^{mix} will visit any significant state with a large enough probability.

Theorem A.23 (Theorem 3.3 by Jin et al. 2020c)**.** *There exists an absolute constant $c > 0$ such that for any $\varepsilon' > 0$ and $\delta \in (0, 1)$, if we set the parameter $N_0 \geqslant cS^2AH^4L/\varepsilon'$ where $L = \log(SAH/(\delta\varepsilon'))$, then with probability at least $1 - \delta/3$ the following event holds*

$$\mathcal{E}^{\text{RF-RL-Explore}}(\delta, \varepsilon') = \left\{ \forall (s, h) \in S_{\varepsilon'}, \forall a \in \mathcal{A}, \forall \pi : \frac{d_h^\pi(s, a)}{\mu_h(s, a)} \leqslant 2SAH \right\},$$

where we denote the visitation distribution of policy π^{mix} by $\mu_h(s, a) = d_h^{\pi^{\text{mix}}}(s, a)$.

Remark A.24. Note that the space complexity of RL-Explore-Ent is very large since we need to store all the intermediate policies in order to construct π^{mix}.

This policy π^{mix} is then used to collect N new independent trajectories $(z_n)_{n\in[N]}$ where $z_n = (s_1^n, a_1^n, \ldots, s_H^n, a_H^n, s_{H+1}^n)$ by following the policy π^{mix} in the original MDP. Next define the set $\mathcal{D} = \{(s_h^n, a_h^n, s_{h+1}^n), h \in [H], n \in [N]\}$ consisting of the transitions in the sampled trajectories.

Planning phase. Given the transitions collected in the exploration phase we estimate the transition probability distributions and then plan in the estimated MDP with the Bellman equations for MTEE to obtain a maximum trajectory entropy policy.

Using the dataset $\mathcal{D}$, we construct estimates of transition probabilities $\{\widehat{p}_h\}_{h\in[H]}$. We first define the number of visits of a state action pair (s,a) at step h and the number of transitions for (s,a) at step h to a states s' observed in the dataset $\mathcal{D}$,

$$n_h(s,a) = \sum_{n=1}^{N} \mathbb{1}\{(s_h^n, a_h^n) = (s,a)\} \quad n_h(s'|s,a) = \sum_{n=1}^{N} \mathbb{1}\{(s_h^n, a_h^n, s_{h+1}^n) = (s,a,s')\},$$

The transitions are estimated using the maximum likelihood method:

$$\widehat{p}_h(s'|s,a) = \begin{cases} \frac{n_h(s'|s,a)}{n_h(s,a)} & n_h(s,a) > 0 \\ \frac{1}{S} & n_h(s,a) = 0 \end{cases}. \tag{A.14}$$

Given these estimates, we can define an empirical version of the regularized Bellman equations for MTEE

$$\begin{aligned} \widehat{Q}^{\pi}_{\lambda,h}(s,a) &= r_h(s,a) + \kappa\mathcal{H}(\widehat{p}_h(s,a)) + \widehat{p}_h\widehat{V}^{\pi}_{\lambda,h+1}(s,a) \\ \widehat{V}^{\pi}_{\lambda,h}(s) &= \pi\widehat{Q}^{\pi}_{\lambda,h}(s) - \lambda\Phi(\pi). \end{aligned} \tag{A.15}$$

Then the output policy $\widehat{\pi}$ is the solution to the optimal regularized Bellman equations

$$\begin{aligned} \widehat{Q}^{\star}_{\lambda,h}(s,a) &= r_h(s,a) + \kappa\mathcal{H}(\widehat{p}_h(s,a)) + \widehat{p}_h\widehat{V}^{\star}_{\lambda,h+1}(s,a) \\ \widehat{V}^{\star}_{\lambda,h}(s) &= \max_{\pi}\Big\{\pi\widehat{Q}^{\star}_{\lambda,h}(s) - \lambda\Phi(\pi)\Big\} \\ \widehat{\pi}_h(s) &= \arg\max_{\pi}\Big\{\pi\widehat{Q}^{\star}_{\lambda,h}(s) - \lambda\Phi(\pi)\Big\}. \end{aligned} \tag{A.16}$$

We call this algorithm `RL-Explore-Ent`. Notably, we can extend this algorithm to the setting of the changing rewards by solving (A.16) with new reward functions $r_h(s,a)$. The detailed description of the algorithm is presented in Algorithm 2. The only difference between our algorithm and `RF-RL-Explore` by Jin et al. (2020c) is the use of a smaller number of trajectories N and solving regularized Bellman equations instead of usual one.

A.5.2 Concentration Events

In this section we describe all required concentration events.

Let $\beta^{\mathrm{conc}}\colon (0,1)\times\mathbb{N}\to\mathbb{R}_+$ and $\beta^{\mathrm{cnt}}\colon (0,1)\to\mathbb{R}_+$ be some functions defined later on in Lemma A.25. We define the following favorable events

$$\mathcal{E}^{\mathrm{conc}}(N,\delta) \triangleq \Bigg\{ \forall h\in[H], \forall G\colon \mathcal{S}\to[0,H\mathrm{R}_{\max}], \forall \nu\colon \mathcal{S}\to\mathcal{A}: \\ \mathbb{E}_{(s,a)\sim\mu_h}\Big[([\widehat{p}_{h'} - p_{h'}]G(s,a))^2\mathbb{1}\{\nu(s)=a\}\Big] \leqslant \frac{CH^2\mathrm{R}_{\max}^2 S\cdot\beta^{\mathrm{conc}}(\delta,N)}{N}\Bigg\},$$

$$\mathcal{E}^{\mathcal{H}}(N,\delta) \triangleq \left\{\forall h\in[H] : \mathbb{E}_{(s,a)\sim\mu_h}\Big[(\mathcal{H}(\widehat{p}_h(s,a)) - \mathcal{H}(p_h(s,a)))^2\Big] \leqslant \frac{12S^2A\log^2(SN)\cdot\beta^{\mathcal{H}}(\delta)}{N}\right\},$$

where C is a some absolute constant. We also introduce two intersections of these events of interest and $\mathcal{E}^{\texttt{RF-RL-Explore}}(\delta)$, defined in Theorem A.23: $\mathcal{G}(N,\delta,\varepsilon') \triangleq \mathcal{E}^{\mathrm{conc}}(N,\delta)\cap\mathcal{E}^{\mathcal{H}}(\delta)\cap\mathcal{E}^{\texttt{RF-RL-Explore}}(\delta,\varepsilon')$. We prove that for the right choice of the functions $\beta^{\mathrm{conc}}, \beta^{\mathcal{H}}$ the above events hold with high probability.

Lemma A.25. *For any $\delta \in (0,1), \varepsilon' > 0, N \in \mathbb{N}$ and for the following choices of functions β,*

$$\beta^{\mathrm{conc}}(\delta, N) \triangleq \log(3AH\mathrm{R}_{\max}N/\delta),$$
$$\beta^{\mathcal{H}}(\delta) \triangleq \log(12SAH/\delta),$$

it holds that

$$\mathbb{P}[\mathcal{E}^{\mathrm{conc}}(\delta)] \geqslant 1-\delta/3, \qquad \mathbb{P}[\mathcal{E}^{\mathcal{H}}(\delta)] \geqslant 1-\delta/3.$$

In particular, $\mathbb{P}[\mathcal{G}(\delta, N, \varepsilon')] \geqslant 1-\delta$.

Proof. Holds from an application of Lemma D.20, Lemma D.19, Theorem A.23 and union bound. □

A.5.3 Sample Complexity Proof

In this section we provide the sample complexity result of `RL-Explore-Ent` algorithm in the simple BPI setting and in the reward free setting.

Theorem A.26. *Algorithm* `RL-Explore-Ent` *with parameters $N_0 = \Omega\Big(\frac{H^7S^3Ar_A^2\cdot\mathrm{R}_{\max}^2\cdot L}{\varepsilon\lambda}\Big)$ and $N \geqslant \Omega\Big(\frac{H^6S^3Ar_A^2\mathrm{R}_{\max}^2L^3}{\varepsilon\lambda}\Big)$ is (ε,δ)-PAC for the best policy identification in regularized MDPs, where $L = \log(SAH/(\varepsilon\lambda\delta))$. The sample complexity is bounded by*

$$\widetilde{\mathcal{O}}\left(\frac{H^8S^4Ar_A^2\mathrm{R}_{\max}^2}{\varepsilon\lambda}\right).$$

Proof. Let us start from exploiting the strong convexity of the regularizer. This property is given by Lemma A.29

$$V^{\star}_{\lambda,1}(s_1) - V^{\widehat{\pi}}_{\lambda,1}(s_1) \leqslant \frac{r_A^2}{2\lambda}\sum_{h=1}^{H}\mathbb{E}_{\widehat{\pi}}\Big[\max_{a\in\mathcal{A}}\Big(\widehat{Q}^{\widehat{\pi}}_{\lambda,h} - Q^{\star}_{\lambda,h}\Big)^2(s_h,a)\Big|s_1\Big].$$

Next we study each separate term in this decomposition. By the definition of $\widehat{\pi}$ and $\pi^{\star}$ we have

$$\widehat{Q}^{\pi^{\star}}_{\lambda,h}(s,a) - Q^{\pi^{\star}}_{\lambda,h}(s,a) \leqslant \widehat{Q}^{\widehat{\pi}}_{\lambda,h}(s,a) - Q^{\star}_{\lambda,h}(s,a) \leqslant \widehat{Q}^{\widehat{\pi}}_{\lambda,h}(s,a) - Q^{\widehat{\pi}}_{\lambda,h}(s,a),$$

thus

$$\Big(\widehat{Q}^{\pi^{\star}}_{\lambda,h}(s,a) - Q^{\pi^{\star}}_{\lambda,h}(s,a)\Big)^2 \leqslant \max\Big\{\Big(\widehat{Q}^{\widehat{\pi}}_{\lambda,h}(s,a) - Q^{\star}_{\lambda,h}(s,a)\Big)^2, \Big(\widehat{Q}^{\widehat{\pi}}_{\lambda,h}(s,a) - Q^{\widehat{\pi}}_{\lambda,h}(s,a)\Big)^2\Big\},$$

and by an inequality $\max\{a,b\} \leqslant a+b$ for positive a,b we have

$$\Big(\widehat{Q}^{\widehat{\pi}}_{\lambda,h}(s,a) - Q^{\star}_{\lambda,h}(s,a)\Big)^2 \leqslant \Big(\widehat{Q}^{\widehat{\pi}}_{\lambda,h}(s,a) - Q^{\widehat{\pi}}_{\lambda,h}(s,a)\Big)^2 + \Big(\widehat{Q}^{\pi^{\star}}_{\lambda,h}(s,a) - Q^{\pi^{\star}}_{\lambda,h}(s,a)\Big)^2.$$

Therefore, the policy error decomposes as follows

$$V^{\star}_{\lambda,1}(s_1) - V^{\widehat{\pi}}_{\lambda,1}(s_1) \leqslant \frac{r_A^2}{2\lambda}\sum_{h=1}^{H}\Bigg(\mathbb{E}_{\widehat{\pi}}\Big[\max_{a\in\mathcal{A}}\Big(\widehat{Q}^{\widehat{\pi}}_{\lambda,h}(s_h,a) - Q^{\widehat{\pi}}_{\lambda,h}(s_h,a)\Big)^2 \mid s_1\Big] + \mathbb{E}_{\widehat{\pi}}\Big[\max_{a\in\mathcal{A}}\Big(\widehat{Q}^{\pi^{\star}}_{\lambda,h}(s_h,a) - Q^{\pi^{\star}}_{\lambda,h}(s_h,a)\Big)^2 \mid s_1\Big]\Bigg).$$

Next we assume that the event $\mathcal{G}(N, \delta, \varepsilon')$ holds for the values N and ε' that will be specified later. Then Lemma A.30 applied $2H$ times yields

$$V^{\star}_{\lambda,1}(s_1) - V^{\widehat{\pi}}_{\lambda,1}(s_1) \leqslant \frac{r_A^2}{\lambda}\left(\frac{48S^3H^4A\mathrm{R}_{\max}^2 \log^2(N)\log(12SAH/\delta)}{N} + \frac{4CH^6\mathrm{R}_{\max}^2 S^2A \cdot (\log(3AH\mathrm{R}_{\max}/\delta) + \log(N))}{N} + 2SH^3\mathrm{R}_{\max}^2 \cdot \varepsilon'\right).$$

Next we take

$$\varepsilon' = \frac{\lambda\varepsilon}{4r_A^2 SH^3\mathrm{R}_{\max}^2},$$

that requires to take $N_0 \geqslant \frac{cH^7S^3Ar_A^2\cdot\mathrm{R}_{\max}^2\cdot L}{\varepsilon\lambda}$ for an absolute constant $c > 0$ and $L = \log(SAH/\delta) + \log(1/(\varepsilon\lambda))$. This yields $\widetilde{\mathcal{O}}(H^8S^4Ar_A^2\mathrm{R}_{\max}^2/(\varepsilon\lambda))$ sample complexity of the first phase, since we need N_0 samples for each $(s,h) \in \mathcal{S}\times[H]$. Under this choice, we have

$$V^{\star}_{\lambda,1}(s_1) - V^{\widehat{\pi}}_{\lambda,1}(s_1) \leqslant \varepsilon/2 + \frac{r_A^2}{\lambda N}\cdot\Big((48+4C)S^3AH^6\mathrm{R}_{\max}^2\log^2(N)\cdot\log(12SAH\mathrm{R}_{\max}/\delta)\Big).$$

To make the second part smaller than ε, we have to analyze the following inequality

$$\log^2(N)/N \cdot B \leqslant \varepsilon$$

and upper bound its minimal solution that we will call $N^\star$. To do it, we first use a simple numeric bound $\log(N) \leqslant 4N^{1/4}$ and obtain a simple estimate $N \geqslant 256B^2/\varepsilon^2$, thus the minimal solution $N^\star \leqslant 256B^2/\varepsilon^2$. Therefore, we can assume that $\log(N^\star) \leqslant 2\log(16B/\varepsilon) = \mathcal{O}(\log(SAH\mathrm{R}_{\max}/\varepsilon + \log(1/\delta))$, thus taking

$$N \geqslant N^\star = \Omega\left(\frac{H^6S^3Ar_A^2\mathrm{R}_{\max}^2\log^2(SAH\mathrm{R}_{\max}/\varepsilon + \log(1/\delta))\cdot\log(SAH\mathrm{R}_{\max}/\delta)}{\varepsilon\lambda}\right).$$

is enough to guarantee that the policy error is smaller than ε. □

Notice that in the proof we do not rely on one particular reward function, since the only we need is conditioning on event $\mathcal{G}(\delta)$ that does not depend on the particular reward function.

Corollary A.27. *Algorithm* RL-Explore-Ent *for a choice* $N_0 = \Omega\big(\frac{H^7S^3Ar_A^2\cdot\mathrm{R}_{\max}^2\cdot L}{\varepsilon\lambda}\big)$ *and* $N \geqslant \Omega\big(\frac{H^6S^3Ar_A^2\mathrm{R}_{\max}^2L^3}{\varepsilon\lambda}\big)$ *outputs* ε*-optimal policies for an arbitrary number of reward functions in regularized MDPs. The sample complexity is bounded by*

$$\widetilde{\mathcal{O}}\left(\frac{H^8S^4Ar_A^2\mathrm{R}_{\max}^2}{\varepsilon\lambda}\right).$$

Finally, we provide a formal proof for application of this algorithm to the MTEE problem, that is a simple application of the results above.

Theorem A.28. *Algorithm* RL-Explore-Ent *with parameters*

$$N_0 = \Omega\left(\frac{H^7S^3A\cdot L^3}{\varepsilon}\right) \quad \text{and} \quad N = \Omega\left(\frac{H^6S^3AL^5}{\varepsilon}\right)$$

is (ε,δ)*-PAC for the MTEE problem, where* $L = \log(SAH/(\varepsilon\delta))$*. The total sample complexity* $SHN_0 + N$ *is bounded by*

$$\widetilde{\mathcal{O}}\left(\frac{H^8S^4A}{\varepsilon}\right).$$

Proof. Fix $\Phi(\pi) = -\mathcal{H}(\pi), \kappa = \lambda = 1$ and $r_{\max} = 0$. By 1-strong convexity of $-\mathcal{H}(\pi)$ with respect to ℓ_1-norm, its dual is 1-strongly convex with respect to ℓ_∞ norm, yielding $r_A = 1$. Also we have $\mathrm{R}_{\max} = \log(SA)$, thus by Theorem A.26 we conclude the statement. □

A.5.4 Technical Lemmas

Lemma A.29. *Let π be a greedy policy with respect to regularized Q-values $\overline{Q}_{\lambda,h}(s,a) : \pi_h(s) = \arg\max_\pi \{\pi \overline{Q}_{\lambda,h}(s) - \lambda\Phi(\pi)\}$. Then the following error decomposition holds*

$$V^\star_{\lambda,1}(s_1) - V^\pi_{\lambda,1}(s_1) \leqslant \frac{r_A^2}{2\lambda}\mathbb{E}_\pi\left[\sum_{h=1}^H \max_{a\in\mathcal{A}}\left(\overline{Q}_{\lambda,h} - Q^\star_{\lambda,h}\right)^2(s_h,a) \mid s_1\right],$$

where r_A is a constant defined in (A.6).

Proof. First, we formulate the statement dependent of h

$$V^\star_{\lambda,h}(s) - V^\pi_{\lambda,h}(s) \leqslant \frac{r_A^2}{2\lambda}\mathbb{E}_\pi\left[\sum_{h'=h}^H \max_{a\in\mathcal{A}}\left(\overline{Q}_{\lambda,h'} - Q^\star_{\lambda,h'}\right)^2(s_{h'},a) \mid s_h = s\right]. \tag{A.17}$$

Notice that for $h = 1$ and $s = s_1$ the initial statement is recovered. We proceed by induction over h. The initial case $h = H + 1$ is trivial, next we assume that the statement (A.17) is true for any $h' > h$.

We start the analysis from understanding the policy error by applying the smoothness of F_λ for any h.

$$\begin{aligned}
V^\star_{\lambda,h}(s) - V^\pi_{\lambda,h}(s) &= F_\lambda(Q^\star_{\lambda,h}(s,\cdot)) - \left(\pi_h Q^\pi_{\lambda,h}(s,\cdot) - \lambda\Phi(\pi_h(s))\right) \\
&\leqslant F_\lambda(\overline{Q}_h)(s) + \langle \nabla F_\lambda(\overline{Q}_h(s,\cdot)), Q^\star_{\lambda,h}(s,\cdot) - \overline{Q}_h(s,\cdot)\rangle + \frac{1}{2\lambda}\|\overline{Q}_h - Q^\star_{\lambda,h}\|^2_*(s) \\
&\quad - \left(\pi_h Q^\pi_{\lambda,h}(s,\cdot) - \lambda\Phi(\pi_h(s))\right).
\end{aligned}$$

Next we recall that

$$\pi_h(s) = \nabla F(\overline{Q}_h(s,\cdot)), \quad F(\overline{Q}_h)(s) = \pi_h\overline{Q}_h(s) - \lambda\Phi(\pi_h(s)),$$

thus we have

$$F(\overline{Q}_h)(s) - \left(\pi_h Q^\pi_{\lambda,h}(s,\cdot) - \lambda\Phi(\pi_h(s))\right) = \pi_h[\overline{Q}_h - Q^\pi_{\lambda,h}](s)$$

and, by Bellman equations

$$\begin{aligned}
V^\star_{\lambda,h}(s) - V^\pi_{\lambda,h}(s) &\leqslant \pi_h\left[Q^\star_{\lambda,h} - Q^\pi_{\lambda,h}\right](s) + \frac{1}{2\lambda}\|\overline{Q}_h - Q^\star_{\lambda,h}\|^2_*(s) \\
&\leqslant \pi_h p_h\left[V^\star_{\lambda,h+1} - V^\pi_{\lambda,h+1}\right](s) + \frac{1}{2\lambda}\|\overline{Q}_h - Q^\star_{\lambda,h}\|^2_*(s).
\end{aligned}$$

Applying norm equivalence (A.6) we have

$$V^\star_{\lambda,h}(s) - V^\pi_{\lambda,h}(s) \leqslant \mathbb{E}_\pi\left[\frac{r_A^2}{2\lambda}\|\overline{Q}_h - Q^\star_{\lambda,h}\|^2_\infty(s_h) + V^\star_{\lambda,h+1}(s_{h+1}) - V^\pi_{\lambda,h+1}(s_{h+1}) \mid s_h = s\right].$$

By induction hypothesis we conclude the statement. □

Lemma A.30. *For any policy π the following holds on event $\mathcal{G}(N,\delta,\varepsilon')$ defined in Lemma A.25 for any $h \in [H]$*

$$\begin{aligned}
\mathbb{E}_{\widehat{\pi}}\left[\max_{a\in\mathcal{A}}\left(\widehat{Q}^\pi_{\lambda,h} - Q^\pi_{\lambda,h}\right)^2(s_h,a) \mid s_1\right] &\leqslant \frac{48S^3H^3A\mathrm{R}^2_{\max}\log^2(N)\beta^{\mathcal{H}}(\delta)}{N} + \frac{4CH^5\mathrm{R}^2_{\max}S^2A\cdot\beta^{\mathrm{conc}}(\delta,N)}{N} \\
&\quad + 2SH^2\mathrm{R}^2_{\max}\cdot\varepsilon'.
\end{aligned}$$

Proof. By performance-difference Lemma A.11 and form of the rewards stated in (A.15) we have for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$

$$\widehat{Q}^{\pi}_{\lambda,h}(s,a) - Q^{\pi}_{\lambda,h}(s,a) = \kappa \mathbb{E}_{\pi}\left[\sum_{h'=h}^{H} \mathcal{H}(\widehat{p}_{h'}(s_{h'}, a_{h'})) - \mathcal{H}(p_{h'}(s_{h'}, a_{h'})) \mid (s_h, a_h) = (s,a)\right] + \mathbb{E}_{\pi}\left[\sum_{h'=h}^{H} [\widehat{p}_{h'} - p_{h'}]\widehat{V}^{\pi}_{\lambda,h'+1}(s_{h'}, a_{h'}) \mid (s_h, a_h) = (s,a)\right].$$

Next we analyze all required expectation for one fixed value $h \in [H]$. By Jensen's inequality and a simple algebraic inequality $(a+b)^2 \leqslant 2a^2 + 2b^2$

$$\left(\widehat{Q}^{\pi}_{\lambda,h}(s,a) - Q^{\pi}_{\lambda,h}(s,a)\right)^2 \leqslant 2\kappa^2 \mathbb{E}_{\pi}\left[\left(\sum_{h'=h}^{H} \mathcal{H}(\widehat{p}_{h'}(s_{h'}, a_{h'})) - \mathcal{H}(p_{h'}(s_{h'}, a_{h'})\right)^2 \middle| (s_h, a_h) = (s,a)\right] + 2\mathbb{E}_{\pi}\left[\left(\sum_{h'=h}^{H} [\widehat{p}_{h'} - p_{h'}]\widehat{V}^{\pi}_{\lambda,h'+1}(s_{h'}, a_{h'})\right)^2 \middle| (s_h, a_h) = (s,a)\right].$$

By Cauchy–Schwarz inequality we have the final form

$$\left(\widehat{Q}^{\pi}_{\lambda,h}(s,a) - Q^{\pi}_{\lambda,h}(s,a)\right)^2 \leqslant 2H\mathbb{E}_{\pi}\left[\sum_{h'=h}^{H} \kappa^2 (\mathcal{H}(\widehat{p}_{h'}(s_{h'}, a_{h'})) - \mathcal{H}(p_{h'}(s_{h'}, a_{h'})))^2 + \sum_{h'=h}^{H} \left([\widehat{p}_{h'} - p_{h'}]\widehat{V}^{\pi}_{\lambda,h'+1}\right)^2 (s_{h'}, a_{h'}) \mid (s_h, a_h) = (s,a)\right] \tag{A.18}$$

To connect this conditional expectation in the right-hand side of (A.18) with conditional expectation over $\widehat{\pi}$, we define the following policy

$$\widetilde{\pi}_{h'}(a|s) = \begin{cases} \widehat{\pi}_{h'}(a|s) & h' < h \\ \mathbb{1}\{a = \arg\max_{a \in \mathcal{A}} \left\{\left(\widehat{Q}^{\pi}_{h}(s,a) - Q^{\pi}_{h}(s,a)\right)^2\right\} & h' = h \\ \pi_h(a|s) & h' > h. \end{cases}$$

Since we do not change the policy π for steps greater than h, we can replace π with $\widetilde{\pi}$ in (A.18). Additionally, since this policy is equal to $\widehat{\pi}$ for the first $h-1$ steps, the distribution $d^{\widehat{\pi}}_h(s_h)$ is equal to $d^{\widetilde{\pi}}_h(s_h)$:

$$\begin{aligned} d^{\widehat{\pi}}_h(s_h) &= \sum_{s_1, a_1, \ldots, s_{h-1}, a_{h-1}} \widehat{\pi}_1(a_1|s_1)\left(\prod_{h'=1}^{h-1} p_{h'-1}(s_{h'}|s_{h'-1}, a_{h'-1})\widehat{\pi}_{h'}(a_{h'}|s_{h'})\right) \cdot p_{h-1}(s_h|s_{h-1}, a_{h-1}) \\ &= \sum_{s_1, a_1, \ldots, s_{h-1}, a_{h-1}} \widetilde{\pi}_1(a_1|s_1)\left(\prod_{h'=1}^{h-1} p_{h'-1}(s_{h'}|s_{h'-1}, a_{h'-1})\widetilde{\pi}_{h'}(a_{h'}|s_{h'})\right) \cdot p_{h-1}(s_h|s_{h-1}, a_{h-1}) \\ &= d^{\widetilde{\pi}}_h(s_h). \end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}_{\widehat{\pi}}\Big[\max_{a\in\mathcal{A}}\Big(\widehat{Q}^{\pi}_{\lambda,h}-Q^{\pi}_{\lambda,h}\Big)^2(s_h,a)\mid s_1\Big] &= \sum_s d^{\widehat{\pi}}_h(s)\max_{a\in\mathcal{A}}\Big(\widehat{Q}^{\pi}_{\lambda,h}-Q^{\pi}_{\lambda,h}\Big)^2(s,a)\\ &= \sum_s d^{\widetilde{\pi}}_h(s)\max_{a\in\mathcal{A}}\Big(\widehat{Q}^{\pi}_{\lambda,h}-Q^{\pi}_{\lambda,h}\Big)^2(s,a)\\ &= \mathbb{E}_{\widetilde{\pi}}\Big[\max_{a\in\mathcal{A}}\Big(\widehat{Q}^{\pi}_{\lambda,h}-Q^{\pi}_{\lambda,h}\Big)^2(s_h,a)\mid s_1\Big]\\ &= \mathbb{E}_{\widetilde{\pi}}\Big[\Big(\widehat{Q}^{\pi}_{\lambda,h}-Q^{\pi}_{\lambda,h}\Big)^2(s_h,a_h)\mid s_1\Big].\end{aligned}$$

Next we show that in (A.18) we can make a change of policy from π to $\widetilde{\pi}$. It is enough to show that the required marginal distributions are equal for all $h'\geqslant h$, i.e. for any $(s,a)\in\mathcal{S}\times\mathcal{A}$ it holds $\mathbb{P}_{\pi}[(s_{h'},a_{h'})|(s_h,a_h)]=\mathbb{P}_{\widetilde{\pi}}[(s_{h'},a_{h'})|(s_h,a_h)]$. For $h'=h$ this probability is an indicator on (s_h,a_h), so it does not depend on policy. For the general case $h'>h$ we can use Markov property and imply

$$\begin{aligned}\mathbb{P}_{\pi}[(s_{h'},a_{h'})|(s_h,a_h)] &= \sum_{(s_{h+1},a_{h+1},\dots,s_{h'-1},a_{h'-1})}\prod_{i=h}^{h'-1}\pi_{i+1}(a_{i+1}|s_{i+1})p_i(s_{i+1}|s_i,a_i)\\ &= \sum_{(s_{h+1},a_{h+1},\dots,s_{h'-1},a_{h'-1})}\prod_{i=h}^{h'-1}\widetilde{\pi}_{i+1}(a_{i+1}|s_{i+1})p_i(s_{i+1}|s_i,a_i)\\ &= \mathbb{P}_{\widetilde{\pi}}[(s_{h'},a_{h'})|(s_h,a_h)].\end{aligned}$$

Therefore, we can make change of measure in (A.18) and obtain

$$\begin{aligned}\mathbb{E}_{\widehat{\pi}}\Big[\max_{a\in\mathcal{A}}\Big(\widehat{Q}^{\pi}_{\lambda,h}-Q^{\pi}_{\lambda,h}\Big)^2(s_h,a)\mid s_1\Big] \leqslant 2H\mathbb{E}_{\widetilde{\pi}}\Bigg[\mathbb{E}_{\widetilde{\pi}}\Bigg[\sum_{h'=h}^{H}\kappa^2\big(\mathcal{H}(\widehat{p}_{h'}(s_{h'},a_{h'}))-\mathcal{H}(p_{h'}(s_{h'},a_{h'}))\big)^2\mid s_h,a_h\Bigg]\\ +\mathbb{E}_{\widetilde{\pi}}\Bigg[\sum_{h'=h}^{H}\Big([\widehat{p}_{h'}-p_{h'}]\widehat{V}^{\pi}_{\lambda,h'+1}\Big)^2(s_{h'},a_{h'})\mid s_h,a_h\Bigg]\;\Bigg|\;s_1\Bigg].\end{aligned}$$

By the properties of conditional expectation we can eliminate the inner expectation and obtain the following upper bound

$$\begin{aligned}\mathbb{E}_{\widehat{\pi}}\Big[\max_{a\in\mathcal{A}}\Big(\widehat{Q}^{\pi}_{\lambda,h}-Q^{\pi}_{\lambda,h}\Big)^2(s_h,a)\mid s_1\Big] \leqslant 2H\kappa^2\sum_{h'=h}^{H}\mathbb{E}_{\widetilde{\pi}}\Big[\big(\mathcal{H}(\widehat{p}_{h'}(s_{h'},a_{h'}))-\mathcal{H}(p_{h'}(s_{h'},a_{h'}))\big)^2\;\Big|\;s_1\Big]\\ +2H\sum_{h'=h}^{H}\mathbb{E}_{\widetilde{\pi}}\Bigg[\Big([\widehat{p}_{h'}-p_{h'}]\widehat{V}^{\pi}_{\lambda,h'+1}\Big)^2(s_{h'},a_{h'})\;\Bigg|\;s_1\Bigg].\end{aligned}$$

Applying Lemma A.31 and the definition of $\mathcal{E}^{\mathcal{H}}(\delta)$ we have

$$\begin{aligned}\mathbb{E}_{\widetilde{\pi}}\Big[\big(\mathcal{H}(\widehat{p}_{h'}(s_{h'},a_{h'}))-\mathcal{H}(p_{h'}(s_{h'},a_{h'}))\big)^2\mid s_1\Big] &\leqslant 2SAH\mathbb{E}_{(s,a)\sim\mu_{h'}}\Big[\big(\mathcal{H}(\widehat{p}_{h'}(s,a))-\mathcal{H}(p_{h'}(s,a))\big)^2\Big]\\ &\quad+S\log^2(S)\varepsilon'\\ &\leqslant\frac{24S^3HA\log^2(SN)\beta^{\mathcal{H}}(\delta)}{N}+S\log^2(S)\varepsilon'.\end{aligned}$$

In the same way by Lemma A.31 and the definition of event $\mathcal{E}^{\mathrm{conc}}(N,\delta)$

$$\begin{aligned}\mathbb{E}_{\widetilde{\pi}}\Big[\Big([\widehat{p}_{h'}-p_{h'}]\widehat{V}^{\pi}_{\lambda,h'+1}\Big)^2(s_{h'},a_{h'})\mid s_1\Big] &\leqslant 2SAH\mathbb{E}_{(s,a)\sim\mu_{h'}}\Big[\Big([\widehat{p}_{h'}-p_{h'}]\widehat{V}^{\pi}_{\lambda,h'+1}\Big)^2(s,a)\Big]+SH^2\mathrm{R}^2_{\max}\varepsilon'\\ &\leqslant\frac{2CH^3\mathrm{R}^2_{\max}S^2A\cdot\beta^{\mathrm{conc}}(\delta,N)}{N}+SH^2\mathrm{R}^2_{\max}\varepsilon'.\end{aligned}$$

Combining these two upper bounds, we have

$$\mathbb{E}_{\widehat{\pi}}\left[\max_{a\in\mathcal{A}}\left(\widehat{Q}^{\pi}_{\lambda,h}-Q^{\pi}_{\lambda,h}\right)^2(s_h,a)\mid s_1\right]\leqslant\frac{48S^3H^3A\kappa^2\log^2(SN)\beta^{\mathcal{H}}(\delta)}{N}+\frac{4CH^5\mathrm{R}^2_{\max}S^2A\cdot\beta^{\mathrm{conc}}(\delta,N)}{N}+S(H^2\mathrm{R}^2_{\max}+\kappa^2\log^2(S))\varepsilon'.$$

Since $\kappa^2\log^2(s)\leqslant\mathrm{R}^2_{\max}$ and $H\geqslant 1$, we conclude the statement. □

Lemma A.31. *For any bounded function $f\colon\mathcal{S}\times\mathcal{A}\to\mathbb{R}_+$, $f(s,a)\leqslant B$ for any policy π and step h on event $\mathcal{E}^{\texttt{RF-RL-Explore}}(\delta,\varepsilon')$ the following holds*

$$\mathbb{E}_{\pi}[f(s_h,a_h)|s_1]\leqslant 2SAH\mathbb{E}_{(s,a)\sim\mu_h}[f(s,a)]+BS\varepsilon'.$$

Proof. Recall $S_{\varepsilon',h}$ be a set of all ε'-significant states (see Definition A.22) at step h. Then we can rewrite this expectation as follows

$$\mathbb{E}_{\pi}[f(s_h,a_h)|s_1]=\sum_{a\in\mathcal{A},s\in S_{\varepsilon',h}}d^{\pi}_h(s,a)f(s,a)+\sum_{a\in\mathcal{A},s\notin S_{\varepsilon',h}}d^{\pi}_h(s,a)f(s,a).$$

For the first sum by Theorem A.23 we have $d^{\pi}_h(s,a)\leqslant 2SAH\mu_h(s,a)$, thus

$$\sum_{a\in\mathcal{A},s\in S_{\varepsilon',h}}d^{\pi}_h(s,a)f(s,a)\leqslant 2SAH\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mu_h(s,a)f(s,a)=2SAH\mathbb{E}_{(s,a)\sim\mu_h}[f(s,a)].$$

For the second sum we apply $f(s,a)\leqslant B$ and the fact that for all states that are not ε'-significant under any policy $d^{\pi}_h(s)\leqslant\varepsilon'$:

$$\sum_{a\in\mathcal{A},s\notin S_{\varepsilon',h}}d^{\pi}_h(s,a)f(s,a)\leqslant B\sum_{a\in\mathcal{A},s\notin S_{\varepsilon',h}}d^{\pi}_h(s,a)=B\sum_{s\notin S_{\varepsilon',h}}d^{\pi}_h(s)\leqslant BS\varepsilon'.$$

□

A.6 Faster Rates for Visitation Entropy

A.6.1 Algorithm description

Let us start from the description of the modified algorithm RegEntGame. It has a similar game-theoretical foundation as it aims at solving the following minimax game

$$\max_{d\in\mathcal{K}_p}\mathcal{H}_{\text{visit}}(d)=\max_{d\in\mathcal{K}_p}\min_{\bar{d}\in\mathcal{K}}\sum_{(h,s,a)}d_h(s,a)\log\frac{1}{\bar{d}_h(s,a)}=\min_{\bar{d}\in\mathcal{K}}\max_{d\in\mathcal{K}_p}\sum_{(h,s,a)}d_h(s,a)\log\frac{1}{\bar{d}_h(s,a)}.$$

As for usual EntGame, there are two players in the game. On the one hand, the min player, or forecaster player, tries to predict which state-action pairs the max player will visit to minimize $\mathrm{KL}(d_h,\bar{d}_h)$. On the other hand, the max player, or sampler player, is rewarded for visiting state-action pairs that the forecaster player did not predict correctly.

We now describe the algorithm RegEntGame for MVEE. In this algorithm, we let a forecaster player and a sampler player compete for T episodes long. Let us first define the two players.

Forecaster-player. The forecaster player remains exactly the same as for usual EntGame algorithm, see the corresponding section in the main text.

Regularized Sampler-player. For the sampler player we exploit strong convexity of visitation entropy. The running time of the sampler player will be divided onto two stages, as it was done in RL-Explore-Ent algorithm.

Exploration phase. Before the start of the game, the sampler-player uses some preprocessing time in order to explore the environment to learn a simple (non-markovian) preliminary exploration policy π^{mix}. This policy is used to construct an accurate enough estimates of transition probabilities. This policy is obtained, as in RF-RL-Explore and RL-Explore-Ent, by learning for each state-action pair (s,ah), a policy that reliably reaches this action pair (s,a) at step h. This can be done by running any regret minimization algorithm for the sparse reward function putting reward one at state s at step h and zero otherwise. The policy π^{mix} is defined as the mixture of the aforementioned policies.

Planning phase. The second phase is starting during the running time of the algorithm. Since RL-Explore-Ent algorithm is essentially reward-free in a sense of working with an arbitrary reward functions.

For each episode t during the game we define the empirical regularized Bellman equations

$$\begin{aligned}\widehat{Q}_h^t(s,a)&=\log\frac{1}{\bar{d}_h^{t+1}(s)}+\widehat{p}_h^t\widehat{V}_{h+1}^t(s,a)\\ \widehat{V}_h^t(s)&=\max_{\pi\in\Delta(\mathcal{A})}\{\pi\widehat{Q}_h^t(s,a)+\mathcal{H}(\pi)\},\end{aligned}\tag{A.19}$$

where $\widehat{V}_{H+1}^t=0$. The sampler player then follows $d^{\pi^{t+1}}$ where π^{t+1} is greedy with respect to the regularized Q-values, that is, $\pi_h^{t+1}(s)\in\arg\max_{\pi\in\Delta_A}\{\pi\widehat{Q}_h^t(s)+\mathcal{H}(\pi)\}$. This choice of sampler player will be clear in the analysis below.

Sampling rule. At each episode t, the policy π^t of the sampler-player is used as a sampling rule to generate a new trajectory.

Decision rule. After T episodes we output a non-Markovian policy $\widehat{\pi}$ defined as the mixture of the policies $\{\pi^t\}_{t\in[T]}$, that is, to obtain a trajectory from $\widehat{\pi}$ we first sample uniformly at random $t \in [T]$ and then follow the policy π^t. Note that the visitation distribution of $\widehat{\pi}$ is exactly the average $d^{\widehat{\pi}} = (1/T)\sum_{t\in[T]} d^{\pi^t}$.

Remark that the stopping rule of `RegEntGame` is deterministic and equals to $\mathfrak{t} = T$. The complete procedure is detailed in Algorithm 7.

Algorithm 7 `RegEntGame`

Input: Number of episodes T, number of exploration episodes N_0, number of transition samples N, prior counts n_0.
# Preliminary exploration
for $(s', h') \in \mathcal{S} \times [H]$ **do**
  Form rewards $r_h(s, a) = \mathbb{1}\{s = s', h = h'\}$.
  Run `EULER` (Zanette and Brunskill 2019) with rewards r_h over N_0 iterates and collect all policies $\Pi_{s',h'}$.
  Modify $\pi \in \Pi_{s',h'} : \ \pi_{h'}(a|s') = 1/A$ for all $a \in \mathcal{A}$.
end for
Construct a uniform mixture policy π^{mix} over all $\{\Pi_{s,h} : (s, h) \in \mathcal{S} \times [H]\}$.
Sample N independent trajectories $\{z_n\}_{n\in[N]}$ using π^{mix} in the original MDP.
Construct from $\{z_n\}_{n\in[N]}$ the estimates $\widehat{p}_h$ as in (A.14).
for $t \in [T]$ **do**
  # Forecaster-player
  Update pseudo counts $\bar{n}_h^{t-1}(s, a)$ and predict $\bar{d}_h^t(s, a)$.
  # Sampler-player
  Compute π^t by regularized planning (A.19) with rewards $\log\left(1/\bar{d}_h^t(s)\right)$ and entropy regularization.
  # Sampling
  for $h \in [H]$ **do**
    Play $a_h^t \sim \pi_h^t(s_h^t)$
    Observe $s_{h+1}^t \sim p_h(s_h^t, a_h^t)$
  end for
  Update counts and transition estimates.
end for
Output $\widehat{\pi}$ the uniform mixture of $\{\pi^t\}_{t\in[T]}$.

A.6.2 Analysis

We first define the regrets of each players obtained by playing T times the games. For the forecaster-player, for any $\bar{d} \in \mathcal{K}$, we define

$$\mathfrak{R}^T_{\mathrm{Fore}}(\bar{d}) \triangleq \sum_{t=1}^{T} \sum_{h,s,a} \widetilde{d}_h^t(s, a)\left(\log \frac{1}{\bar{d}_h^t(s, a)} - \log \frac{1}{\bar{d}_h(s, a)}\right)$$

where $\widetilde{d}_h^t(s,a) \triangleq \mathbb{1}\Big\{(s_h^t, a_h^t) = (s,a)\Big\}$ is a sample from $d_h^{\pi^t}(s,a)$. Similarly for the sampler-player, for any $d \in \mathcal{K}_p$, we define a *regularized regret*

$$\mathfrak{R}_{\mathrm{Samp}}^T(d) \triangleq \sum_{t=1}^T \Bigg(\sum_{h,s,a} \Big[d_h(s,a) - d_h^{\pi^t}(s,a)\Big] \log \frac{1}{\bar{d}_h^t(s,a)} \\ - \sum_{h,s} \Big[d_h(s)\,\mathrm{KL}(\pi(s), \bar{\pi}_h^t(s)) - d_h^{\pi^t}(s)\,\mathrm{KL}(\pi_h^t(s), \bar{\pi}_h^t(s))\Big]\Bigg),$$

where corresponding policies are defined as $\pi_h(a|s) = d_h(s,a)/d_h(s)$ and $\bar{\pi}_h^t(a|s) = \bar{d}_h^t(s,a)/\bar{d}_h^t(s)$ for $d_h(s) = \sum_a d_h(s,a)$ and $\bar{d}_h^t(s) = \sum_a \bar{d}_h^t(s,a)$.

Recall that the visitation distribution of the policy π returned by `RegEntGame` is the average of the visitation distributions of the sampler-player $d_h^{\widehat{\pi}}(s,a) = \widehat{d}_h^T(s,a) \triangleq (1/T)\sum_{t=1}^T d_h^{\pi^t}(s,a)$. We also denote by $\mathring{d}_h^T(s,a) \triangleq (1/T)\sum_{t=1}^T \widetilde{d}^t(s,a)$ the average of the 'sample' visitation distributions.

We now relate the difference between the optimal visitation entropy and the visitation entropy of the outputted policy $\widehat{\pi}$ to the regrets of the two players. Indeed, using $\mathcal{H}(p) = \sum_{i\in[n]} p_i \log(1/q_i) - \mathrm{KL}(p,q)$ for all $(p,q) \in (\Delta_n)^2$ and

$$\begin{aligned}
\mathrm{KL}(d_h^{\pi^\star}, \bar{d}_h^t) &= \sum_{s,a} d_h^{\pi^\star}(s,a) \log\left(\frac{d_h^{\pi^\star}(s,a)}{\bar{d}_h^t(s,a)}\right) \\
&= \sum_{s} d_h^{\pi^\star}(s) \log\left(\frac{d_h^{\pi^\star}(s)}{\bar{d}_h^t(s)}\right) + \sum_{s} d_h^{\pi^\star}(s) \sum_a \pi_h^\star(a|s) \log\left(\frac{\pi_h^\star(a|s)}{\bar{\pi}_h^t(a|s)}\right) \\
&\geqslant \sum_{s} d_h^{\pi^\star}(s)\,\mathrm{KL}(\pi_h^\star(s), \bar{\pi}_h^t(s))\,.
\end{aligned}$$

This inequality could be treated as a strong convexity of visitation entropy with respect to trajectory entropy since $\mathrm{KL}(d_h^{\pi^\star}, \bar{d}_h^t)$ is a *Bregman divergence* with respect to $\mathcal{H}_{\mathrm{visit}}$, and the final average of $\mathrm{KL}(\pi_h^\star(s), \boldsymbol{\pi}_h^t(s))$ is a Bregman divergence with respect to $\mathcal{H}_{\mathrm{traj}}$ (up to linearities).

Applying this inequality, we have

$$\begin{aligned}
T\Big(\mathcal{H}_{\mathrm{visit}}(d^{\pi^\star}) - \mathcal{H}_{\mathrm{visit}}(d^{\widehat{\pi}})\Big) &\leqslant \sum_{t=1}^T \Bigg(\sum_{h,a,s} d_h^{\pi^\star}(s,a) \log \frac{1}{\bar{d}_h^t(s,a)} - \sum_{h,s} d_h^{\pi^\star}(s)\,\mathrm{KL}(\pi_h^\star(s), \bar{\pi}_h^t(s))\Bigg) \\
&\quad - \sum_{t=1}^T \widetilde{d}_h^t(s,a) \log \frac{1}{\mathring{d}_h^T(s,a)} + T\Big(\mathcal{H}_{\mathrm{visit}}(\mathring{d}^T) - \mathcal{H}_{\mathrm{visit}}(\widehat{d}^T)\Big) \\
&\leqslant \mathfrak{R}_{\mathrm{Samp}}^T(d^{\pi^\star}) + \underbrace{\sum_{t=1}^T \sum_{h,s,a} \Big(d_h^{\pi^t}(s,a) - \widetilde{d}_h^t(s,a)\Big) \log \frac{1}{\bar{d}_h^t(s,a)}}_{\mathrm{Bias}_1} \\
&\quad + \mathfrak{R}_{\mathrm{Fore}}^T(\mathring{d}^T) + \underbrace{T\Big(\mathcal{H}_{\mathrm{visit}}(\mathring{d}^T) - \mathcal{H}_{\mathrm{visit}}(\widehat{d}^T)\Big)}_{\mathrm{Bias}_2}\,.
\end{aligned}$$

It remains to upper bound each terms separately in order to obtain a bound on the gap. Notably, only the sampler player result changes in comparison to `EntGame`.

A.6.3 Regret of the Sampler-Player

We start from introducing new notation. Let $\mathcal{M}_t = (\mathcal{S}, \mathcal{A}, \{p_h\}_{h\in[H]}, \{r_h^t\}_{h\in[H]}, s_1)$ be a sequence of entropy-regularized MDPs where reward defined as $r_h^t(s,a) = \log(1/\bar{d}_h^t(s))$. Define $Q_h^{\pi,t}(s,a)$

and $V_h^{\pi,t}(s,a)$ as a action-value and value functions of a policy π on a MDP $\mathcal{M}_t$. Notice that the value-function of initial state in this case could be written as follows (see Appendix A.3)

$$\begin{aligned} V_1^{\pi,t}(s_1) &= \sum_{h,s,a} d_h^\pi(s,a) \log\left(\frac{1}{\bar{d}_h^t(s,a)}\right) - \sum_{h,s} d_h^\pi(s)\,\mathrm{KL}(\pi_h(s), \bar{\pi}_h^t(s)) \\ &= \sum_{h,s,a} d_h^\pi(s,a) \log\left(\frac{1}{\bar{d}_h^t(s)}\right) + \sum_{h,s} d_h^\pi(s)\mathcal{H}(\pi_h(s)), \end{aligned}$$

also see Appendix A.4 for more exposition. Therefore, the regret for the sampler-player could be rewritten in the terms of the regret for this sequence of entropy-regularized MDPs

$$\mathfrak{R}_{\mathrm{Samp}}^T(d^\pi) = \sum_{t=1}^T V_1^{\pi,t}(s_1) - V_1^{\pi^t,t}(s_1).$$

We notice that our approach does not gives a regret minimizer algorithm in a classical sense, however analysis shows us that we can control the sum of policy error with respect to *any* reward function.

Lemma A.32. *Let* $N_0 = \Omega\left(\frac{H^7S^3A\cdot\log^2(T+SA)\cdot L}{\varepsilon}\right)$ *and* $N = \Omega\left(\frac{H^6S^3A\log^2(T+SA)L^3}{\varepsilon}\right)$. *Then with probability at least* $1-\delta/2$, *the regret of the sampler player is bounded as*

$$\mathfrak{R}_{\mathrm{Samp}}^T(d^{\pi^\star}) \leqslant \varepsilon/2 \cdot T$$

after

$$\widetilde{\mathcal{O}}\left(\frac{H^8S^4A}{\varepsilon}\right)$$

episodes of pure exploration.

Proof. From Corollary A.27 under the choice of parameters $\lambda = 1, \kappa = 0$ and reward function $r_h^t(s,a) = \log(1/\bar{d}_h^t(s))$ for each iteration, that is bounded by $\log(T+SA)$, we have that for any reward function the sub-optimality gap is bounded by $\varepsilon/2$. The total number of episodes of pure exploration is equal to $N_0SH + N$. □

A.6.4 Proof of Theorem 2.7

We state the version of this theorem with all prescribed dependencies factors.

Theorem A.33. *Fix some* $\varepsilon > 0$ *and* $\delta \in (0,1)$. *Then for* $n_0 = 1$,

$$N_0 = \Omega\left(\frac{H^7S^3A\cdot L^3}{\varepsilon}\right), \quad N = \Omega\left(\frac{H^6S^3AL^5}{\varepsilon}\right), \quad T = \Omega\left(\frac{H^2SAL^3}{\varepsilon^2} + \frac{H^2S^2A^2L^2}{\varepsilon}\right)$$

with $L = \log(SAH/\delta\varepsilon)$, *the algorithm* `RegEntGame` *is* (ε,δ)*-PAC. Its total sample complexity is equal to* $SH\cdot N_0 + N + T$, *that is,*

$$\mathfrak{t} = \widetilde{\mathcal{O}}\left(\frac{H^2SA}{\varepsilon^2} + \frac{H^8S^4A}{\varepsilon}\right).$$

Proof. We start from writing down the decomposition defined in the beginning of the appendix

$$T(\mathcal{H}_{\mathrm{visit}}(d^{\pi^\star,\mathrm{VE}}) - \mathcal{H}_{\mathrm{visit}}(\widehat{d^\pi})) \leqslant \mathfrak{R}_{\mathrm{Samp}}^T(d^{\pi^\star,\mathrm{VE}}) + \mathfrak{R}_{\mathrm{Fore}}^T(\mathring{d}^T) + \mathrm{Bias}_1 + \mathrm{Bias}_2.$$

By Lemma A.32 with probability at least $1-\delta/2$ it holds

$$\mathfrak{R}^T_{\mathrm{Samp}}(d^{\pi^\star,\mathrm{VE}}) \leqslant \varepsilon T/2.$$

By Lemma A.1

$$\mathfrak{R}^T_{\mathrm{Fore}}(\mathring{d}^T) \leqslant HSA \log\Big(\mathrm{e}(T+1)\Big).$$

By Lemma A.6 with probability at least $1-\delta/2$

$$\mathrm{Bias}_1 + \mathrm{Bias}_2 \leqslant 3\log(SAT)\left(\sqrt{TH\log(4/\delta)} + H\sqrt{SAT\log(3T)}\right).$$

By union bound all these inequalities hold simultaneously with probability at least $1-\delta$. Combining all these bounds we get

$$\begin{aligned} T(\mathcal{H}_{\mathrm{visit}}(d^{\pi^\star,\mathrm{VE}}) - \mathcal{H}_{\mathrm{visit}}(d^{\widehat{\pi}})) \leqslant 3\log(SAT)&\left(\sqrt{TH\log(4/\delta)} + H\sqrt{SAT\log(3T)}\right) \\ &+ HSA\log(\mathrm{e}(T+1)) + \varepsilon T/2. \end{aligned}$$

Therefore, it is enough to choose T such that $\mathcal{H}_{\mathrm{visit}}(d^{\pi^\star,\mathrm{VE}}) - \mathcal{H}_{\mathrm{visit}}(d^{\widehat{\pi}})$ is guaranteed to be less than ε. In this case `RegEntGame` become automatically (ε,δ)-PAC. It is equivalent to find a maximal T such that

$$\varepsilon T/2 \leqslant 3\log(SAT)\left(\sqrt{TH\log(4/\delta)} + H\sqrt{SAT\log(3T)}\right) + HSA\log(\mathrm{e}(T+1))).$$

and add 1 to it. We start from obtaining a loose bound to eliminate logarithmic factors in T.

First, we assume that $T \geqslant 1$, thus $T+1 \leqslant 2T$. Additionally, let us use inequality $\log(x) \leqslant x^\beta/\beta$ for any $x > 0$ and $\beta > 0$. We obtain

$$\varepsilon T \leqslant 48(SAT)^{1/8}\left(\sqrt{T^{3/2}H\log(4/\delta)} + H\sqrt{4SAT^{3/2}}\right) + 16/7 \cdot HSA(2\mathrm{e}T)^{7/8}$$

that could be relaxed as follows

$$\varepsilon T^{1/8} \leqslant 48(SA)^{1/8}(H\log(4/\delta)^{1/2} + H(SA)^{5/8} + 11HSA$$

thus we can define $\gamma = 8\log\Big((48(SA)^{1/8}(H\log(4/\delta)^{1/2} + H(SA)^{5/8} + 11HSA)/\varepsilon\Big) = \mathcal{O}(L)$ for which $\log(T) \leqslant \gamma$. Therefore

$$\varepsilon T/2 \leqslant 3(\log(SA)+\gamma)\sqrt{T}\left(\sqrt{H\log(4/\delta)} + \sqrt{SAH^2(\log(3)+L)}\right) + HSA(1+2\gamma).$$

Solving this quadratic inequality, we obtain the minimal required T to guarantee $\mathcal{H}_{\mathrm{visit}}(d^{\pi^\star,\mathrm{VE}}) - \mathcal{H}_{\mathrm{visit}}(d^{\widehat{\pi}}) \leqslant \varepsilon$. In particular,

$$T = \Omega\left(\frac{H^2SAL^3}{\varepsilon^2} + \frac{H^2S^2A^2L^2}{\varepsilon}\right).$$

□

A.7 Additional Experiments

In this section we provide details on experiments and additional experiments.

First, we describe our baselines in details.

- Random policy $\pi_h(a|s) = 1/A$ for any $h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$;
- Optimal MVEE policy. First we compute the solution to the following convex program

$$\max_{d \in \mathcal{K}_p} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} -\left(\frac{1}{H}\sum_{h=1}^{H} d_h(s,a)\right) \log\left(\frac{1}{H}\sum_{h=1}^{H} d_h(s,a)\right),$$

  where $\mathcal{K}_p$ is a polytope of admissible visitation distributions defined in Section 2.2. This problem exactly corresponds to the setting visitation entropy by Hazan et al. (2019). Then we compute the optimal MVEE policy by normalization of the visitation distribution $\pi_h(a|s) = \frac{d_h(s,a)}{\sum_{a \in \mathcal{A}} d_h(s,a)}$.
- Optimal MTEE policy that was computed by solving regularized Bellman equations with $\lambda = 1$ and the entropy of transition kernels as rewards.

The last two baselines requires the knowledge of the true transition kernel $p_h(s'|s,a)$. Additionally, we perform experiments with `RF-UCRL` algorithm by Kaufmann et al. (2021) but the final visitation numbers were very close to the visitation numbers of `EntGame` algorithm and we do not report them. The data collection procedure for `EntGame` and `UCBVI-Ent` is following.

- For `EntGame` algorithm we report visitation counts of states of the algorithm during the learning. It is an appropriate choice since this counts corresponds to the empirical density $\mathring{d}_h^T(s,a) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{(s_h^t, a_h^t) = (s,a)\}$ which converges to $d_h^{\widehat{\pi}}(s,a) = \frac{1}{T}\sum_{t=1}^{T} d_h^{\pi^t}(s,a)$.
- For `UCBVI-Ent` algorithm we have to separate stages. At the first stage the algorithm interacts with environment to learn the final policy during N interactions with the environment, and all the plots present the visitations counts only for the final policy during another N interactions when the policy is fixed.

Remark A.34. We do not compare `EntGame` with `MaxEnt` and `Toc-UCRL2` because of their similarity. The only difference in the these algorithms is the way how density estimation is performed. More precisely, in the `MaxEnt` algorithm the average of state-action visitation distributions of all the policies played is estimated from scratch at the end of each epoch with additional trajectories. In `EntGame` we also estimate this average but with all trajectories collected until now and without need of additional fresh trajectories. Thus, `EntGame` can be seen as a version of `MaxEnt` that reuses past trajectories instead of collecting new ones. The latter feature is helpful in practice but will not change the rate. That is why we decided not to include `MaxEnt` in the experiments. Similar comments hold for `Toc-UCRL2`.

Double Chain. The experiment described in Section 2.6 was preformed on Double Chain environment described by Kaufmann et al. (2021). This MDP consists of states $\mathcal{S} = \{0, \dots, L-1\}$, where L is the length of the chain, the actions $\mathcal{A} = \{l, r\}$ which corresponds to the transition to the left (action l) or to the right (action r). Additionally, while taking the actions, there is 10% probability of moving to the opposite direction. The agent starts at the middle of the chain $s_1 = (L-1)/2$. For experiments we run each algorithm to collect $N = 100000$ samples during episodes of horizon $H = 20$ and then report the average and confidence intervals over 48 random seeds.

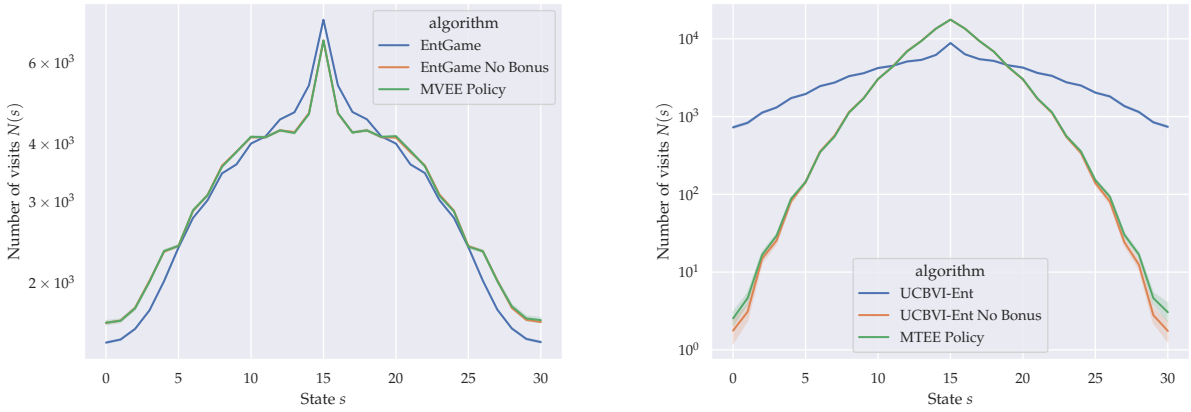

Figure A.1: Number of state visits for $N = 100000$ samples in the Double Chain MDP for `EntGame` and `UCBVI-Ent` algorithms with and without bonuses.

In Section 2.6 we conclude that the exploration bonuses have the important role in the state visitations of `UCBVI-Ent` algorithm and make it close to `RF-UCRL` by Kaufmann et al. (2021). As an ablation study we preform the same experiments for algorithms without bonuses. The results are presented in Figure A.1.

In particular, we see that `EntGame` algorithm without bonuses converges to the optimal MVEE policy in terms of the visitation densities that is slightly more spread than `EntGame` algorithm with bonuses. Notice that algorithm has its own exploration mechanism since the rewards are equal to $\log((t + SA)/(n^t(s, a) + 1))$ that is tightly connected to exploration bonuses. `UCBVI-Ent` algorithm also has its own exploration method that induced by soft-max policies. In particular, soft-max policies are tightly connected to the Boltzmann exploration with step-size equal to 1 (Sutton 1990). However, it is well-known (Cesa-Bianchi et al. 2017) that the Boltzmann exploration with fixed step-size did not provide efficient way to explore.

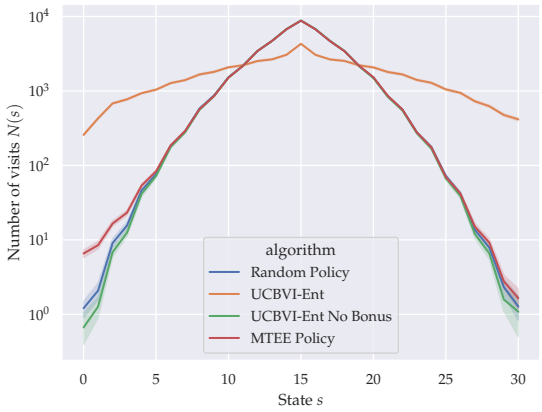

Figure A.2: Number of state visits for $N = 50000$ samples in the Double Chain MDP with resampling for the `UCBVI-Ent` algorithm with and without bonuses.

Double Chain with Resampling. To test the exploration mechanism of `UCBVI-Ent` without bonuses, we slightly modify Double Chain environment: now the visitation of the left end of the

chain leads to uniform resampling over all states. The result is presented in Figure A.2.

In particular, we observe that in this situation UCBVI-Ent algorithm without bonuses still acts like a random policy due to low exploration for the ends of the chain, whereas the optimal MTEE policy visits the left part of the chain more often. This observation imply that the additional exploration mechanism for UCBVI-Ent algorithm is required and the exploration problem in regularized MDPs is non-trivial.

GridWorld. As an additional experiment to verify our findings we perform additional experiments on the environment Grid World as it presented in Figure A.3. The state space is a set of discrete points in a 21×21 grid. For each state there are 4 possible actions: left, right, up or down, and for each action there is a 5% probability to move to the wrong direction. The initial state s_1 is the middle of the grid. For this experiment we use $N = 60000$ samples and report the average over 12 random seeds.

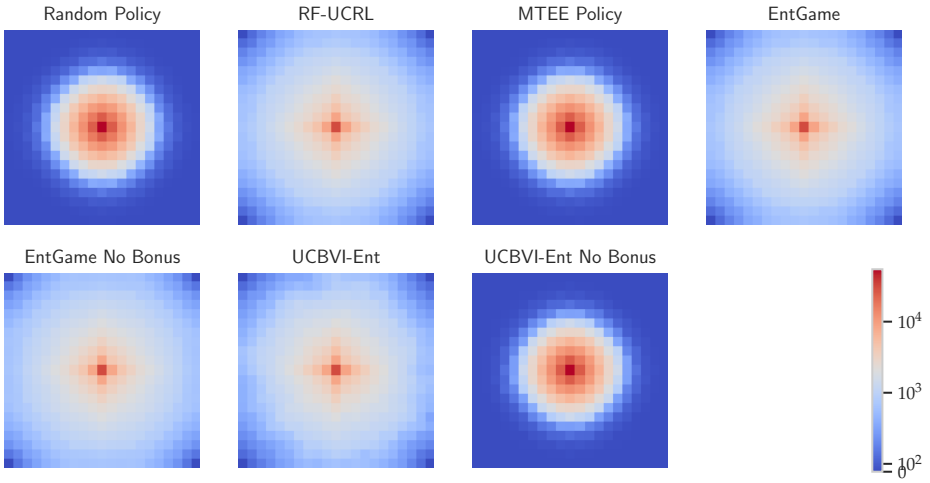

Figure A.3: Number of state visits for $N = 60000$ samples in the GridWorld MDP for EntGame and UCBVI-Ent algorithms with and without bonuses.

Here we see the similar effect as on simpler environment: the UCBVI-Ent algorithm produces slightly less "spread" policy than EntGame or RF-UCRL algorithms. It is connected to the fact that in this case the limiting MTEE policy of UCBVI-Ent is again almost uniform policy due to near-deterministic structure of the MDP. Additionally, we remark that in this case the optimal MVEE policy is much harder to compute due to numerical issues, however, the EntGame algorithm without bonuses produces slightly more uniform distribution that coincides with the effect we observe during experiments with a Double Chain environment.

Appendix B

Complements on Chapter 3

Contents

B.1 Notation

For any $(s,a) \in \mathcal{S}$, transition kernel $p(s,a) \in \mathcal{P}(\mathcal{S})$ and $f\colon \mathcal{S} \to \mathbb{R}$, define $pf(s,a) = \mathbb{E}_{p(s,a)}[f]$ and $\mathrm{Var}_p[f](s,a) = \mathrm{Var}_{p(s,a)}[f]$. For any $s \in \mathcal{S}$, policy $\pi(s) \in \mathcal{P}(\mathcal{S})$ and $f\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, set $\pi f(s) = \mathbb{E}_{a\sim\pi(s)}[f(s,a)]$ and $\mathrm{Var}_\pi f(s) = \mathrm{Var}_{a\sim\pi(s)}[f(s,a)]$. For a MDP $\mathcal{M}$, a policy π and a sequence of function $(f_h,\, h \in [H])$, define $\mathbb{E}_\pi[\sum_{h'=h}^H f(s_{h'},a_{h'})|s_h]$ as a conditional expectation of $\sum_{h'=h}^H f(s_{h'},a_{h'})$ with respect to the sigma-algebra $\mathcal{F}_h = \sigma\{(s_{h'},a_{h'})|h' \leqslant h\}$, where for any $h \in [H]$, we have $a_h \sim \pi(s_h)$, $s_{h+1} \sim p_h(s_h,a_h)$.

Table B.1: Table of notation use throughout the paper

Notation	**Meaning**
$\mathcal{S}$	state space of size S
$\mathcal{A}$	action space of size A
d	dimension of linear MDP
H	length of one episode
s_1	initial state
$\mathfrak{t}$	stopping time
$\mathcal{T}$	trajectory space, $\mathcal{T} \triangleq (\mathcal{S} \times \mathcal{A})^H$
ε	desired accuracy of solving the problem
δ	desired upper bound on failure probability
$p_h(s'\|s,a)$	probability transition
$r_h(s,a)$	reward function
$V_h^\pi, V_h^\star$	value of policy π and optimal value
$Q_h^\pi, Q_h^\star$	Q-value of policy π and optimal Q-value
$V_{\widetilde{\pi},\lambda,h}^{\widetilde{\pi}}, V_{\widetilde{\pi},\lambda,h}^{\star}$	regularized value of policy π and optimal regularized value
$Q_{\widetilde{\pi},\lambda,h}^{\widetilde{\pi}}, Q_{\widetilde{\pi},\lambda,h}^{\star}$	regularized Q-value of policy π and optimal regularized Q-value
Π, Π_h	space of all policies and space of policies on step h
$\Pi_\gamma, \Pi_{h,\gamma}$	space of all policies and policies on step h with minimal probability γ
$\mathcal{D}_{\mathrm{E}}$	expert dataset of size N^{E}: $\mathcal{D}_{\mathrm{E}} \triangleq \{\mathring{\tau}_i = (s_1^i, a_1^i, \ldots, s_H^i, a_H^i),\ i \in [N^{\mathrm{E}}]\}$
π^{E}	expert policy
$\pi^{\mathrm{E},\kappa}$	κ-greedy version of the expert policy
ε_{E}	sub-optimality gap of the expert policy: $V_1^\star(s_1) - V_1^{\pi^{\mathrm{E}}}(s_1) \leqslant \varepsilon_{\mathrm{E}}$
π^{BC}	behavior cloning policy
$\mathcal{R}_h$	regularizer for behavior cloning
$\mathcal{F}$	class of policies for behavior cloning
$d_{\mathcal{F}}$	covering dimension of one-step policy class for behavior cloning
s_h^t	state that was visited at h step during t episode
a_h^t	action that was picked at h step during t episode
$r^\star$	true reward function in a preference-based model
σ	link function, see Assumption 4
ζ	linearity measure of link function, see Assumption 4
π^{S}	sampling policy for generation preference dataset
$\mathcal{D}_{\mathrm{RM}}$	preference dataset of size N^{RM} : $\mathcal{D}_{\mathrm{RM}} \triangleq \{(\tau_0^k, \tau_1^k, o^k)\}$
$\mathcal{G}$	class of trajectory rewards for reward modeling
$d_{\mathcal{G}}$	bracketing dimension of the induced preference models
π^{RL}	policy for BPI with demonstration
π^{RLHF}	policy for preference-based BPI with demonstration
$\mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta)$	sample complexity for BPI with demonstration
$\mathcal{C}(\varepsilon, \lambda, \delta)$	sample complexity for regularized BPI

We define trajectory KL-divergence between two policies $\pi = \{\pi_h\}_{h\in[H]}, \pi' = \{\pi_h\}_{h\in[H]}$ as follows

$$\mathrm{KL}_{\mathrm{traj}}(\pi, \pi') = \mathbb{E}_\pi \left[\sum_{h=1}^{H} \mathrm{KL}(\pi_h(s_h), \pi'_h(s_h)) \right].$$

We write $f(S, A, H, \varepsilon) = \mathcal{O}(g(S, A, H, \varepsilon, \delta))$ if there exist $S_0, A_0, H_0, \varepsilon_0, \delta_0$ and constant $C_{f,g}$ such that for any $S \geqslant S_0, A \geqslant A_0, H \geqslant H_0, \varepsilon < \varepsilon_0, \delta < \delta_0, f(S, A, H, T, \delta) \leqslant C_{f,g} \cdot g(S, A, H, T, \delta)$. We write $f(S, A, H, \varepsilon, \delta) = \widetilde{\mathcal{O}}(g(S, A, H, \varepsilon, \delta))$ if $C_{f,g}$ in the previous definition is poly-logarithmic in $S, A, H, 1/\varepsilon, 1/\delta$.

Coverings, packings, and bracketings. A pair $(\mathcal{X}, \rho)$ is called pseudometric space with a metric $\rho\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ if ρ satisfies $\rho(x, x) = 0$ for all $x \in \mathcal{X}$, ρ is symmetric, that is, $\forall x, y \in \mathcal{X} : \rho(x, y) = \rho(y, x)$, and ρ satisfies triangle inequality $\forall x, y, z : \rho(x, y) + \rho(y, z) \geqslant \rho(x, z)$.

Definition B.1 (ε-covering and packing)**.** Let $(\mathcal{X}, \rho)$ be a (pseudo)metric space with a metric $\rho\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. The ε-covering number $\mathcal{N}(\varepsilon, \mathcal{X}, \rho)$ is the size of the minimal ε-cover of $(\mathcal{X}, \rho)$, that is,

$$\mathcal{N}(\varepsilon, \mathcal{X}, \rho) = \min_{X \subseteq \mathcal{X}} \{|X| : \forall y \in \mathcal{X}\ \exists x \in X : \rho(y, x) \leqslant \varepsilon\}\,.$$

The ε-packing number $\mathcal{P}(\varepsilon, \mathcal{X}, \rho)$ is the size of the maximal ε-separated set of $(\mathcal{X}, \rho)$,

$$\mathcal{P}(\varepsilon, \mathcal{X}, \rho) = \max_{X \subseteq \mathcal{X}} \{|X| : \forall x \neq y \in X : \rho(x, y) > \varepsilon\}\,.$$

Definition B.2 (ε-bracketing)**.** Let $\mathcal{F}\colon \mathcal{X} \to \mathbb{R}$ be a function class endowed with a norm $\|\cdot\|$. Given two functions $\ell, u\colon \mathcal{X} \to \mathbb{R}$, a bracket $[\ell, u]$ is a set of all functions $f \in \mathcal{F}$ such that $\ell(x) \leqslant f(x) \leqslant u(x)$ for all $x \in \mathcal{X}$. A ε-bracket is a bracket $[\ell, u]$ such that $\|\ell - u\| \leqslant \varepsilon$. The ε-bracketing number $\mathcal{N}_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the cardinality of the minimal set of ε-brackets needed to cover $\mathcal{F}$,

$$\mathcal{N}_{[]}(\mathcal{F}, \|\cdot\|) = \min_{N} \{|N| \mid \forall f \in \mathcal{F}\ \exists [\ell, u] \in N : \ell(x) \leqslant f(x) \leqslant u(x) \forall x \in \mathcal{X}, \|\ell - u\| \leqslant \varepsilon\}\,.$$

B.2 Behavior Cloning

In this appendix, we gather the proofs of the results for behavior cloning presented in Section 3.3.

B.2.1 Proof for General setting

In this appendix, we provide the proof of Theorem 3.3.

Theorem (Restatement of Theorem 3.3)**.** *Assume Assumptions 1-2 and that* $0 \leqslant \mathcal{R}_h(\pi_h) \leqslant M$ *for all* $h \in [H]$*, for any policy* $\pi \in \mathcal{F}_h$*. Let* π^{BC} *be a solution to (3.1). Then with probability at least* $1-\delta$ *the behavior policy* π^{BC} *satisfies*

$$\mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}}\|\pi^{\mathrm{BC}}) \leqslant \frac{6d_{\mathcal{F}}H\cdot(\log(Ae^3/(A\gamma\wedge\kappa))\cdot\log(2HN^{\mathrm{E}}R_{\mathcal{F}}/(\gamma\delta))}{N^{\mathrm{E}}} + \frac{2HM}{N^{\mathrm{E}}} + \frac{18\kappa}{1-\kappa}\,.$$

Proof. We commence by defining the one-step trajectory KL-divergence as follows:

$$\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi_h^{\mathrm{BC}}) = \mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E}}(a_h|s_h)}{\pi_h^{\mathrm{BC}}(a_h|s_h)}\right)\right].$$

In particular, by the linearity of expectation, the following holds

$$\mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}}\|\pi^{\mathrm{BC}}) = \sum_{h=1}^{H}\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi_h^{\mathrm{BC}})\,.$$

Recall the definition of the κ-greedy version of the expert policy

$$\pi_h^{\mathrm{E},\kappa}(a|s) = (1-\kappa)\pi_h^{\mathrm{E}}(a|s) + \frac{\kappa}{A} = (1-\kappa)\cdot\left(\pi_h^{\mathrm{E}}(a|s) + \frac{\kappa}{(1-\kappa)A}\right).$$

Next, we can decompose the one-step trajectory KL-divergence as follows

$$\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi_h^{\mathrm{BC}}) = \mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi_h^{\mathrm{BC}}(a_h|s_h)}\right)\right] + \mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E}}(a_h|s_h)}{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}\right)\right].$$

For the second term, we have

$$\mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E}}(a_h|s_h)}{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}\right)\right] = \mathbb{E}_{\pi^{\mathrm{E}}}\left[\underbrace{\log\left(\frac{\pi_h^{\mathrm{E}}(a_h|s_h)}{\pi_h^{\mathrm{E}}(a_h|s_h)+\kappa/(A(1-\kappa))}\right)}_{\leqslant 0}\right] - \log(1-\kappa) \leqslant \frac{\kappa}{1-\kappa}\,,$$

where the last inequality follows from the fact that $(1-x)\log(1-x) \geqslant -x$ for any $x < 1$, by convexity of the function $x \mapsto x\log x$. Next, we decompose the smoothed version of the one-step trajectory KL to the sum of stochastic and empirical terms,

$$\begin{aligned}\mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi_h^{\mathrm{BC}}(a_h|s_h)}\right)\right] &= \frac{1}{N^{\mathrm{E}}}\sum_{t=1}^{N^{\mathrm{E}}}\left(\mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi_h^{\mathrm{BC}}(a_h|s_h)}\right)\right] - \log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h^t|s_h^t)}{\pi_h^{\mathrm{BC}}(a_h^t|s_h^t)}\right)\right)\\ &\quad + \frac{1}{N^{\mathrm{E}}}\sum_{t=1}^{N^{\mathrm{E}}}\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h^t|s_h^t)}{\pi_h^{\mathrm{BC}}(a_h^t|s_h^t)}\right).\end{aligned}$$

To upper bound the first term we apply Lemma B.5 and obtain with probability at least $1-\delta$

$$\frac{1}{N^{\mathrm{E}}}\sum_{t=1}^{N^{\mathrm{E}}}\left(\mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi_h^{\mathrm{BC}}(a_h|s_h)}\right)\right]-\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h^t|s_h^t)}{\pi_h^{\mathrm{BC}}(a_h^t|s_h^t)}\right)\right)$$
$$\leqslant\sqrt{\frac{2\log(\mathrm{e}^2/\gamma)\,\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi_h^{\mathrm{BC}})\cdot d(\log(2N^{\mathrm{E}}R_{\mathcal{F}}/\gamma)+\log(1/\delta))}{N^{\mathrm{E}}}}$$
$$+\frac{5(\log(A\mathrm{e}^3/(A\gamma\wedge\kappa))\cdot d_{\mathcal{F}}(\log(2N^{\mathrm{E}}R_{\mathcal{F}}/\gamma)+\log(1/\delta)))}{3N^{\mathrm{E}}}+\frac{8\kappa}{1-\kappa}\,.$$

To control the second term, we first notice that since $\mathcal{F}$ has a product structure, then by a simple observation

$$\{\pi_h\}_{h=1}^H=\operatorname*{arg\,min}_{\pi_1\in\mathcal{F}_1,\ldots,\pi_H\in\mathcal{F}_H}\sum_{h=1}^H\mathcal{L}_h(\pi_h)\iff\forall h\in[H]:\pi_h=\operatorname*{arg\,min}_{\pi_h\in\mathcal{F}_h}\mathcal{L}_h(\pi_h)$$

for any functions $\{\mathcal{L}_h\}_{h=1}^H$, the MLE estimation (3.1) implies

$$\pi_h^{\mathrm{BC}}\in\operatorname*{arg\,min}_{\pi_h\in\mathcal{F}_h}\sum_{i=1}^{N^{\mathrm{E}}}\log\frac{1}{\pi_h(a_h^i|s_h^i)}+\mathcal{R}_h(\pi_h)\,,$$

therefore the following holds

$$\sum_{t=1}^{N^{\mathrm{E}}}\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h^t|s_h^t)}{\pi_h^{\mathrm{BC}}(a_h^t|s_h^t)}\right)\leqslant M\quad+\left\{\sum_{t=1}^{N^{\mathrm{E}}}\log\left(\frac{1}{\pi_h^{\mathrm{BC}}(a_h^t|s_h^t)}\right)+\mathcal{R}_h(\pi_h^{\mathrm{BC}})\right\}$$
$$-\left\{\sum_{t=1}^{N^{\mathrm{E}}}\log\left(\frac{1}{\pi_h^{\mathrm{E},\kappa}(a_h^t|s_h^t)}\right)+\mathcal{R}_h(\pi_h^{\mathrm{E},\kappa})\right\}\leqslant M\,.$$

Thus, we have

$$\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi_h^{\mathrm{BC}})\leqslant\sqrt{\frac{2\log(\mathrm{e}^2/\gamma)\,\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi_h^{\mathrm{BC}})\cdot d_{\mathcal{F}}(\log(2N^{\mathrm{E}}R_{\mathcal{F}}/\gamma)+\log(1/\delta))}{N^{\mathrm{E}}}}$$
$$+\frac{5(\log(A\mathrm{e}^3/(A\gamma\wedge\kappa))\cdot d_{\mathcal{F}}(\log(2N^{\mathrm{E}}R_{\mathcal{F}}/\gamma)+\log(1/\delta)))}{3N^{\mathrm{E}}}+\frac{M}{N^{\mathrm{E}}}+\frac{9\kappa}{1-\kappa}\,.$$

This means that $\sqrt{\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi_h^{\mathrm{BC}})}$ satisfies a quadratic inequality of the form $x^2\leqslant ax+b$. Since $ax\leqslant(a^2+x^2)/2$, we further have $x^2\leqslant a^2+2b$. As a result

$$\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi_h^{\mathrm{BC}})\leqslant\frac{6d_{\mathcal{F}}\log(A\mathrm{e}^3/(A\gamma\wedge\kappa))\cdot\log(2N^{\mathrm{E}}R_{\mathcal{F}}/(\gamma\delta))}{N^{\mathrm{E}}}+\frac{2M}{N^{\mathrm{E}}}+\frac{18\kappa}{1-\kappa}\,.$$

To conclude the statement, we apply a union bound over $h\in[H]$ and sum over the final upper bound. □

B.2.2 Proofs for Finite setting

We recall that for finite MDPs we chose a logarithmic regularizer $\mathcal{R}_h(\pi_h)=\sum_{s,a}\log(1/\pi_h(a|s))$ and the policy class $\mathcal{F}=\{\pi\in\Pi:\pi_h(a|s)\geqslant 1/(N^{\mathrm{E}}+A)\}$. One can check that Assumptions 1-2 holds for these choices and that $0\leqslant\mathcal{R}_h(\pi_h)\leqslant SA\log(A)$. Then we can apply Theorem 3.3 to obtain the following bound for finite MDPs.

Corollary (Restatement of Corollary 3.4)**.** *For all $N^{\mathrm{E}} \geqslant A$, for function class $\mathcal{F}$ and regularizer $(\mathcal{R}_h)_{h\in[H]}$ defined above, with probability at least $1-\delta$,*

$$\mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}}\|\pi^{\mathrm{BC}}) \leqslant \frac{6SAH\cdot\log(2\mathrm{e}^4 N^{\mathrm{E}})\cdot\log(12H(N^{\mathrm{E}})^2/\delta)}{N^{\mathrm{E}}} + \frac{18AH}{N^{\mathrm{E}}}\,.$$

Proof. Let us start by observing that $\mathcal{F}_h \subseteq \Delta(\mathcal{A})^{\mathcal{S}}$ is a subset of a unit ball in ℓ_∞-norm. Therefore by a standard result in the covering numbers for finite-dimensional Banach spaces (see i.e. Problem 5.5 by Handel 2016)

$$\log\mathcal{N}(\varepsilon,\mathcal{F}_h,\|\cdot\|_\infty) \leqslant SA\log(3/\varepsilon)\,.$$

Thus, the parametric classes $\{\mathcal{F}_h\}_{h\in[H]}$ satisfies Assumption 1 with constants $d_{\mathcal{F}} = SA, R_{\mathcal{F}} = 3, \gamma = 1/(N^{\mathrm{E}}+A)$. Next, we notice that for *any* expert policy, Assumption 2 is satisfied with $\kappa = A/(N^{\mathrm{E}}+A)$ for this parametric family. Thus, we can apply Theorem 3.3 and get

$$\begin{aligned}\mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}}\|\pi^{\mathrm{BC}}) &\leqslant \frac{6SAH\cdot\log((N^{\mathrm{E}}+A)\mathrm{e}^3)\cdot\log(6HN^{\mathrm{E}}(N^{\mathrm{E}}+A)/\delta)}{N^{\mathrm{E}}}\\ &\quad+\frac{2SAH\log(A)}{N^{\mathrm{E}}}+\frac{18AH}{N^{\mathrm{E}}}\,.\end{aligned}$$

By upper bounding the first and the second terms under the assumption $N^{\mathrm{E}} \geqslant A$ we conclude the statement. □

B.2.3 Proofs for Linear setting

We start from a natural example when Assumption 3 is fulfilled, and the sub-optimality error ε_{E} is small.

Lemma B.3. *Assume that the MDP $\mathcal{M}$ is linear (see Definition 3.2) and consider the regularized MDP with uniform policy $\widetilde{\pi}(a|s) = \mathcal{U}\mathrm{nif}[A]$ and with a coefficient λ (see Appendix B.5 for more exposition). Then the optimal regularied policy $\pi^\star_{\widetilde{\pi},\lambda,h}$ satisfies Assumption 3 with a constant $R = H\sqrt{d}/\lambda$. Moreover, this policy is $\lambda H\log(A)$-optimal.*

Proof. At first, by Proposition B.22, it holds that for an optimal policy $\pi^\star_{\widetilde{\pi},\lambda,h}$ there are some weights $w_h^\star$ such that $Q_h^\star(s,a) = \langle\psi(s,a), w_h^\star\rangle$ and, moreover, $\|w_h^\star\| \leqslant H\sqrt{d}$.

Then we notice that from the regularized Bellman equations, it holds

$$\begin{aligned}\pi^\star_{\widetilde{\pi},\lambda,h}(a|s) &= \arg\max_\pi\Big\{\pi Q^\star_{\widetilde{\pi},\lambda,h}(s) - \lambda\,\mathrm{KL}(\pi\|\mathcal{U}\mathrm{nif}[A])\Big\}\\ &= \arg\max_\pi\left\{\pi\left[\frac{1}{\lambda}Q^\star_{\widetilde{\pi},\lambda,h}\right](s) - \mathrm{KL}(\pi\|\mathcal{U}\mathrm{nif}[A])\right\} = \frac{\exp\Big(\langle\psi(s,a),\frac{1}{\lambda}w_h^\star\rangle\Big)}{Z(s)}\,.\end{aligned}$$

Therefore, Assumption 3 is satisfied with $R = H\sqrt{d}/\lambda$ for $\pi^{\mathrm{E}} = \pi^\star_\lambda$.

To verify the suboptimality of this policy, we notice that $\pi^\star_{\widetilde{\pi},\lambda}$ satisfies

$$\pi^\star_{\widetilde{\pi},\lambda} = \arg\max_{\pi\in\Pi}\{V_1^\pi(s_1) - \lambda\,\mathrm{KL}_{\mathrm{traj}}(\pi\|\widetilde{\pi})\}\,,$$

therefore

$$\begin{aligned}V_1^\star(s_1) - \lambda\,\mathrm{KL}_{\mathrm{traj}}(\pi^\star\|\widetilde{\pi})\} &\leqslant V_1^{\pi^\star_{\widetilde{\pi},\lambda}}(s_1) - \lambda\,\mathrm{KL}_{\mathrm{traj}}(\pi^\star_{\widetilde{\pi},\lambda}\|\mathcal{U}\mathrm{nif}[A])\\ &\Rightarrow V_1^\star(s_1) - V_1^{\pi^\star_{\widetilde{\pi},\lambda}}(s_1) \leqslant \lambda H\log(A)\,.\end{aligned}$$

□

Next, we provide the result for linear MDPs under Assumption 3, using the parametric assumption given in (3.2).

Corollary (Restatement of Corollary 3.7)**.** *Under Assumption 3, for all $N^{\mathrm{E}} \geqslant A$, for the function class $\mathcal{F}$ defined in (3.2) and regularizer $\mathcal{R}_h = 0$, for all $h \in [H]$, with probability at least $1-\delta$,*

$$\mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}} \| \pi^{\mathrm{BC}}) \leqslant \frac{6dH \cdot \log(2\mathrm{e}^3 N^{\mathrm{E}}) \cdot \log(48H(N^{\mathrm{E}})^2 R/\delta)}{N^{\mathrm{E}}} + \frac{18AH}{N^{\mathrm{E}}} \,.$$

Proof. We start by checking that Assumption 1 holds. By construction of $\mathcal{F}_h$ in (3.2), we have

$$\inf_{\pi_h \in \mathcal{F}_h} \inf_{(s,a) \in \mathcal{S} \times \mathcal{A}} \pi_h(a|s) \geqslant \frac{1}{N^{\mathrm{E}} + A} \,.$$

Next, we have to consider the covering dimension of the hypothesis set. First, we notice that for any two policies $\pi_h, \mu_h \in \mathcal{F}_h$ we have

$$|\pi_h(a|s) - \mu_h(a|s)| = |(1-\kappa)\pi'_h(a|s) - (1-\kappa)\mu'_h(a|s)| = (1-\kappa)|\pi'_h(a|s) - \mu'_h(a|s)| \,,$$

where $\pi'_h, \mu'_h \in \mathcal{F}'_h$ for $\mathcal{F}'_h$ defined as follows

$$\mathcal{F}'_h = \left\{ \pi_h(a|s) = \frac{\exp(\psi(s,a)^\intercal w_h)}{\sum_{a' \in \mathcal{A}} \exp(\psi(s,a')^\intercal w_h)} : \|w_h\|_2 \leqslant R \right\}.$$

Thus, it is sufficient to compute the covering number for $\mathcal{F}'_h$. Let us define

$$\Phi(a|s, w_h) = \exp\{\langle \psi(s,a), w_h \rangle\}, \quad Z(s, w_h) = \sum_{a \in \mathcal{A}} \Phi(a|s, w_h) \,.$$

Then, let w_h, w'_h be weight vectors corresponding to π'_h and μ'_h, respectively. Then

$$\begin{aligned}
|\pi'_h(a|s) - \mu'_h(a|s)| &= \left| \frac{\Phi(a|s,w_h)}{Z(s,w_h)} - \frac{\Phi(a|s,w'_h)}{Z(s,w'_h)} \right| \\
&= \left| \frac{\Phi(a|s,w_h) - \Phi(a|s,w'_h)}{Z(s,w_h)} - \Phi(a|s,w'_h) \left[\frac{1}{Z(s,w'_h)} - \frac{1}{Z(s,w_h)} \right] \right| \\
&\leqslant \frac{\Phi(a|s,w_h)}{Z(s,w_h)} \left| 1 - \frac{\Phi(a|s,w'_h)}{\Phi(a|s,w_h)} \right| + \frac{\Phi(a|s,w'_h)}{Z(s,w'_h)} \left| 1 - \frac{Z(s,w'_h)}{Z(s,w_h)} \right| .
\end{aligned}$$

Next, we analyze both terms separately. For the first term, we have

$$\left| 1 - \frac{\Phi(a|s,w'_h)}{\Phi(a|s,w_h)} \right| = |1 - \exp\{\langle \psi(s,a), w'_h - w_h \rangle\}| \,.$$

We notice that the absolute value of the expression under exponent is upper-bounded by $\|w_h - w'_h\|_2$. Let us assume that $\|w_h - w'_h\|_2 \leqslant 1$, then by the inequality $|1 - e^x| \leqslant 2|x|$ for any $|x| \leqslant 1$, we have

$$\left| 1 - \frac{\Phi(a|s,w'_h)}{\Phi(a|s,w_h)} \right| \leqslant 2\|w_h - w'_h\|_2 \,. \tag{B.1}$$

For the second term, we have by the definition of the normalization constant

$$\begin{aligned}
\left| 1 - \frac{Z(s,w'_h)}{Z(s,w_h)} \right| &= \left| \frac{\sum_{a'} \Phi(a'|s,w_h)[1 - \Phi(a'|s,w'_h)/\Phi(a'|s,w_h)]}{Z(s,w_h)} \right| \\
&\leqslant \frac{\sum_{a'} \Phi(a'|s,w_h) \cdot |1 - \Phi(a'|s,w'_h)/\Phi(a'|s,w_h)|}{Z(s,w_h)} \leqslant 2\|w_h - w'_h\|_2 \,,
\end{aligned}$$

where in the end we applied (B.1). Finally, we have, for any policies π'_h and μ'_h such that the corresponding weights w_h and w'_h satisfies $\|w_h - w'_h\|_2 \leqslant 1$, that

$$|\pi_h(a|s) - \mu_h(a|s)| \leqslant |\pi'_h(a|s) - \mu'_h(a|s)| \leqslant 4\|w_h - w'_h\|_2 \,. \tag{B.2}$$

Now we construct an ε-net for $\varepsilon \in (0,1)$. Let $\mathcal{N}_{\varepsilon/4}(W, \|\cdot\|_2)$ be a $\varepsilon/4$-net in the space of weights $W = \{w_h \in \mathbb{R}^d : \|w_h\|_2 \leqslant R\}$. It satisfies (see i.e. Handel 2016)

$$\log \mathcal{N}(\varepsilon/4, W, \|\cdot\|_2)| \leqslant d \log(12R/\varepsilon) \,.$$

Next, we show that policies with weights that correspond to a covering of size $\mathcal{N}(\varepsilon/4, W_h, \|\cdot\|_2)$ forms an ε-net in $\mathcal{F}_h$. Let $\pi_h \in \mathcal{F}_h$ be an arbitrary policy with parameter w_h. Let w'_h be in the covering of size $\mathcal{N}(\varepsilon/4, W, \|\cdot\|_2)$ be a parameter that satisfies $\|w_h - w'_h\|_2 \leqslant \varepsilon/4 \leqslant 1$. Let us fix μ_h as a policy corresponding to w'_h. Since $\|w_h - w'_h\|_2 \leqslant 1$, (B.2) is applicable. Thus

$$\|\pi_h - \mu_h\|_\infty = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\pi_h(a|s) - \mu_h(a|s)| \leqslant 4\|w_h - w'_h\|_2 \leqslant \varepsilon \,.$$

Therefore, policies that correspond to an $\varepsilon/4$-net in w_h form an ε-net in $\mathcal{F}$ and we have an upper bound on the size of the ε-net. As a result, $\mathcal{F}_h$ satisfies Assumption 1 with a dimension $d_{\mathcal{F}} = d$, a scaling factor $R_{\mathcal{F}} = 12R$ and $\gamma = 1/(N^{\mathrm{E}} + A)$. Additionally, by construction of $\mathcal{F}_h$ and Assumption 3, the last Assumption 2 holds with $\kappa = A/(N^{\mathrm{E}} + A)$. Therefore, we can apply Theorem 3.3 and obtain with probability at least $1 - \delta$

$$\mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}} \| \pi^{\mathrm{BC}}) \leqslant \frac{6dH \cdot (\log(\mathrm{e}^3(N^{\mathrm{E}} + A)) \cdot (\log(24HN^{\mathrm{E}}(N^{\mathrm{E}} + A)R/\delta)))}{N^{\mathrm{E}}} + \frac{18AH}{N^{\mathrm{E}}} \,.$$

Using of $A \leqslant N^{\mathrm{E}}$ concludes the statement. □

B.2.4 Concentration Results

In this section, we state important results on the concentration of the stochastic error for the risk estimates.

Recall the definition of the κ-greedy version of the expert policy as follows

$$\pi_h^{\mathrm{E},\kappa}(a|s) = (1-\kappa)\pi_h^{\mathrm{E}}(a|s) + \frac{\kappa}{A} = (1-\kappa) \cdot \left(\pi_h^{\mathrm{E}}(a|s) + \frac{\kappa}{(1-\kappa)A}\right) .$$

Lemma B.4. *Let π^{E} be a fixed expert policy. Let $(s_h^t, a_h^t)_{t=1}^N$ be an i.i.d. sequence of state-action pairs generated by following the policy π^{E} at step h. For $\gamma \in (0, 1/A)$ let π a policy such that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ it holds $\pi_h(a|s) \geqslant \gamma$. Then for any $\delta \in (0,1)$ and any $\kappa < 1/2$ with probability at least $1 - \delta$*

$$\left|\frac{1}{N}\sum_{t=1}^N \log\left(\frac{\pi_{h,\sigma}^{\mathrm{E}}(a_h^t|s_h^t)}{\pi_h(a_h^t|s_h^t)}\right) - \mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi_h(a_h|s_h)}\right)\right]\right| \leqslant \sqrt{\frac{2\log(\mathrm{e}^2/\gamma)\,\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}|\pi_h)\log(2/\delta)}{N}}$$
$$+ \frac{2\log(A\mathrm{e}^3/(A\gamma \wedge \kappa)\cdot\log(2/\delta)}{3N} + \frac{5\kappa}{1-\kappa} \,.$$

Proof. As a first step, we can apply Bernstein inequality

$$\left|\frac{1}{N}\sum_{t=1}^N \log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h^t|s_h^t)}{\pi_h(a_h^t|s_h^t)}\right) - \mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi_h(a_h|s_h)}\right)\right]\right| \leqslant \sqrt{\frac{2\mathrm{Var}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi_h(a_h|s_h)}\right)\right]\log(2/\delta)}{N}}$$
$$+ \frac{2(\log(1/\gamma) \vee \log(A/\kappa)\cdot\log(2/\delta)}{3N} \,.$$

Next, we want to upper bound a variance in terms of $\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}} \| \pi_h)$. We start from a bound of square root variance in terms of the second moment and Minkowski inequality

$$\begin{aligned}
\sqrt{\mathrm{Var}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi_h(a_h|s_h)}\right)\right]} &\leqslant \sqrt{\mathbb{E}_{\pi^{\mathrm{E}}}\left[\left(\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi_h(a_h|s_h)}\right)\right)^2\right]} \\
&= \sqrt{\mathbb{E}_{\pi^{\mathrm{E}}}\left[\left(\log\left(\frac{\pi_h^{\mathrm{E}}(a|s)}{\pi_h(a|s)}\right) + \log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a|s)}{\pi_h^{\mathrm{E}}(a|s)}\right)\right)^2\right]} \\
&\leqslant \sqrt{\mathbb{E}_{\pi^{\mathrm{E}}}\left[\left(\log\left(\frac{\pi_h^{\mathrm{E}}(a_h|s_h)}{\pi_h(a_h|s_h)}\right)\right)^2\right]} && = \sqrt{(\mathbf{A})} \\
&+ \sqrt{\mathbb{E}_{\pi^{\mathrm{E}}}\left[\left(\log\left(1 + \frac{\kappa}{(1-\kappa)A\pi_h^{\mathrm{E}}(a|s)}\right) + \log(1-\kappa)\right)^2\right]}. && = \sqrt{(\mathbf{B})}
\end{aligned}$$

Term (A). The result below directly follows from Lemma 4 of Yang and Barron (1998). However, for completeness, we prove it here.

First, we notice that

$$(\mathbf{A}) = \mathbb{E}_{\pi^{\mathrm{E}}}\left[\sum_{a\in\mathcal{A}} \pi_h^{\mathrm{E}}(a|s_h) \log^2\left(\frac{\pi_h^{\mathrm{E}}(a|s_h)}{\pi_h(a|s_h)}\right)\right].$$

To analyze this term, we define an f-divergence (Sason and Verdú 2016) for a function f as follows

$$D_f(\pi_h^{\mathrm{E}}(s) \| \pi_h(s)) = \sum_{a\in\mathcal{A}} f\left(\frac{\pi_h^{\mathrm{E}}(a|s)}{\pi_h(a|s)}\right)\pi_h(a|s)\,.$$

In particular, $\mathrm{KL}(\pi_h^{\mathrm{E}}(s), \pi_h(s)) = D_g(\pi_h^{\mathrm{E}}(s)\|\pi_h(s))$ for $g(t) = t\log t + (1-t)$ and, moreover for $f(t) = t\log^2(t)$

$$D_f(\pi_h^{\mathrm{E}}(s) \| \pi_h(s)) = \sum_{a\in\mathcal{A}} \pi_h^{\mathrm{E}}(a|s) \log^2\left(\frac{\pi_h^{\mathrm{E}}(a|s)}{\pi_h(a|s)}\right).$$

Then we notice that g and f are non-negative function and, moreover, its argument t takes values in $(0,1) \cup (1, 1/\gamma]$ since for $t = 1$ both functions are zero. First, we analyze the ratio for f and g for any $t \in (0,1)$

$$r(t) = \frac{f(t)}{g(t)} = \frac{t\log^2(t)}{t\log t + (1-t)}\,.$$

To bound this function, let us prove that it is monotone for all $t \in (0,1)$

$$r'(t) = \frac{\overbrace{\log(t)}^{\leqslant 0} \cdot \overbrace{((t+1)\log(t) + 2(1-t))}^{\leqslant 0}}{(t\log t + (1-t))^2} \geqslant 0\,.$$

Thus for any $t \in (0,1)$

$$r(t) \leqslant \lim_{t\to 1} \frac{t\log^2(t)}{t\log t + (1-t)} = 2\,.$$

Next, we analyze the segment $t \in (1, 1/\gamma]$.

$$r(t) = \frac{t\log^2(t)}{t\log t + (1-t)} \leqslant \log(t) + 2 \leqslant \log(1/\gamma) + 2 = \log(\mathrm{e}^2/\gamma)\,.$$

since

$$t\log^2(t) \leqslant t\log^2(t) + (1-t)\log t + 2t\log(t) + 2(1-t) \iff (t+1)\log(t) + 2(1-t) \geqslant 0 \quad \forall t > 1\,.$$

Therefore we have for any $t \in (0,1) \cup (1, 1/\gamma]$ and as a simple corollary

$$f(t) \leqslant \log(\mathrm{e}^2/\gamma) \cdot g(t) \Rightarrow D_f(\pi_h^{\mathrm{E}}(s)\|\pi_h(s)) \leqslant \log(\mathrm{e}^2/\gamma) \cdot \mathrm{KL}(\pi_h^{\mathrm{E}}(s)\|\pi_h(s))\,.$$

Finally, we have

$$(\mathbf{A}) \leqslant \log(\mathrm{e}^2/\gamma)\mathbb{E}_{\pi^{\mathrm{E}}}\Big[\mathrm{KL}(\pi_h^{\mathrm{E}}(s_h), \pi_h(s_h))\Big] = \log(\mathrm{e}^2/\gamma)\,\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi_h)\,.$$

Term (B). We can rewrite this term as follows using inequality $(a+b)^2 \leqslant 2a^2 + 2b^2$

$$(\mathbf{B}) \leqslant 2\mathbb{E}_{\pi^{\mathrm{E}}}\left[\sum_{a:\pi_h^{\mathrm{E}}(a|s_h)>0} \pi_h^{\mathrm{E}}(a|s_h)\log^2\left(1 + \frac{\kappa}{(1-\kappa)A\pi_h^{\mathrm{E}}(a_h|s_h)}\right)\right] + 2\left(\frac{\kappa}{1-\kappa}\right)^2.$$

Next we analyze the function $g(x) = x\log^2(1+\varepsilon/x)$. Its derivative is equal to

$$g'(x) = \log\left(1+\frac{\varepsilon}{x}\right)\cdot\left(\log\left(1+\frac{\varepsilon}{x}\right) - \frac{2\varepsilon}{x+\varepsilon}\right).$$

Since $\varepsilon > 0$. we can define $x^\star$ as a root of equation $g'(x) = 0$ for $x > 0$. Notice that it will be maximum of $g(x)$, thus for $\varepsilon > 0$

$$g(x) \leqslant g(x^\star) = x^\star\left(\log\left(1+\frac{\varepsilon}{x^\star}\right)\right)^2 = x^\star\frac{4\varepsilon^2}{(x^\star+\varepsilon)^2} \leqslant \frac{4\varepsilon^2}{x^\star+\varepsilon} \leqslant 4\varepsilon\,.$$

Therefore

$$(\mathbf{B}) \leqslant \frac{8\kappa}{1-\kappa} + 2\left(\frac{\kappa}{1-\kappa}\right)^2.$$

Final bound on variance. Combining these two bounds, we have

$$\sqrt{\mathrm{Var}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi_h(a_h|s_h)}\right)\right]} \leqslant \sqrt{\log(\mathrm{e}^2/\gamma)\cdot\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi_h)} + \sqrt{8\kappa/(1-\kappa) + 2\kappa^2/(1-\kappa)^2}\,.$$

Using this bound on variance, we can bound the main stochastic term

$$\sqrt{\frac{2\mathrm{Var}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi_h(a_h|s_h)}\right)\right]\log(2/\delta)}{N}} \leqslant \sqrt{\frac{2\log(\mathrm{e}^2/\gamma)\,\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi_h)\log(2/\delta)}{N}} + 2\sqrt{\frac{(4\kappa/(1-\kappa) + \kappa^2/(1-\kappa)^2)\log(2/\delta)}{N}}\,.$$

Next, we use an inequality $2ab \leqslant a^2 + b^2$ to obtain

$$2\sqrt{(4\kappa/(1-\kappa) + \kappa^2/(1-\kappa)^2)\cdot\frac{\log(2/\delta)}{N}} \leqslant (4\kappa/(1-\kappa) + \kappa^2/(1-\kappa)^2) + \frac{2\log(2/\delta)}{N}\,.$$

Finally, since $k/(1-\kappa) \leqslant 1$, we conclude the statement. □

Next we define $\Pi_{\gamma,h}$ a set of all one-step policies $\pi_h(a|s)$ such that

$$\inf_{(s,a)\in\mathcal{S}\times\mathcal{A}} \pi_h(a|s) \geqslant \gamma\,.$$

This set forms a metric space with a metric induced by ℓ_∞-norm

$$\|\pi_h - \pi'_h\|_\infty = \sup_{(s,a)\times\mathcal{S}\times\mathcal{A}} |\pi_h(a|s) - \pi'_h(a|s)|\,.$$

Lemma B.5. *Let $\mathcal{F}$ be sub-space of $\Pi_{\gamma,h}$ with an induced metric, such that it satisfies for all $\varepsilon \in (0,1)$*

$$\log|\mathcal{N}(\varepsilon,\mathcal{F},\|\cdot\|_\infty)| \leqslant d\log(R/\varepsilon)\,.$$

for some positive constants $R, d > 0$. Then with probability at least $1-\delta$ the following holds for all $\pi_h \in \mathcal{F}$ simultaneously

$$\begin{aligned}\left|\frac{1}{N}\sum_{t=1}^{N}\left(\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h^t|s_h^t)}{\pi_h(a_h^t|s_h^t)}\right) - \mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi_h(a_h|s_h)}\right)\right]\right)\right| \\ \leqslant \sqrt{\frac{2\log(\mathrm{e}^2/\gamma)\,\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi_h)\cdot d(\log(2NR/\gamma)+\log(1/\delta))}{N}} \\ + \frac{5(\log(A\mathrm{e}^3/(\gamma\wedge\sigma))\cdot d(\log(2NR/\gamma)+\log(1/\delta)))}{3N} + \frac{8\kappa}{1-\kappa}\,.\end{aligned}$$

Proof. Let $\mathcal{N}_\varepsilon$ be a minimal ε-net of $\mathcal{F}$ for ε that will be specified later. Combining Lemma B.4 with a union bound over $\mathcal{N}_\varepsilon$ we have for any $\pi'_h \in \mathcal{N}_\varepsilon$ with probability at least $1-\delta$

$$\begin{aligned}\left|\frac{1}{N}\sum_{t=1}^{N}\left(\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h^t|s_h^t)}{\pi'_h(a_h^t|s_h^t)}\right) - \mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi'_h(a_h|s_h)}\right)\right]\right)\right| \\ \leqslant \sqrt{\frac{2\log(\mathrm{e}^2/\gamma)\,\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi'_h)\cdot d\log(2R/(\varepsilon\delta))}{N}} \\ + \frac{2(\log(A\mathrm{e}^3/(A\gamma\wedge\kappa))\cdot d\log(2R/(\varepsilon\delta))}{3N} + \frac{5\kappa}{1-\kappa}\,.\end{aligned} \tag{B.3}$$

Next, we select an arbitrary policy $\pi_h \in \mathcal{F}$ and let $\pi'_h \in \mathcal{N}_\varepsilon$ be ε-close policy to π_h. Then

$$\begin{aligned}\left|\frac{1}{N}\sum_{t=1}^{N}\left(\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h^t|s_h^t)}{\pi_h(a_h^t|s_h^t)}\right) - \mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi_h(a_h|s_h)}\right)\right]\right)\right| \\ \leqslant \left|\frac{1}{N}\sum_{t=1}^{N}\left(\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h^t|s_h^t)}{\pi'_h(a_h^t|s_h^t)}\right) - \mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi'_h(a_h|s_h)}\right)\right]\right)\right| \\ + \left|\frac{1}{N}\sum_{t=1}^{N}\left(\log\left(\frac{\pi'_h(a_h^t|s_h^t)}{\pi_h(a_h^t|s_h^t)}\right) - \mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi'_h(a_h|s_h)}{\pi_h(a_h|s_h)}\right)\right]\right)\right|.\end{aligned}$$

We start from bounding the second term, which could be done as follows

$$\left|\frac{1}{N}\sum_{t=1}^{N}\left(\log\left(\frac{\pi'_h(a_h^t|s_h^t)}{\pi_h(a_h^t|s_h^t)}\right) - \mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi'_h(a_h|s_h)}{\pi_h(a_h|s_h)}\right)\right]\right)\right| \leqslant 2\max_{s,a}\left|\log\left(\frac{\pi'_h(a|s)}{\pi_h(a|s)}\right)\right|.$$

Next, we use simple inequalities

$$\log\left(\frac{\pi'_h(a|s)}{\pi_h(a|s)}\right) = \log\left(1+\frac{\pi'_h(a|s)-\pi_h(a|s)}{\pi_h(a|s)}\right) \leqslant \frac{|\pi'_h(a|s)-\pi_h(a|s)|}{\gamma} \leqslant \frac{\varepsilon}{\gamma}\,,$$

and, in the opposite direction, we can use the same reasoning

$$\log\left(\frac{\pi'_h(a|s)}{\pi_h(a|s)}\right) = -\log\left(\frac{\pi_h(a|s)}{\pi'_h(a|s)}\right) \geqslant -\frac{\varepsilon}{\gamma}\,.$$

Thus, applying (B.3) for the first term we obtain

$$\begin{aligned}\left|\frac{1}{N}\sum_{t=1}^{N}\left(\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h^t|s_h^t)}{\pi_h(a_h^t|s_h^t)}\right) - \mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi_h(a_h|s_h)}\right)\right]\right)\right| \\ \leqslant \sqrt{\frac{2\log(\mathrm{e}^2/\gamma)\,\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi'_h)\cdot d\log(2R/(\varepsilon\delta))}{N}} \\ + \frac{2(\log(A\mathrm{e}^3/(A\gamma\wedge\kappa))\cdot d\log(2R/(\varepsilon\delta))}{3N} + \frac{5\kappa}{1-\kappa} + 2\varepsilon/\gamma\,.\end{aligned}$$

Next, we use a similar inequality to obtain

$$\begin{aligned}\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi'_h) &= \mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E}}(a_h|s_h)}{\pi'_h(a_h|s_h)}\right)\right] \\ &= \mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi_h) + \mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h(a_h|s_h)}{\pi'_h(a_h|s_h)}\right)\right] \\ &\leqslant \mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi_h) + \frac{\varepsilon}{\gamma}\,.\end{aligned}$$

Finally, applying inequalities $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ and $2ab \leqslant a^2 + b^2$

$$\begin{aligned}\left|\frac{1}{N}\sum_{t=1}^{N}\left(\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h^t|s_h^t)}{\pi_h(a_h^t|s_h^t)}\right) - \mathbb{E}_{\pi^{\mathrm{E}}}\left[\log\left(\frac{\pi_h^{\mathrm{E},\kappa}(a_h|s_h)}{\pi_h(a_h|s_h)}\right)\right]\right)\right| \\ \leqslant \sqrt{\frac{2d\log(\mathrm{e}^2/\gamma)\,\mathrm{KL}_{\mathrm{traj}}(\pi_h^{\mathrm{E}}\|\pi_h)\log(2R/(\varepsilon\delta))}{N}} \\ + \frac{d\log(\mathrm{e}^2/\gamma)\cdot\log(2R/(\varepsilon\delta))}{N} + \frac{\varepsilon}{\gamma} \\ + \frac{2(\log(A\mathrm{e}^3/(A\gamma\wedge\kappa))\cdot d\log(2R/(\varepsilon\delta))}{3N} + \frac{5\kappa}{1-\kappa} + 2\varepsilon/\gamma\,.\end{aligned}$$

We conclude the statement by rearranging the terms and taking $\varepsilon = \gamma\cdot\kappa/(1-\kappa)$. □

B.2.5 Proof of Lower Bounds

General setup

In this section, we provide a lower bound on estimation in KL-divergence using a framework of Chapter 2 by Tsybakov (2008). Our goal is to obtain a lower bound on minimax risk that is defined as follows

$$\inf_{\widehat{\pi}}\sup_{\pi\in\mathcal{F}}\mathbb{E}_{\tau_1,\ldots,\tau_N\sim\pi}[\mathrm{KL}_{\mathrm{traj}}(\pi\|\widehat{\pi})]\,,$$

where infimum is taken over all estimators that map the sampled trajectories $(\tau_1,\ldots,\tau_N)$ to a policy from the hypothesis class $\mathcal{F} \triangleq \mathcal{F}_1\times\ldots\mathcal{F}_H$.

Let us consider a specific type of MDPs where the transition kernel $p_h(s,a)$ does not depend on a state-action pair (s,a): $\forall(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H], \forall A\in\mathcal{F}_{\mathcal{S}} : p_h(A|s,a) = \mu_h(A)$ for fixed measures μ_h. In particular, for $h=1$ we always have $\mu_h = \delta_{s_1}$ is a Dirac measure at initial state s_1.

Then, we define over the space of all policies the following specific distance defined through the Hellinger distance

$$\rho_h(\pi_h, \pi'_h) = \sqrt{\mathbb{E}_{s\sim\mu_h}[d^2_{\mathcal{H}}(\pi_h, \pi'_h)]}, \qquad d^2_{\mathcal{H}}(\pi_h, \pi'_h) = \sum_{a\in\mathcal{A}} \left(\sqrt{\pi_h(a|s)} - \sqrt{\pi'_h(a|s)}\right)^2$$

and we define the following distance for the space of full policies (the triangle inequality follows from Minkowski inequality)

$$\rho(\pi, \pi') = \sqrt{\sum_{h=1}^{H} \rho^2_h(\pi_h, \pi'_h)}\,. \tag{B.4}$$

Next, we impose the following metric-specific assumption for our hypothesis classes

Assumption 5. For all $h \in [H]$ a of the function class $\mathcal{F}_h$ with respect to the metric ρ_h satisfies

$$\forall \varepsilon \in (0,1) : \log \mathcal{P}(\varepsilon, \mathcal{F}_h, \rho_h) \geqslant d_h \log(R/\varepsilon)$$

for constants $d_h \geqslant 0$ and $R > 0$.

In particular, Lemma B.10 implies that $\log \mathcal{P}(\varepsilon, \mathcal{F}, \rho) \geqslant \sum_{h=1}^{H} d_h \log(R/\varepsilon)$.

Theorem B.6. *Let Assumption 5 holds and let us define $D = \sum_{h=1}^{H} d_h$. Also we assume that for any $\pi \in \mathcal{F}$ it holds $\pi_h(s,a) \geqslant \gamma$ for $\gamma \in (0, 1/A)$. Let us assume $D \geqslant 5$ and $n \geqslant \mathrm{e}^2 D/R^2$. Then, the following minimax lower bound holds*

$$\min_{\widehat{\pi}} \max_{\pi\in\mathcal{F}} \mathbb{E}[\mathrm{KL}_{\mathrm{traj}}(\pi\|\widehat{\pi})] \geqslant \frac{D}{16N\log(\mathrm{e}^2/\gamma)}\,.$$

Proof. First, we notice by the first part of Lemma B.11

$$\min_{\widehat{\pi}} \max_{\pi\in\mathcal{F}} \mathbb{E}\left[\frac{\log(\mathrm{e}^2/\gamma)N}{D}\,\mathrm{KL}_{\mathrm{traj}}(\pi\|\widehat{\pi})\right] \geqslant \min_{\widehat{\pi}} \max_{\pi\in\mathcal{F}} \mathbb{E}\left[\frac{\log(\mathrm{e}^2/\gamma)N}{D}\rho^2(\pi,\widehat{\pi})\right],$$

where the expectation is taken with respect to a sample $\tau_1, \ldots, \tau_N$. Next, we can follow the general reduction scheme, see Chapter 2.2 by Tsybakov (2008). By Markov inequality

$$\min_{\widehat{\pi}} \max_{\pi\in\mathcal{F}} \mathbb{E}\left[\frac{n\log(\mathrm{e}^2/\gamma)}{D}\rho^2(\pi,\widehat{\pi})\right] \geqslant \frac{1}{4}\min_{\widehat{\pi}} \max_{\pi\in\mathcal{F}} \mathbb{P}\left[\rho(\pi,\widehat{\pi}) \geqslant \sqrt{\frac{D}{4N\log(\mathrm{e}^2/\gamma)}}\right].$$

Next, we use the reduction to a finite hypothesis class. Define $\zeta = \sqrt{D/(4\log(\mathrm{e}^2/\gamma)N)}$ and take P is a maximal ζ-separated set of size $M+1 = \mathcal{P}(\zeta, \mathcal{F}, \rho)$ and enumerate all the policies in it as $\pi_0, \ldots, \pi_M$. Therefore we obtain

$$\begin{aligned}
\min_{\widehat{\pi}} \max_{\pi\in\mathcal{F}} \mathbb{P}_{\tau_1,\ldots,\tau_N\sim\pi}&\left[\rho(\pi,\widehat{\pi}) \geqslant \sqrt{\frac{D}{4N\cdot\log(\mathrm{e}^2/\gamma)}}\right] \\
&\geqslant \min_{\widehat{\pi}} \max_{j\in\{0,\ldots,M\}} \mathbb{P}_{\tau_1,\ldots,\tau_N\sim\pi_j}\left[\rho(\pi_j,\widehat{\pi}) \geqslant \sqrt{\frac{D}{4N\cdot\log(\mathrm{e}^2/\gamma)}}\right].
\end{aligned}$$

Let us define $\psi^\star = \arg\min_{j=0,\ldots,M} \rho(\pi_j, \widehat{\pi})$. Then we have that if $\psi^\star \neq j$, then

$$2\rho(\pi_j,\widehat{\pi}) \geqslant \rho(\pi_j,\widehat{\pi}) + \rho(\pi_{\psi^\star},\widehat{\pi}) \geqslant \rho(\pi_j,\pi_{\psi^\star})\,.$$

Since $j \neq \pi^\star$, then by definition of ζ-separable set we have $\rho(\pi_j, \widehat{\pi}) \geqslant \zeta/2$. As a result

$$\min_{\widehat{\pi}} \max_{j \in \{0,\dots,M\}} \mathbb{P}_{\tau_1,\dots,\tau_N \sim \pi_j}\left[\rho(\pi_j, \widehat{\pi}) \geqslant \sqrt{\frac{D}{4N \cdot \log(\mathrm{e}^2/\gamma)}}\right] \geqslant \min_{\widehat{\pi}} \max_{j \in \{0,\dots,M\}} \mathbb{P}_{\tau_1,\dots,\tau_N \sim \pi_j}[\psi^\star \neq j]\,.$$

Finally, taking infimum over all hypothesis tests, we obtain

$$\min_{\widehat{\pi}} \max_{\pi \in \mathcal{F}} \mathbb{P}\left[\rho(\pi, \widehat{\pi}) \geqslant \sqrt{\frac{D}{4N \cdot \log(\mathrm{e}^2/\gamma)}}\right] \geqslant \inf_{\psi} \max_{j=0,\dots,M} \mathbb{P}_j[\psi \neq j] \triangleq p_{e,M}\,.$$

To lower bound the right-hand side, we apply Proposition 2.3 by Tsybakov (2008). Notice that the maximal ε-packing is ε-net (see Lemma 4.2.6 by Vershynin (2018)). Therefore, by Lemma B.11 we have

$$\frac{1}{M}\sum_{i=1}^{M} \mathrm{KL}(\mathbb{P}_{\tau_1,\dots,\tau_N \sim \pi_i} \| \mathbb{P}_{\tau_1,\dots,\tau_N \sim \pi_0}) = \frac{N}{M}\sum_{i=1}^{M} \mathrm{KL}_{\mathrm{traj}}(\pi_i \| \pi_0) \leqslant n \log(\mathrm{e}^2/\gamma)\frac{1}{M}\sum_{i=1}^{M} \rho^2(\pi_i, \pi_0) \leqslant D \triangleq \alpha_\star\,.$$

Thus by Proposition 2.3 by Tsybakov (2008)

$$p_{e,M} \geqslant \sup_{0<\tau<1}\left[\frac{\tau M}{1+\tau M}\left(1 + \frac{\alpha_\star + \sqrt{\alpha_\star}}{\log(\tau)}\right)\right].$$

Next, we select $\tau_\star$ in a way such that

$$\frac{\alpha_\star + \sqrt{\alpha_\star/2}}{\log(\tau_\star)} = -\frac{1}{2} \iff \log(\tau_\star) = -\frac{1}{2}\left(\alpha_\star + \sqrt{\alpha_\star/2}\right).$$

Therefore we have

$$p_{e,M} \geqslant \frac{1}{2}\frac{\exp(\log(M) - 1/2(\alpha_\star + \sqrt{\alpha_\star/2}))}{1 + \exp(\log(M) - 1/2(\alpha_\star + \sqrt{\alpha_\star/2}))}\,.$$

Notice that the function $f(x) = \exp(x)/(1+\exp(x))$ monotonically increasing. Therefore, it is enough to bound the expression under the exponent from below.

Let us assume that $\alpha_\star \geqslant 5 \iff D \geqslant 5$. Then we have

$$\begin{aligned}\log(M) - 1/2(\alpha^\star + \sqrt{\alpha_\star/2}) &\geqslant \log(M+1) - \frac{2+\sqrt{2}}{4}\alpha_\star - \log(2) \\ &\geqslant D\log\left(R\sqrt{\frac{4\log(\mathrm{e}^2/\gamma)N}{D}}\right) - D \\ &\geqslant \frac{D}{2}\left(\log\left(\frac{4\log(\mathrm{e}^2/\gamma)R^2 \cdot N}{D}\right) - 2\right).\end{aligned}$$

To show that the expression above is non-negative, it is enough to guarantee

$$\log(N) + \log(R^2/D) \geqslant 2 \iff N \geqslant \mathrm{e}^2 D/R^2\,.$$

Under this condition, we have $p_{e,M} \geqslant 1/4$ concluding the statement. □

Finite MDPs

For the case of finite MDPs, we additionally specialize the distributions μ_h as a uniform over $\mathcal{S}$ for all $h > 1$ and $\mu_1 = \delta_{s_1}$.

Lemma B.7. *Let* $\Delta_{A,\gamma} = \{x \in \mathbb{R}^A : \sum_{i=1}^n x_i = 1, x_i \geqslant \gamma\}$ *with* $\gamma < 1/A$. *Then we have for any* $\varepsilon \in (0,1)$

$$\log \mathcal{P}(\varepsilon, \Delta_{A,\gamma}, d_{\mathcal{H}})| \geqslant (A-1)\log((1-A\gamma)/(2\varepsilon))\,.$$

Proof. Let $S = \{x \in \mathbb{R}^A : \sum_{i=1}^n x_i^2 = 1, x_i \geqslant 0\}$. Then there is a mapping $\varphi\colon (S, \|\cdot\|_2) \to (\Delta_A, d_{\mathcal{H}})$ that defines an isometry between these two metric spaces:

$$\forall x, y \in S : \|x - y\|_2 = d_{\mathcal{H}}(\varphi(x), \varphi(y)), \quad \varphi(x) = \sqrt{x}\,,$$

where the square root is applied component-wise. Therefore, it is enough to estimate the packing number of the preimage of $\Delta_{A,\gamma}$ that is defined as follows $S_\gamma(1) = \{x \in \mathbb{R}^A : \sum_{i=1}^A x_i^2 = 1, x_i \geqslant \sqrt{\gamma}\}$ with the same Euclidean metric.

The next step is to proceed with a shift $x \mapsto x + \sqrt{\gamma}$ that will be isometry between $S_\gamma(1)$ and $S_0(1 - A\gamma)$.

Next, we can lower bound the ℓ_2-distance over the sphere by the ℓ_2 distance over the first $A-1$ coordinates and therefore it is enough to consider the packing number of $S_0^\circ(1-A\gamma) = \mathcal{B}(0,1) \cap \{y \in \mathbb{R}^{A-1} : \sum_{i=1}^{A-1} y_i^2 \leqslant 1 - A\gamma\}$.

Finally, we apply the volume argument. In particular, it is enough to compute the volume of $S_0'(1 - A\gamma)$. To do it, we notice that we can represent the ball of radius $1 - A\gamma$ by 2^{A-1} copy of $S_0'(1 - A\gamma)$. Thus

$$\operatorname{vol}(S_0'(1 - A\gamma)) = \frac{(1 - A\gamma)^{A-1}}{2^{A-1}} \cdot \operatorname{vol}(B_2^{A-1})\,.$$

Finally we have by Proposition 4.2.12 by Vershynin (2018)

$$\mathcal{P}(\varepsilon, \Delta_{A,\gamma}, d_{\mathcal{H}})| \geqslant \left(\frac{1 - A\gamma}{2\varepsilon}\right)^{A-1}.$$

□

Lemma B.8. *Let* $(\mathcal{X}, \rho)_{i=1}^K$ *be a metric space such that* $\log|\mathcal{P}_\varepsilon(\mathcal{X}, \rho)| \geqslant d\log(R/\varepsilon)$ *for* $d \geqslant 1$. *and define on the space* $\mathcal{X}^K$ *the following metric* $\rho(x,y) = \frac{1}{K}\sum_{i=1}^K \rho(x_i, y_i)$. *Then*

$$\log \mathcal{P}(\varepsilon, \mathcal{X}^K, \rho) \geqslant dK/2 \cdot \log(R/(8\varepsilon))\,.$$

Proof. Consider the maximal ε-separable set P of the space K. This set could be considered as a finite alphabet of size $q \geqslant (R/\varepsilon)^d$. Let us consider the set P^K as the set of words in alphabet P of size q of length K with a Hoeffding distance. Then we notice that if there is two words $(x, y) \in P^K$ that have Hoeffding distance at least αK for some constant $\alpha \in (0,1)$, then

$$\rho(x,y) = \frac{1}{K}\sum_i \rho(x_i, y_i) \geqslant \alpha\varepsilon.$$

Therefore, if we consider an αK separable set in P^K in terms of Hoeffding distance, it will automatically be a $\alpha\varepsilon$-separable set in the original space $\mathcal{X}^K$. To find such a set, we use the Gilbert–Varshamov bound from coding theory. As a result

$$\mathcal{P}(\alpha\varepsilon, P^K, \rho) \geqslant \frac{1}{\sum_{j=1}^{\lceil \alpha K \rceil} \binom{K}{j}(1-1/q)^j \cdot (1/q)^{K-j}}\,.$$

The denominator could be interpreted as follows: Let $X_1, \dots, X_K$ be $\mathcal{B}\mathrm{er}(1/q)$ random variables.

$$\sum_{j=1}^{\lceil \alpha K \rceil} \binom{K}{j} (1-1/q)^j \cdot (1/q)^{K-j} = \mathbb{P}\left[\sum_{i=1}^{K} X_i \geqslant (1-\alpha)K\right].$$

To upper bound the last probability, we can apply the Chernoff–Hoeffding theorem

$$\mathbb{P}\left[\frac{1}{K}\sum_{i=1}^{K} X_i \geqslant 1/q + (1-\alpha-1/q)\right] \leqslant \exp(-\operatorname{kl}(1-\alpha\|1/q)\cdot K).$$

Take $\alpha = 1/2$, then we have

$$\operatorname{kl}(1/2\|1/q) = \frac{1}{2}\log\left(\frac{q}{2}\right) + \frac{1}{2}\log\left(\frac{q}{q-1}\right) - \frac{1}{2}\log(2) \geqslant \frac{1}{2}\log\left(\frac{q}{4}\right).$$

Thus, we have since $d \geqslant 1$

$$\mathcal{P}(\varepsilon/2, P^K, \rho) \geqslant \exp\{K/2 \cdot \log(q/4)\} \geqslant \exp\{dK/2 \cdot \log(R/(4\varepsilon))\}.$$

By rescaling ε we conclude the statement. □

Corollary B.9. *Assume that $\gamma \leqslant 1/(2A)$. Let us define $\mathcal{F}_h = \Delta^S_{A,\gamma}$ and $\mathcal{F} = \mathcal{F}_1 \times \dots \mathcal{F}_H$. Then Assumption 5 holds with constants $d_h = (A-1)S/2$ for all $h > 1$, $d_1 = (A-1)$ and $R = 1/32$.*

As a result, as soon as $H \geqslant 2, A \geqslant 2$, $HSA \geqslant 40$ and $n \geqslant 512\mathrm{e}^2 HSA$ the following minimax lower bound holds

$$\min_{\widehat{\pi}} \max_{\pi \in \mathcal{F}} \mathbb{E}_{\tau_1,\dots,\tau_N \sim \pi}[\mathrm{KL}_{\mathrm{traj}}(\pi\|\widehat{\pi})] \geqslant \frac{HSA}{128N\log(\mathrm{e}^2/\gamma)}.$$

Technical lemmas

Lemma B.10. *Let $\{(\mathcal{X}_i, \rho_i)\}_{i=1,\dots,K}$ be a collection of relaxed pseudometric spaces that satisfy*

$$\forall \varepsilon \in (0,1) : \log \mathcal{P}(\varepsilon, \mathcal{X}_i, \rho_i) \geqslant d_i \log(R/\varepsilon)$$

for some constants $0 \leqslant d_1 \leqslant d_2$, Then the product space $\mathcal{X} = \mathcal{X}_1 \times \dots \mathcal{X}_K$ with a pseudometric $\rho((x_1,\dots,x_K),(y_1,\dots,y_K)) = \sqrt{\sum_{i=1}^K \rho_i^2(x_i,y_i)}$ satisfies

$$\forall \varepsilon \in (0,1) : \log \mathcal{P}(\varepsilon, \mathcal{X}, \rho) \geqslant \sum_{i=1}^{K} d_i \log(R/\varepsilon).$$

Proof. Let $P_1, \dots, P_K$ be a maximal ε-separated set in the corresponding spaces $\mathcal{M}_1, \dots, \mathcal{M}_K$. Then we want to show that the set $P_1 \times \dots \times P_K$ is also ε-separated set in the product space.

Let $\mathbf{x} \neq \mathbf{x}'$ be two point in P. Then

$$\rho(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{i=1}^{K} \rho_i^2(x_i, x_i')} \geqslant \varepsilon$$

since $\mathbf{x}$ and $\mathbf{x}'$ are different in at least one coordinate. As a result, we have

$$\log \mathcal{P}(\varepsilon, \mathcal{M}_i, \rho_i)| \geqslant \sum_{i=1}^{K} \log \mathcal{P}(\varepsilon, \mathcal{M}_i, \rho_i) \geqslant \sum_{i=1}^{K} d_i \log(R/\varepsilon).$$

□

Lemma B.11. *Let $\pi, \pi' \in \Pi_\gamma$. Let ρ be an averaged Hellinger distance distance defined in (B.4). Then the following inequality holds*

$$\rho^2(\pi, \pi') \leqslant \mathrm{KL}_{\mathrm{traj}}(\pi \| \pi') \leqslant \log(\mathrm{e}^2/\gamma)\rho^2(\pi, \pi')\,.$$

Proof. It is enough to show that for two measures $p, q \in \Delta_n$ such that $\min_i p_i/q_i \geqslant \gamma$ the following holds

$$d^2_{\mathcal{H}}(p, q) \leqslant \mathrm{KL}(p\|q) \leqslant \log(\mathrm{e}^2/\gamma) d^2_{\mathcal{H}}(p, q)\,.$$

The lower bound holds by Lemma 2.4 of Tsybakov (2008), and the upper bound holds by Lemma 4 of Yang and Barron (1998).

□

B.2.6 Imitation Learning Guarantees

In this appendix, we present guarantees that give behavior cloning procedure in the setting of imitation learning for finite MDPs and compare obtained results to (Ross and Bagnell 2010; Rajaraman et al. 2020).

General expert. Using Pinsker inequality in the space of trajectories (see Lemma B.13) and the fact the expert policy is ε_{E}-optimal we deduce the following bound on the optimality gap of the behavior cloning policy with probability at least $1 - \delta$

$$V_1^\star(s_1) - V_1^{\pi^{\mathrm{BC}}}(s_1) \leqslant \varepsilon_{\mathrm{E}} + \widetilde{\mathcal{O}}\left(\sqrt{\frac{SAH^3}{N^{\mathrm{E}}}}\right).$$

We remark that we obtain a similar rate as in BPI, see for example Ménard et al. (2021), where instead of observing N^{E} demonstrations, we collect the same number of trajectories (also observing the rewards) by interacting sequentially with the MDPs. This seems a bit counter-intuitive since we expect to learn faster by directly observing the expert.

However, we get an improved rate for the deterministic expert using the following variance-aware Pinkser inequality.

Lemma B.12. *Let π and π' be arbitrary policies. Then, the following upper bound holds*

$$V_1^\pi(s_1) - V_1^{\pi'}(s_1) \leqslant \sqrt{4\mathbb{E}_\pi\left[\sum_{h=1}^H \mathrm{Var}_\pi Q_h^{\pi'}(s_h)\right] \cdot \mathrm{KL}_{\mathrm{traj}}(\pi\|\pi')} + 5H\,\mathrm{KL}_{\mathrm{traj}}(\pi\|\pi')\,.$$

Proof. Let us start from Lemma B.15 and Lemma D.9.

$$\begin{aligned}
V_1^\pi(s_1) - V_1^{\pi'}(s_1) &= \mathbb{E}_\pi\left[\sum_{h=1}^H [\pi_h - \pi'_h] Q_h^{\pi'}(s_h)\right] \\
&\leqslant \mathbb{E}_\pi\left[\sum_{h=1}^H \sqrt{2\mathrm{Var}_{\pi'} Q_h^{\pi'}(s_h) \cdot \mathrm{KL}(\pi_h(s_h)\|\pi'_h(s_h))} + \frac{H}{3}\,\mathrm{KL}(\pi_h(s_h)\|\pi'_h(s_h))\right] \\
&\leqslant \sqrt{2\mathbb{E}_\pi\left[\sum_{h=1}^H \mathrm{Var}_{\pi'} Q_h^{\pi'}(s_h)\right]} \cdot \sqrt{\mathrm{KL}_{\mathrm{traj}}(\pi\|\pi')} + \frac{H}{3}\,\mathrm{KL}_{\mathrm{traj}}(\pi\|\pi')\,,
\end{aligned}$$

where in the last line we have applied Cauchy-Schwartz inequality. Next we apply Lemma D.10 and obtain

$$\mathbb{E}_\pi\left[\sum_{h=1}^H \mathrm{Var}_{\pi'} Q_h^{\pi'}(s_h)\right] \leqslant 2\mathbb{E}_\pi\left[\sum_{h=1}^H \mathrm{Var}_\pi Q_h^{\pi'}(s_h)\right] + 4H^2\,\mathrm{KL}_{\mathrm{traj}}(\pi\|\pi')\,.$$

By inequality $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ for any $a, b \geqslant 0$ we have

$$V_1^{\pi}(s_1) - V_1^{\pi'}(s_1) \leqslant \sqrt{4\mathbb{E}_{\pi}\left[\sum_{h=1}^{H} \operatorname{Var}_{\pi} Q_h^{\pi'}(s_h)\right] \cdot \mathrm{KL}_{\mathrm{traj}}(\pi \| \pi') + 5H\,\mathrm{KL}_{\mathrm{traj}}(\pi \| \pi')}\,.$$

□

Deterministic expert. If we assume that the expert policy is deterministic, for example, a deterministic optimal policy, then we can improve the bound on the optimality gap since the variance term in Lemma B.12 is zero,

$$V_1^{\star}(s_1) - V_1^{\pi^{\mathrm{BC}}}(s_1) \leqslant \varepsilon_{\mathrm{E}} + \widetilde{\mathcal{O}}\left(\frac{SAH^2}{N^{\mathrm{E}}}\right).$$

Ross and Bagnell (2010) also consider behavior cloning with a deterministic expert and provide a bound in terms of the classification-type error of the behavior cloning policy to imitate the expert

$$V_1^{\star}(s_1) - V_1^{\pi^{\mathrm{BC}}}(s_1) \leqslant \varepsilon_{\mathrm{E}} + e_{\mathrm{E}}(\pi^{\mathrm{BC}}) \text{ where } e_{\mathrm{E}}(\pi^{\mathrm{BC}}) = \frac{1}{H}\mathbb{E}^{\pi^{\mathrm{BC}}}\left[\sum_{h=1}^{H} \mathbb{1}\{\pi^{\mathrm{E}}(a_h|s_h) \neq 1\}\right].$$

We can easily recover our bound on the optimality gap of the behavior cloning policy from their bound by noting that $e_{\mathrm{E}}(\pi^{\mathrm{BC}}) \leqslant \mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}} \| \pi^{\mathrm{BC}})/H$, see Lemma B.14 for a proof. In particular, we also remark that the bound scales quadratically with the horizon H and linearly with the number of actions and states SA.

By comparing this bound to the lower bound in Theorem 1.1 of Rajaraman et al. (2020) we see that it is optimal in its dependence on S, H and N. Additional dependence on a number of actions comes from the fact that our behavior cloning algorithm always outputs stochastic policy and obtains additional dependence on a number of actions.

We would like to underline that the bound in Lemma B.12 directly does not give any insights on the performance of the algorithm in the case of non-deterministic optimal expert, whereas Rajaraman et al. (2020) provides $\widetilde{\mathcal{O}}(SH^2/N^{\mathrm{E}})$ guarantees using a non-regularized behavior cloning algorithm. It is connected to the fact that for the optimal policy, it is enough to determine the subset of the support of $\pi^{\star}(s)$ to achieve the policy with the same value.

Finally, we would like to emphasize that our approach is directly generalized to arbitrary parametric function approximation setting, whereas the approach of Rajaraman et al. (2020) could be applied only in the setting of finite MDPs.

Technical Lemmas for Imitation Learning

Lemma B.13. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \{p_h\}_{h=1}^H, \{r_h\}_{h=1}^H, s_1)$ be a finite MDP and let π and π' be two any policies in $\mathcal{M}$. Then*

$$V_1^{\pi}(s_1) - V_1^{\pi'}(s_1) \leqslant H\sqrt{\mathrm{KL}_{\mathrm{traj}}(\pi \| \pi')/2}\,.$$

Proof. Let us define the trajectory distribution $q^{\pi}(\tau)$ for $\tau = (s_1, a_1, \ldots, s_H, a_H)$. Then by the chain rule for KL-divergence we have $\mathrm{KL}(q^{\pi} \| q^{\pi'}) = \mathrm{KL}_{\mathrm{traj}}(\pi \| \pi')$.

Since $r_h \in [0, 1]$, we may apply a variational formula for total variation distance and Pinkser's inequality

$$\begin{aligned}\left|V_1^{\pi}(s_1) - V_1^{\pi'}(s_1)\right| &\leqslant \left|\sum_{h=1}^{H} \mathbb{E}_{\pi}[r_h(s_h, a_h)] - \mathbb{E}_{\pi'}[r_h(s_h, a_h)]\right| \\ &\leqslant H\,\mathrm{TV}(q^{\pi}, q^{\pi'}) \leqslant H\sqrt{\mathrm{KL}_{\mathrm{traj}}(\pi \| \pi')/2}\,.\end{aligned}$$

□

Let us assume that a policy π is deterministic, then define the following notion of the imitation learning error

$$e_\pi(\pi') \triangleq \frac{1}{H}\sum_{h=1}^{H}\mathbb{E}_\pi\left[\sum_{a\in\mathcal{A}}\pi'(a|s_h)\mathbb{1}\{a \neq \pi(s_h)\}\right].$$

Notice that

$$\sum_{a\in\mathcal{A}}\pi'(a|s_h)\mathbb{1}\{a \neq \pi(s_h)\} = \frac{1}{2}\sum_{a\in\mathcal{A}}|\pi'(a|s_h) - \pi(a|s_h)| = \mathrm{TV}(\pi(s_h), \pi'(s_h))\,.$$

Therefore, this quantity could be decomposed as follows

$$e_\pi(\pi') = \frac{1}{H}\sum_{h=1}^{H}\mathbb{E}_\pi[\mathrm{TV}(\pi(s_h), \pi'(s_h))]\,.$$

Lemma B.14. *Let π be a deterministic policy and π' be any policy. Then*

$$e_\pi(\pi') \leqslant \frac{\mathrm{KL}_{\mathrm{traj}}(\pi\|\pi')}{H}\,.$$

Proof. If the policy π is deterministic, then

$$\begin{aligned}\mathrm{KL}(\pi_h(s_h)\|\pi'_h(s_h)) &= \log\left(\frac{1}{\pi'_h(a_h|s_h)}\right) = \log\left(\frac{1}{\pi'_h(\pi_h(s_h)|s_h)}\right)\\ &= \log\left(\frac{1}{1-\sum_{a\in\mathcal{A}}\pi'_h(a|s_h)\mathbb{1}\{a\neq\pi_h(s_h)\}}\right)\\ &= -\log(1-\mathrm{TV}(\pi_h(s_h), \pi'_h(s_h)))\,.\end{aligned}$$

By an inequality $\log(1-x)\leqslant -x$ for any $x>0$ we have $\mathrm{KL}(\pi_h(s_h)\|\pi'_h(s_h)) \geqslant \mathrm{TV}(\pi_h(s_h), \pi'_h(s_h))$. □

The following lemma is known as performance-difference lemma (see, e.g., Kakade and Langford (2002) for a statement in the discounted setting). We provide proof for completeness.

Lemma B.15. *Let π and π' be arbitrary policies. Then, the following decomposition holds*

$$V_1^\pi(s_1) - V_1^{\pi'}(s_1) = \mathbb{E}_\pi\left[\sum_{h=1}^{H}[\pi_h - \pi'_h]Q_h^{\pi'}(s_h)\right].$$

Proof. Let us proceed by backward induction over $h\in[H]$. We want to show the following bound

$$V_h^\pi(s) - V_h^{\pi'}(s) = \mathbb{E}_\pi\left[\sum_{h'=1}^{H}[\pi_{h'} - \pi'_{h'}]Q_{h'}^{\pi'}(s_{h'})|s_h = s\right].$$

For $h = H+1$ both sides of the equation above are equal to zero. Let us assume that the statement holds for any $h' > h$. Then we have

$$\begin{aligned}V_h^\pi(s) - V_h^{\pi'}(s) &= \pi Q_h^\pi(s) - \pi' Q_h^{\pi'}(s) = \pi[Q_h^\pi - Q_h^{\pi'}](s) + [\pi-\pi']Q^{\pi'}(s)\\ &= \mathbb{E}_\pi\left[V_{h+1}^\pi(s_{h+1}) - V_{h+1}^{\pi'}(s_{h+1}) + [\pi-\pi']Q^{\pi'}(s_h)|s_h = s\right].\end{aligned}$$

By the induction hypothesis and the tower property of mathematical expectation, we conclude the statement. □

B.3 Proof for Demonstration-regularized RL

Here for completeness we present the proof of Theorem 3.13.

Theorem (Restatement of Theorem 3.13)**.** *For all $\varepsilon > 0$, $\delta \in (0,1)$, the* `UCBVI-Ent+` */*`LSVI-UCB-Ent` *algorithms defined in Appendix B.4.4 /Appendix B.5.4 are (ε,δ)-PAC for the regularized BPI with sample complexity*

$$\mathcal{C}(\varepsilon,\delta) = \widetilde{\mathcal{O}}\left(\frac{H^5 S^2 A}{\lambda\varepsilon}\right) \textit{ (finite)} \qquad \mathcal{C}(\varepsilon,\delta) = \widetilde{\mathcal{O}}\left(\frac{H^5 d^2}{\lambda\varepsilon}\right) \textit{ (linear).}$$

Additionally, assume that the expert policy is $\varepsilon_{\mathrm{E}} = \varepsilon/2$-optimal and satisfies Assumption 3 in the linear case. Let π^{BC} be the behavior cloning policy obtained using corresponding function sets described in Section 3.3. Then demonstration-regularized RL based on `UCBVI-Ent+` */*`LSVI-UCB-Ent` *with parameters $\varepsilon_{\mathrm{RL}} = \varepsilon/4$, $\delta_{\mathrm{RL}} = \delta/2$ and $\lambda = \widetilde{\mathcal{O}}\big(N^{\mathrm{E}}\varepsilon/(SAH)\big)$ / $\widetilde{\mathcal{O}}\big(N^{\mathrm{E}}\varepsilon/(dH)\big)$ is (ε,δ)-PAC for BPI with demonstration in finite / linear MDPs and has sample complexity of order*

$$\mathcal{C}(\varepsilon,N^{\mathrm{E}},\delta) = \widetilde{\mathcal{O}}\left(\frac{H^6 S^3 A^2}{N^{\mathrm{E}}\varepsilon^2}\right) \textit{ (finite)} \qquad \mathcal{C}(\varepsilon,N^{\mathrm{E}},\delta) = \widetilde{\mathcal{O}}\left(\frac{H^6 d^3}{N^{\mathrm{E}}\varepsilon^2}\right) \textit{ (linear)}.$$

Proof. The first part of the statement is a combination of Theorem B.21 and Theorem B.28. The second part of the statement follows from an upper bound on $\varepsilon_{\mathrm{KL}}^2$ by a behavior cloning (see Appendix B.2.2 and Appendix B.2.3) and Theorem 3.10. □

B.4 Best Policy Identification in Regularized Finite MDPs

In this appendix, we present and analyze the UCBVI-Ent+ algorithm for regularized BPI.

B.4.1 Preliminaries

First, we will detail the general setting of KL-regularized MDPs.

Given some reference policy $\widetilde{\pi}$ and some regularization parameter $\lambda > 0$, instead of looking at the usual value function of a policy π, we consider the trajectory Kullback-Leibler divergence regularized value function $V^{\pi}_{\widetilde{\pi},\lambda,1}(s_1) \triangleq V^{\pi}_1(s_1) - \lambda \mathrm{KL}_{\mathrm{traj}}(\pi, \widetilde{\pi})$. In this case, the value function of the policy π is penalized for moving too far from the reference policy $\widetilde{\pi}$. Interestingly, we can compute the value of policy π and the optimal value thanks to the regularized Bellman equations

$$
\begin{aligned}
Q^{\pi}_{\widetilde{\pi},\lambda,h}(s,a) &= r_h(s,a) + p_h V^{\pi}_{\widetilde{\pi},\lambda,h+1}(s,a)\,, \\
V^{\pi}_{\widetilde{\pi},\lambda,h}(s) &= \pi_h Q^{\pi}_{\widetilde{\pi},\lambda,h}(s) - \lambda\,\mathrm{KL}(\pi_h(s)\|\widetilde{\pi}_h(s))\,, \\
Q^{\star}_{\widetilde{\pi},\lambda,h}(s,a) &= r_h(s,a) + p_h V^{\star}_{\widetilde{\pi},\lambda,h+1}(s,a)\,, \\
V^{\star}_{\widetilde{\pi},\lambda,h}(s) &= \max_{\pi\in\Delta_A}\Big\{\pi Q^{\star}_{\widetilde{\pi},\lambda,h}(s) - \lambda\,\mathrm{KL}(\pi_h(s)\|\widetilde{\pi}_h(s))\Big\}, \\
\pi^{\star}_{\widetilde{\pi},\lambda,h}(s) &= \arg\max_{\pi\in\Delta_A}\Big\{\pi Q^{\star}_{\widetilde{\pi},\lambda,h}(s) - \lambda\,\mathrm{KL}(\pi_h(s)\|\widetilde{\pi}_h(s))\Big\}\,,
\end{aligned}
\tag{B.5}
$$

where $V^{\pi}_{\widetilde{\pi},\lambda,H+1} = V^{\star}_{\widetilde{\pi},\lambda,H+1} = 0$. Note that for π the uniform policy we recover the entropy-regularized Bellman equations.

Next we define a convex conjugate to $\lambda\,\mathrm{KL}(\cdot\|\widetilde{\pi}_h(s))$ as $F_{\widetilde{\pi}_h(s),\lambda,h}\colon \mathbb{R}^{\mathcal{A}} \to \mathbb{R}$

$$
F_{\widetilde{\pi}_h(s),\lambda,h}(x) = \max_{\pi\in\Delta(\mathcal{A})}\{\langle \pi, x\rangle - \lambda\,\mathrm{KL}(\pi\|\widetilde{\pi}_h(s))\} = \lambda \log\left(\sum_{a\in\mathcal{A}} \widetilde{\pi}_h(a|s)\exp\{x_a/\lambda\}\right).
$$

and, with a sight abuse of notation, extend the action of this function to the Q-function as follows

$$
V^{\star}_{\widetilde{\pi},\lambda,h}(s) = F_{\widetilde{\pi}_h(s),\lambda,h}(Q^{\star}_{\widetilde{\pi},\lambda,h}(s,\cdot)) = \max_{\pi\in\Delta_{\mathcal{A}}}\Big\{\pi Q^{\star}_{\widetilde{\pi},\lambda,h}(s) - \lambda\,\mathrm{KL}(\pi\|\widetilde{\pi}_h(s))\Big\}\,.
$$

Thanks to the fact that the norm of gradients of $\mathrm{KL}(\pi|\widetilde{\pi}_h(s))$ tends to infinity as π tends to a border of simplex, we have an exact formula for the optimal policy by Fenchel-Legendre transform

$$
\pi^{\star}_{\widetilde{\pi},\lambda,h}(s) = \arg\max_{\pi\in\Delta_{\mathcal{A}}}\Big\{\pi Q^{\star}_{\widetilde{\pi},\lambda,h}(s) - \lambda\,\mathrm{KL}(\pi\|\widetilde{\pi}_h(s))\Big\} = \nabla F_{\widetilde{\pi}_h(s),\lambda,h}(Q^{\star}_{\widetilde{\pi},\lambda,h}(s,\cdot))\,.
$$

Notice that we have $\nabla F_{\widetilde{\pi}_h(s),\lambda,h}(Q^{\star}_{\widetilde{\pi},\lambda,h}(s,\cdot)) \in \Delta(\mathcal{A})$ since the gradient of Φ diverges on the boundary of $\Delta(\mathcal{A})$.

Finally, it is known that the smoothness property of $F_{\widetilde{\pi}_h(s),\lambda,h}$ plays a key role in reduced sample complexity for planning in regularized MDPs (Grill et al. 2019). For our general setting we have that since $\lambda\,\mathrm{KL}(\cdot\|\widetilde{\pi}_h(s))$ is λ-strongly convex with respect to $\|\cdot\|_1$, then $F_{\widetilde{\pi}_h(s),\lambda,h}$ is $1/\lambda$-strongly smooth with respect to the dual norm $\|\cdot\|_\infty$

$$
F_{\widetilde{\pi}_h(s),\lambda,h}(x) \leqslant F_{\widetilde{\pi}_h(s),\lambda,h}(x') + \langle \nabla F_{\widetilde{\pi}_h(s),\lambda,h}(x'), x - x'\rangle + \frac{1}{2\lambda}\|x - x'\|^2_\infty\,.
$$

We notice that KL-divergence is always non-negative and, moreover, $V^{\star}_{\widetilde{\pi},\lambda,h}(s) \geqslant 0$ for any s since the value of the reference policy $\widetilde{\pi}$ is non-negative.

B.4.2 Algorithm Description

In this appendix, we present the UCBVI-Ent+ algorithm, a modification of the algorithm UCBVI-Ent proposed by Tiapkin et al. (2023a), that achieves better rates in the tabular setting. The UCBVI-Ent+ algorithm works by sampling trajectory according to an exploratory version of an optimistic solution of the regularized MDP and is characterized by the following rules.

Sampling rule. To obtain the sampling rule at episode t, we first compute a policy $\bar{\boldsymbol{\pi}}^t$ by optimistic planning in a regularized MDP,

$$\begin{aligned}
\overline{Q}_h^t(s,a) &= \mathrm{clip}\Big(r_h(s,a) + \widehat{p}_h^t \overline{V}_{h+1}^t(s,a) + b_h^{p,t}(s,a), 0, H\Big)\,, \\
\overline{V}_h^t(s) &= \max_{\pi \in \Delta_A} \Big\{\pi \overline{Q}_h^t(s) - \lambda\,\mathrm{KL}(\pi \| \widetilde{\pi}_h(s))\Big\}\,, \\
\bar{\pi}_h^{t+1}(s) &= \arg\max_{\pi \in \Delta_A} \Big\{\pi \overline{Q}_h^t(s) - \lambda\,\mathrm{KL}(\pi \| \widetilde{\pi}_h(s))\Big\}\,,
\end{aligned} \tag{B.6}$$

with $\overline{V}_{H+1}^t = 0$ by convention, where $\widehat{p}^t$ is an estimate of the transition probabilities defined in Appendix B.4.3 and $b^{p,t}$ some bonus term, defined in (B.8), Appendix B.4.4. It takes into account an estimation error for transition probabilities. Then, we define a family of policies aimed to explore actions for which Q-value is not well estimated at a particular step. That is, for $h' \in [0, H]$, the policy $\pi^{t,(h')}$ first follows the optimistic policy $\bar{\boldsymbol{\pi}}^t$ until step h where it selects an action leading to the largest confidence interval for the optimal Q-value,

$$\pi_h^{t,(h')}(a|s) = \begin{cases} \pi_h^{t,(h')}(a|s) = \bar{\pi}_h^t(a|s) & \text{if } h \neq h'\,, \\ \pi_h^{t,(h')}(a|s) = \mathbb{1}\Big\{a \in \arg\max_{a' \in \mathcal{A}} (\overline{Q}_h^t(s,a') - \underline{Q}_h^t(s,a'))\Big\} & \text{if } h = h'\,, \end{cases}$$

where $\underline{Q}^t$ is a lower bound on the optimal regularized Q-value function, see Appendix B.4.4. In particular, for $h' = 0$ we have $\pi^{t,(0)} = \bar{\boldsymbol{\pi}}^t$. The sampling rule is obtained by picking up uniformly at random one policy among the family $\pi^t = \pi^{t,(h')}$, $h' \sim \mathcal{U}\mathrm{nif}[0, H]$. Note that it is equivalent to sampling from a uniform mixture policy $\pi^{\mathrm{mix},t}$ over all $h' \in [0, H]$.

Stopping rule and decision rule. To define the stopping rule, we first recursively build an upper-bound on the difference between the value of the optimal policy and the value of the current optimistic policy $\bar{\boldsymbol{\pi}}^t$,

$$\begin{aligned}
W_h^t(s,a) &= \Big(1 + \frac{1}{H}\Big)\widehat{p}_h^t G_{h+1}^t(s) + b_h^{\mathrm{gap},t}(s,a)\,, \\
G_h^t(s) &= \mathrm{clip}\Big(\bar{\pi}_h^{t+1} W_h^t(s) + \frac{1}{2\lambda}\max_{a \in \mathcal{A}}\Big(\overline{Q}_h^t(s,a) - \underline{Q}_h^t(s,a)\Big)^2, 0, H\Big)\,,
\end{aligned} \tag{B.7}$$

where $b_h^{\mathrm{gap},t}$ is a bonus defined in (B.10), Appendix B.4.4, $\underline{V}^t$ is a lower-bound on the optimal value function defined in Appendix B.4.4 and $G_{H+1}^t = 0$ by convention. Then the stopping time $\mathfrak{t} = \inf\{t \in \mathbb{N} : G_1^t(s_1) \leqslant \varepsilon\}$ corresponds to the first episode when this upper-bound is smaller than ε. At this episode, we return the policy $\widehat{\pi} = \boldsymbol{\pi}^{\mathfrak{t}}$.

The complete procedure is described in Algorithm 8.

B.4.3 Concentration Events

We first define an estimate of the transition kernel. The number of times the state action-pair (s,a) was visited in step h in the first t episodes are $n_h^t(s,a) \triangleq \sum_{i=1}^t \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\}$. Let

Algorithm 8 UCBVI-Ent+

Input: Target precision ε, target probability δ, bonus functions $b^t, b^{t,\mathrm{KL}}$.
while true **do**
  Compute $\bar{\boldsymbol{\pi}}^t$ by optimistic planning with (B.6).
  Compute bound on the gap $G_1^t(s,a)$ with (B.7).
  if $G_1^t(s_1) \leqslant \varepsilon$ **then break**
  Sample $h' \sim \mathcal{U}\mathrm{nif}[H]$ and set $\pi^t = \pi^{t,(h')}$.
  for $h \in [H]$ **do**
    Play $a_h^t \sim \pi_h^t(s_h^t)$
    Observe $s_{h+1}^t \sim p_h(s_h^t, a_h^t)$
  end for
  Update transition estimates $\widehat{p}^t$.
end while
Output policy $\widehat{\pi} = \bar{\boldsymbol{\pi}}^t$.

$n_h^t(s'|s,a) \triangleq \sum_{i=1}^t \mathbb{1}\left\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\right\}$ be the number of transitions from s to s' at step h. The empirical distribution is defined as $\widehat{p}_h^t(s'|s,a) = n_h^t(s'|s,a)/n_h^t(s,a)$ if $n_h^t(s,a) > 0$ and $\widehat{p}_h^t(s'|s,a) \triangleq 1/A$ for all $s' \in \mathcal{S}$ else.

Following the ideas of Ménard et al. (2021), we define the following concentration events. First, we define pseudo-counts as a sum of conditional expectations of the random variables that correspond to counts

$$\overline{n}_h^t(s,a) = \sum_{i=1}^t \widetilde{d}_h^{\pi^t}(s,a)\,,$$

where $\pi^{\mathrm{mix},t}$ is a mixture policy played on t-th step. In particular, we have

$$d_h^{\pi^{\mathrm{mix},t}}(s,a) = \frac{1}{H+1}\sum_{h'=0}^{H} d_h^{\pi^{t,(h')}}(s,a)\,,$$

where $\pi_h^{t,(h')}(s,a)$ is a greedy-modified policy $\bar{\boldsymbol{\pi}}^t$ in h'-step:

$$\pi_h^{t,(h')}(a|s) = \begin{cases} \pi_h^{t,(h')}(a|s) = \bar{\pi}_h^t(a|s) & \text{if } h \neq h' \\ \pi_h^{t,(h')}(a|s) = \mathbb{1}\left\{a = \arg\max_{a' \in \mathcal{A}} (\overline{Q}_h^t(s,a') - \underline{Q}_h^t(s,a'))\right\} & \text{if } h = h' \end{cases},$$

Let $\beta^{\mathrm{KL}}, \beta^{\mathrm{cnt}} : (0,1) \times \mathbb{N} \to \mathbb{R}_+$ be some functions defined later on in Lemma B.16. We define the following favorable events

$$\mathcal{E}^{\mathrm{KL}}(\delta) \triangleq \left\{\forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S}\times\mathcal{A} :\ \mathrm{KL}(\widehat{p}_h^t(s,a)\|p_h(s,a)) \leqslant \frac{\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}\right\},$$

$$\mathcal{E}^{\mathrm{cnt}}(\delta) \triangleq \left\{\forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S}\times\mathcal{A} :\ n_h^t(s,a) \geqslant \frac{1}{2}\overline{n}_h^t(s,a) - \beta^{\mathrm{cnt}}(\delta)\right\},$$

We also introduce an intersection of these events of interest, $\mathcal{G}(\delta) \triangleq \mathcal{E}^{\mathrm{KL}}(\delta) \cap \mathcal{E}^{\mathrm{cnt}}(\delta)$. We prove that for the right choice of the functions $\beta^{\mathrm{KL}}, \beta^{\mathrm{cnt}}$, the above events hold with high probability.

Lemma B.16. *For any $\delta \in (0,1)$ and for the following choices of functions β,*

$$\beta^{\mathrm{KL}}(\delta, n) \triangleq \log(2SAH/\delta) + S\log(\mathrm{e}(1+n))\,,$$
$$\beta^{\mathrm{cnt}}(\delta) \triangleq \log(2SAH/\delta)\,,$$

it holds that

$$\mathbb{P}[\mathcal{E}^{\mathrm{KL}}(\delta)] \geqslant 1-\delta/2, \quad \mathbb{P}[\mathcal{E}^{\mathrm{cnt}}(\delta)] \geqslant 1-\delta/2,$$

In particular, $\mathbb{P}[\mathcal{G}(\delta)] \geqslant 1-\delta$.

Proof. Applying Theorem D.13 and the union bound over $h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$ we get $\mathbb{P}[\mathcal{E}^{\mathrm{KL}}(\delta)] \geqslant 1-\delta/2$.

By Theorem D.14 and union bound, $\mathbb{P}[\mathcal{E}^{\mathrm{cnt}}(\delta)] \geqslant 1-\delta/2$. The union bound over three prescribed events concludes $\mathbb{P}[\mathcal{G}_H(\delta)] \geqslant 1-\delta$ and $\mathbb{P}[\mathcal{G}_B(\delta)] \geqslant 1-\delta$. □

Lemma B.17. *Assume conditions of Lemma B.16. Then on event* $\mathcal{E}^{\mathrm{KL}}(\delta)$*, for any* $f\colon \mathcal{S} \to [0,H], t \in \mathbb{N}, h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$

$$[p_h - \widehat{p}_h^t]f(s,a) \leqslant \sqrt{\frac{2H^2\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}}\,.$$

Proof. By a Hölder and Pinsker inequalities

$$[p_h - \widehat{p}_h^t]f(s,a) \leqslant H\|p_h(s,a) - \widehat{p}_h^t(s,a)\|_1 \leqslant H\sqrt{2\,\mathrm{KL}(\widehat{p}_h^t(s,a)\|p_h(s,a))}\,.$$

We conclude the statement by applying the definition of the event $\mathcal{E}^{\mathrm{KL}}$. □

Lemma B.18. *Assume conditions of Lemma B.16. Then on event* $\mathcal{E}^{\mathrm{KL}}(\delta)$*, for any* $f\colon \mathcal{S} \to [0,H]$*,* $t \in \mathbb{N}, h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$*,*

$$[p_h - \widehat{p}_h^t]f(s,a) \leqslant \frac{1}{H}\widehat{p}_h^t f(s,a) + H\left(\frac{5H\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \wedge 1\right),$$
$$[\widehat{p}_h^t - p_h]f(s,a) \leqslant \frac{1}{H}p_h f(s,a) + H\left(\frac{5H\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \wedge 1\right).$$

Proof. Let us start from the first statement. We apply Lemma D.9 to a function $g = H - f$ and Lemma D.10 to obtain

$$\begin{aligned}[\widehat{p}_h^t - p_h]g(s,a) = [p_h - \widehat{p}_h^t]f(s,a) &\leqslant \sqrt{2\mathrm{Var}_{p_h}[f](s,a)\cdot\mathrm{KL}(\widehat{p}_h^t\|p_h)} + \frac{H}{3}\,\mathrm{KL}(\widehat{p}_h^t\|p_h)\\ &\leqslant 2\sqrt{\mathrm{Var}_{\widehat{p}_h^t}[f](s,a)\cdot\mathrm{KL}(\widehat{p}_h^t\|p_h)} + 4H\,\mathrm{KL}(\widehat{p}_h^t\|p_h)\,.\end{aligned}$$

Since $0 \leqslant f(s) \leqslant H$ we get

$$\mathrm{Var}_{\widehat{p}_h^t}[f](s,a) \leqslant \widehat{p}_h^t[f^2](s,a) \leqslant H\cdot\widehat{p}_h^t f(s,a)\,.$$

Finally, applying $2\sqrt{ab} \leqslant a+b, a, b \geqslant 0$, we obtain the following inequality

$$(p_h - \widehat{p}_h^t)f(s,a) \leqslant \frac{1}{H}\widehat{p}_h^t f(s,a) + 5H^2\,\mathrm{KL}(\widehat{p}_h^t\|p_h)\,.$$

The definition of $\mathcal{E}^{\mathrm{KL}}(\delta)$ implies the first part of the statement. At the same time we have a trivial bound since $f(s) \in [0,H]$

$$[p_h - \widehat{p}_h^t]f(s,a) \leqslant H \leqslant \frac{1}{H}\widehat{p}_h^t f(s,a) + H\,.$$

The second statement holds by the same inequalities. □

B.4.4 Confidence Intervals

Similar to Azar et al. (2017), Zanette and Brunskill (2019), and Ménard et al. (2021), we define the upper confidence bound for the optimal regularized Q-function with Hoeffding bonuses.

Then we have the following sequences defined as follows

$$\begin{aligned}
\overline{Q}_h^t(s,a) &= \operatorname{clip}\Big(r_h(s,a) + \widehat{p}_h^t \overline{V}_{h+1}^t(s,a) + b_h^{p,t}(s,a), 0, H\Big)\,, \\
\boldsymbol{\pi}_h^{t+1}(s) &= \max_{\pi \in \Delta_{\mathcal{A}}} \{\pi \overline{Q}_h^t(s) - \lambda \operatorname{KL}(\pi \| \widetilde{\pi}_h(s))\}\,, \\
\overline{V}_h^t(s) &= \bar{\pi}_h^{t+1} \overline{Q}_h^t(s) - \lambda \operatorname{KL}(\bar{\pi}_h^{t+1}(s) \| \widetilde{\pi}_h(s))\,, \\
\overline{V}_{H+1}^t(s) &= 0\,,
\end{aligned}$$

and the lower confidence bound as follows

$$\begin{aligned}
\underline{Q}_h^t(s,a) &= \operatorname{clip}\Big(r_h(s,a) + \widehat{p}_h^t \underline{V}_h^t(s,a) - b_h^{p,t}(s,a), 0, H\Big) \\
\underline{V}_h^t(s) &= \max_{\pi \in \Delta_{\mathcal{A}}} \{\pi \underline{Q}_h^t(s) - \lambda \operatorname{KL}(\pi \| \widetilde{\pi}_h(s))\}\,, \\
\underline{V}_{H+1}^t(s) &= 0\,,
\end{aligned}$$

where we have two types of transition bonuses that will be specified before use. The Hoeffding bonuses are defined as follows

$$b_h^{p,t}(s,a) \triangleq \sqrt{\frac{2H^2 \beta^{\mathrm{KL}}(\delta, n_h^t(s,a))]}{n_h^t(s,a)}}\,. \tag{B.8}$$

Proposition B.19. *Let $\delta \in (0,1)$. Assume Hoeffding bonuses (B.8). Then on event $\mathcal{G}(\delta)$ for any $t \in \mathbb{N}$, $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$ it holds*

$$\underline{Q}_h^t(s,a) \leqslant Q_{\widetilde{\pi},\lambda,h}^{\star}(s,a) \leqslant \overline{Q}_h^t(s,a)\,, \qquad \underline{V}_{\lambda,h}^t(s) \leqslant V_{\widetilde{\pi},\lambda,h}^{\star}(s) \leqslant \overline{V}_h^t(s)\,.$$

Proof. Proceed by induction over h. For $h = H+1$ the statement is trivial. Now we assume that inequality holds for any $h' > h$ for a fixed $h \in [H]$. Fix a timestamp $t \in \mathbb{N}$ and a state-action pair (s,a) and assume that $\overline{Q}_h^t(s,a) < H$, i.e., no clipping occurs. Otherwise the inequality $Q_{\widetilde{\pi},\lambda,h}^{\star}(s,a) \leqslant \overline{Q}_h^t(s,a)$ is trivial. In particular, it implies $n_h^t(s,a) > 0$.

In this case by Bellman equations (B.5) we have

$$[\overline{Q}_h^t - Q_{\widetilde{\pi},\lambda,h}^{\star}](s,a) = \widehat{p}_h^t \overline{V}_{h+1}^t(s,a) - p_h V_{\widetilde{\pi},\lambda,h+1}^{\star}(s,a) + b_h^{p,t}(s,a)\,.$$

To show that the right-hand side is non-negative, we start from the induction hypothesis

$$[\overline{Q}_h^t - Q_{\widetilde{\pi},\lambda,h}^{\star}](s,a) \geqslant [\widehat{p}_h^t - p_h] V_{\widetilde{\pi},\lambda,h+1}^{\star}(s,a) + b_h^{p,t}(s,a)\,.$$

The non-negativity of the expression above automatically holds from Lemma B.17. To prove the second inequality on Q-value, we proceed exactly the same.

Finally, we have to show the inequality for V-values. To do it, we use the fact that V-value are computed by $F_{\widetilde{\pi}_h(s),\lambda,h}$ applied to Q-value

$$\underline{V}_{\lambda,h}^t(s) = F_{\widetilde{\pi}_h(s),\lambda,h}(\underline{Q}_h^t)(s),\ V_{\widetilde{\pi},\lambda,h}^{\star}(s) = F_{\widetilde{\pi}_h(s),\lambda,h}(Q_{\widetilde{\pi},\lambda,h}^{\star})(s),\ \overline{V}_{\lambda,h}^t(s) = F_{\widetilde{\pi}_h(s),\lambda,h}(\overline{Q}_h^t)(s)\,.$$

Notice that $\nabla F_{\widetilde{\pi}_h(s),\lambda,h}$ takes values in a probability simplex; thus, all partial derivatives of $F_{\widetilde{\pi}_h(s),\lambda,h}$ are non-negative and therefore $F_{\widetilde{\pi}_h(s),\lambda,h}$ is monotone in each coordinate. Thus, since $\underline{Q}_h^t(s,a) \leqslant Q_{\widetilde{\pi},\lambda,h}^{\star}(s,a) \leqslant \overline{Q}_h^t(s,a)$, we have the same inequality $\underline{V}_h^t(s) \leqslant V_{\widetilde{\pi},\lambda,h}^{\star}(s) \leqslant \overline{V}_h^t(s)$. □

B.4.5 Sample Complexity Bounds

In this section, we provide guarantees for the regularization-aware gap that highly depends on the parameter λ.

Let us recall the regularization-aware gap that is defined recursively, starting from $G^t_{H+1} \triangleq 0$ and

$$\begin{aligned} W^t_h(s,a) &= \Big(1 + \frac{1}{H}\Big)\widehat{p}^t_h G^t_{h+1}(s) + b^{\mathrm{gap},t}_h(s,a)\,, \\ G^t_h(s) &= \mathrm{clip}\bigg(\bar{\pi}^{t+1}_h W^t_h(s) + \frac{1}{2\lambda}\max_{a\in\mathcal{A}}\Big(\overline{Q}^t_h(s,a) - \underline{Q}^t_h(s,a)\Big)^2, 0, H\bigg)\,, \end{aligned} \tag{B.9}$$

where the additional bonus is defined as

$$b^{\mathrm{gap},t}_h(s,a) \triangleq \frac{5H^2\beta^{\mathrm{KL}}(\delta, n^t_h(s,a))}{n^t_h(s,a)} \wedge H\,, \tag{B.10}$$

and the corresponding stopping time for the algorithm

$$\mathfrak{t} = \inf\{t \in \mathbb{N} : G^t_1(s_1) \leqslant \varepsilon\}\,. \tag{B.11}$$

The next lemma justifies this choice of the stopping rule.

Lemma B.20. *Assume the choice of Hoeffding bonuses (B.8) and let the event $\mathcal{G}(\delta)$ defined in Lemma B.16 holds. Then for any $t \in \mathbb{N}$, $s \in \mathcal{S}, h \in [H]$*

$$V^{\star}_{\widetilde{\pi},\lambda,h}(s) - V^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h}(s) \leqslant G^t_h(s)\,.$$

Proof. Let us proceed by induction. For $h = H+1$ the statement is trivial. Assume that for any $h' > h$ the statement holds. Also, assume that $G^t_h(s) < H$; otherwise, the inequality on the policy error holds trivially. In particular, it holds that $n^t_h(s,a) > 0$ for all $a \in \mathcal{A}$.

We can start analysis from understanding the policy error by applying the smoothness of $F_{\widetilde{\pi}_h(s),\lambda,h}$.

$$\begin{aligned} V^{\star}_{\widetilde{\pi},\lambda,h}(s) - V^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h}(s) &= F_{\widetilde{\pi}_h(s),\lambda,h}(Q^{\star}_{\widetilde{\pi},\lambda,h}(s,\cdot)) - \Big(\bar{\pi}^{t+1}_h Q^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h}(s) - \lambda\,\mathrm{KL}(\bar{\pi}^{t+1}_h(s)\|\widetilde{\pi}_h(s))\Big) \\ &\leqslant F_{\widetilde{\pi}_h(s),\lambda,h}(\overline{Q}^t_h(s,\cdot)) + \langle \nabla F_{\widetilde{\pi}_h(s),\lambda,h}(\overline{Q}^t_h(s,\cdot)), Q^{\star}_{\widetilde{\pi},\lambda,h}(s,\cdot) - \overline{Q}^t_h(s,\cdot)\rangle \\ &\quad + \frac{1}{2\lambda}\|\overline{Q}^t_h - Q^{\star}_{\widetilde{\pi},\lambda,h}\|^2_{\infty}(s) - \Big(\bar{\pi}^{t+1}_h Q^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h}(s) - \lambda\,\mathrm{KL}(\bar{\pi}^{t+1}_h(s)\|\widetilde{\pi}_h(s))\Big)\,. \end{aligned}$$

Next we recall that

$$\bar{\pi}^{t+1}_h(s) = \nabla F_{\widetilde{\pi}_h(s),\lambda,h}(\overline{Q}^t_h(s,\cdot)), \quad F_{\widetilde{\pi}_h(s),\lambda,h}(\overline{Q}^t_h)(s) = \bar{\pi}^{t+1}_h \overline{Q}^t_h(s) - \lambda\,\mathrm{KL}(\bar{\pi}^{t+1}_h(s)\|\widetilde{\pi}_h(s))\,,$$

thus we have

$$F_{\widetilde{\pi}_h(s),\lambda,h}(\overline{Q}^t_h)(s) - \Big(\bar{\pi}^{t+1}_h Q^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h}(s,\cdot) - \lambda\,\mathrm{KL}(\bar{\pi}^{t+1}_h(s)\|\widetilde{\pi}_h(s))\Big) = \bar{\pi}^{t+1}_h[\overline{Q}^t_h - Q^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h}](s)$$

and, by Bellman equations

$$\begin{aligned} V^{\star}_{\widetilde{\pi},\lambda,h}(s) - V^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h}(s) &\leqslant \bar{\pi}^{t+1}_h\Big[Q^{\star}_{\widetilde{\pi},\lambda,h} - Q^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h}\Big](s) + \frac{1}{2\lambda}\|\overline{Q}^t_h - Q^{\star}_{\widetilde{\pi},\lambda,h}\|^2_{*}(s) \\ &\leqslant \bar{\pi}^{t+1}_h p_h\Big[V^{\star}_{\widetilde{\pi},\lambda,h+1} - V^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h+1}\Big](s) + \frac{1}{2\lambda}\|\overline{Q}^t_h - Q^{\star}_{\widetilde{\pi},\lambda,h}\|^2_{*}(s)\,. \end{aligned}$$

By induction hypothesis we have

$$V^{\star}_{\widetilde{\pi},\lambda,h}(s) - V^{\boldsymbol{\pi}^{t+1}}_{\lambda,h}(s) \leqslant \bar{\pi}_h^{t+1} p_h G^t_{\lambda,h+1}(s) + \frac{1}{2\lambda}\|\overline{Q}^t_h - Q^{\star}_{\widetilde{\pi},\lambda,h}\|^2_*(s)\,.$$

Next, we apply Lemma B.18

$$\begin{aligned} p_h G_{\lambda,h+1}(s,a) &= \widehat{p}_h^t G_{\lambda,h+1}(s,a) + [p_h - \widehat{p}_h^t] G^t_{\lambda,h+1}(s,a) \\ &\leqslant \Big(1 + \frac{1}{H}\Big)\widehat{p}_h^t G^t_{\lambda,h+1}(s,a) + \frac{5H^2\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \wedge H \triangleq W_h^t(s,a)\,, \end{aligned}$$

thus

$$V^{\star}_{\widetilde{\pi},\lambda,h}(s) - V^{\boldsymbol{\pi}^{t+1}}_{\lambda,h}(s) \leqslant \bar{\pi}_h^{t+1} W_h^t(s) + \frac{1}{2\lambda}\|\overline{Q}^t_h - Q^{\star}_{\widetilde{\pi},\lambda,h}\|^2_\infty(s)\,.$$

Finally, by the definition of $\|\cdot\|_\infty$ and Proposition B.19

$$V^{\star}_{\widetilde{\pi},\lambda,h}(s) - V^{\boldsymbol{\pi}^{t+1}}_{\lambda,h}(s) \leqslant \bar{\pi}_h^{t+1} W_h^t(s) + \frac{1}{2\lambda}\max_{a\in\mathcal{A}}\big(\overline{Q}^t_h(s,a) - \underline{Q}^t_h(s,a)\big)^2 \triangleq G_h^t(s)\,.$$

□

Theorem B.21. *Let $\varepsilon > 0$, $\delta \in (0,1)$, $S \geqslant 2$ and $\lambda \leqslant H$. Then* UCBVI-Ent+ *algorithm with Hoeffding bonuses and a stopping rule $\mathfrak{t}$ (B.11) is (ε,δ)-PAC for the best policy identification in regularized MDPs.*

Moreover, the stopping time $\mathfrak{t}$ is bounded as follows

$$\mathfrak{t} = \mathcal{O}\left(\frac{H^5 SA \cdot (\log(SAH/\delta) + SL)\cdot L}{\varepsilon\lambda}\right),$$

where $L = \mathcal{O}(\log(SAH\log(1/\delta)/(\varepsilon\lambda)))$.

Proof. To show that UCBVI-Ent+ is (ε,δ)-PAC we notice that on event $\mathcal{G}(\delta)$ for $\widehat{\pi} = \pi^{\mathfrak{t}}$ by Lemma B.20

$$V^{\star}_{\widetilde{\pi},\lambda,1}(s_1) - V^{\widehat{\pi}}_{\widetilde{\pi},\lambda,1}(s_1) \leqslant G^{\mathfrak{t}}_1(s_1) \leqslant \varepsilon\,,$$

and the event $\mathcal{G}(\delta)$ holds with probability at least $1-\delta$. Next, we show that the sample complexity is bounded by the abovementioned quantity.

Step 1. Bound for $G^t_1(s_1)$ First, we start from bounding $W^t_h(s,a)$ and $G^t_h(s)$. By Lemma B.18 we can define the following upper bound for $W^t_h(s,a)$

$$W_h^t(s,a) \leqslant \Big(1 + \frac{2}{H}\Big) p_h G^t_{h+1}(s,a) + \frac{10H^2\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \wedge 2H\,.$$

Therefore we obtain

$$\begin{aligned} G_h^t(s) \leqslant \mathbb{E}_{\boldsymbol{\pi}^{t+1}}\Bigg[&\Big(1 + \frac{2}{H}\Big) G^t_{h+1}(s_{h+1}) + \frac{10H^2\beta^{\mathrm{KL}}(\delta, n_h^t(s_h,a_h))}{n_h^t(s_h,a_h)} \\ &+ \frac{1}{2\lambda}\max_{a\in\mathcal{A}}\big(\overline{Q}^t_h(s_h,a) - \underline{Q}^t_h(s_h,a)\big)^2 \Big| s_h = s\Bigg]\,, \end{aligned}$$

By rolling out this expression

$$\begin{aligned} G_1^t(s_1) \leqslant \mathbb{E}_{\boldsymbol{\pi}^{t+1}}\Bigg[&\sum_{h=1}^{H}\Big(1 + \frac{2}{H}\Big)^h \left(\frac{10H^2\beta^{\mathrm{KL}}(\delta, n_h^t(s_h,a_h))}{n_h^t(s_h,a_h)} \wedge 2H\right) \\ &+ \Big(1 + \frac{2}{H}\Big)^h \frac{1}{2\lambda}\max_{a\in\mathcal{A}}\big(\overline{Q}^t_h(s_h,a) - \underline{Q}^t_h(s_h,a)\big)^2\Bigg]\,. \end{aligned}$$

Using the fact that $(1+2/H)^h \leqslant \mathrm{e}^2$, we have

$$G_1^t(s_1) \leqslant \underbrace{10\mathrm{e}^2 H^2 \mathbb{E}_{\boldsymbol{\pi}^{t+1}}\left[\sum_{h=1}^{H}\left(\frac{\beta^{\mathrm{KL}}(\delta, n_h^t(s_h,a_h))}{n_h^t(s_h,a_h)} \wedge 1\right)\right]}_{(\mathbf{A})} + \underbrace{\frac{\mathrm{e}^2}{2\lambda}\mathbb{E}_{\boldsymbol{\pi}^{t+1}}\left[\sum_{h=1}^{H}\max_{a\in\mathcal{A}}\left(\overline{Q}_h^t(s_h,a) - \underline{Q}_h^t(s_h,a)\right)^2\right]}_{(\mathbf{B})}.$$

Term (A). The analysis of the term (**A**) follows Ménard et al. (2021): we switch counts to pseudo-counts by Lemma D.5 and obtain

$$(\mathbf{A}) \leqslant 40\mathrm{e}^2 H^2 \sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\boldsymbol{\pi}^{t+1}}(s,a)\frac{\beta^{\mathrm{KL}}(\delta, \overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a)\vee 1}.$$

Term (B). For this term we analyze each summand over h separately. By Lemma D.12

$$\mathbb{E}_{\boldsymbol{\pi}^{t+1}}\left[\max_{a\in\mathcal{A}}\left(\overline{Q}_h^t(s_h,a) - \underline{Q}_h^t(s_h,a)\right)^2\right] = \mathbb{E}_{\pi^{t+1,(h)}}\left[\left(\overline{Q}_h^t(s_h,a_h) - \underline{Q}_h^t(s_h,a_h)\right)^2\right].$$

Next, we analyze the expression under the square. First, we have

$$\overline{Q}_h^t(s_h,a_h) - \underline{Q}_h^t(s_h,a_h) \leqslant 2b_h^{p,t}(s_h,a_h) + \widehat{p}_h^t[\overline{V}_{h+1}^t - \underline{V}_{h+1}^t](s_h,a_h).$$

By Lemma B.17

$$\overline{Q}_h^t(s_h,a_h) - \underline{Q}_h^t(s_h,a_h) \leqslant 4b_h^{p,t}(s_h,a_h) + p_h[\overline{V}_{h+1}^t - \underline{V}_{h+1}^t](s_h,a_h).$$

using $\overline{V}_{\lambda,h+1}^t(s) - \underline{V}_{\lambda,h+1}^t(s) \leqslant \bar{\pi}_{h+1}^t\left[\overline{Q}_{h+1}^t - \underline{Q}_{h+1}^t\right](s)$ and the definition of Hoeffding bonuses (B.8), thus, rolling out this recursion

$$\overline{Q}_h^t(s_h,a_h) - \underline{Q}_h^t(s_h,a_h) \leqslant 5H\cdot\mathbb{E}_{\boldsymbol{\pi}^{t+1}}\left[\sum_{h'=h}^{H}\sqrt{\frac{2\beta^{\mathrm{KL}}(\delta, n_{h'}^t(s_{h'},a_{h'}))}{n_{h'}^t(s_{h'},a_{h'})}\wedge 1}\middle| s_h\right].$$

By Lemma D.5, Jensen inequality, and a change of policy $\boldsymbol{\pi}^{t+1}$ to $\pi^{t+1,(h)}$ by Lemma D.12 we have

$$\overline{Q}_h^t(s_h,a_h) - \underline{Q}_h^t(s_h,a_h) \leqslant 20H^{3/2}\sqrt{\mathbb{E}_{\pi^{t+1,(h)}}\left[\sum_{h'=h}^{H}\frac{2\beta^{\mathrm{KL}}(\delta, \overline{n}_{h'}^t(s_{h'},a_{h'}))}{\overline{n}_{h'}^t(s_{h'},a_{h'})\vee 1}|s_h\right]}.$$

By taking the square, we get

$$\mathbb{E}_{\boldsymbol{\pi}^{t+1}}\left[\max_{a\in\mathcal{A}}\left(\overline{Q}_h^t(s_h,a) - \underline{Q}_h^t(s_h,a)\right)^2\right] \leqslant 400H^3\mathbb{E}_{\pi^{t+1,(h)}}\left[\mathbb{E}_{\pi^{t+1,(h)}}\left[\sum_{h'=h}^{H}\frac{2\beta^{\mathrm{KL}}(\delta, \overline{n}_{h'}^t(s_{h'},a_{h'}))}{\overline{n}_{h'}^t(s_{h'},a_{h'})\vee 1}\middle| s_h\right]\right].$$

The telescoping property of conditional expectation yields the final bound

$$\mathbb{E}_{\boldsymbol{\pi}^{t+1}}\left[\max_{a\in\mathcal{A}}\left(\overline{Q}_h^t(s_h,a) - \underline{Q}_h^t(s_h,a)\right)^2\right] \leqslant \sum_{h'=1}^{H} 400H^3\mathbb{E}_{\pi^{t+1,(h)}}\left[\frac{2\beta^{\mathrm{KL}}(\delta, \overline{n}_{h'}^t(s_{h'},a_{h'}))}{\overline{n}_{h'}^t(s_{h'},a_{h'})\vee 1}\right].$$

Finally, collecting bounds over all $h \in [H]$ we have

$$(\mathbf{B}) \leqslant \frac{200\mathrm{e}^2 H^3}{\lambda} \sum_{h=1}^{H} \sum_{s,a} \sum_{h'=1}^{H} d_{h'}^{\pi^{t,(h)}}(s,a) \frac{\beta^{\mathrm{KL}}(\delta, \overline{n}_{h'}^t(s,a))}{\overline{n}_{h'}^t(s,a) \vee 1} \,.$$

The final bound for an initial gap follows

$$\begin{aligned} G_1^t(s_1) \leqslant 40\mathrm{e}^2 H^2 \sum_{h'=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{h'}^{\pi^{t+1}}(s,a) \frac{\beta^{\mathrm{KL}}(\delta, \overline{n}_{h'}^t(s,a))}{\overline{n}_{h'}^t(s,a) \vee 1} \\ + \frac{200\mathrm{e}^2 H^3}{\lambda} \sum_{h=1}^{H} \sum_{s,a} \sum_{h'=1}^{H} d_{h'}^{\pi^{t,(h)}}(s,a) \frac{\beta^{\mathrm{KL}}(\delta, \overline{n}_{h'}^t(s,a))}{\overline{n}_{h'}^t(s,a) \vee 1} \,. \end{aligned}$$

Since $\lambda \leqslant H$, we have that $H^2 \leqslant H^3/\lambda$. Using a convention $d_{h'}^{\pi^{t+1}}(s,a) = d_{h'}^{\pi^{t+1,(0)}}(s,a)$ we have

$$G_1^t(s_1) \leqslant \frac{240\mathrm{e}^2 H^3}{\lambda} \sum_{h=0}^{H} \sum_{s,a} \sum_{h'=1}^{H} d_{h'}^{\pi^{t,(h)}}(s,a) \frac{\beta^{\mathrm{KL}}(\delta, \overline{n}_{h'}^t(s,a))}{\overline{n}_{h'}^t(s,a) \vee 1} \,.$$

By changing the summation order and noticing that

$$d_{h'}^{\pi^{\mathrm{mix},t}}(s,a) = \frac{1}{H+1} \sum_{h=0}^{H} d_h^{\pi^{t,(h)}}(s,a)$$

for $H+1 \leqslant 2H$ we get

$$G_1^t(s_1) \leqslant \frac{480\mathrm{e}^2 H^4}{\lambda} \sum_{s,a} \sum_{h=1}^{H} d_h^{\pi^{\mathrm{mix},t}}(s,a) \frac{\beta^{\mathrm{KL}}(\delta, \overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a) \vee 1} \,.$$

Step 2. Sum over $t < \mathfrak{t}$. Assume $\mathfrak{t} > 0$. In the case $\mathfrak{t} = 0$, the bound is trivially true. Notice that for any $t < \mathfrak{t}$ we have

$$G_{\lambda,1}^t(s_1) > \varepsilon \,,$$

thus, summing upper bounds on $G_{\lambda,1}^t(s_1)$ over all $t < \mathfrak{t}$ we have

$$\varepsilon(\mathfrak{t}-1) < \sum_{t=1}^{\mathfrak{t}-1} G_{\lambda,1}^t(s_1) \leqslant \frac{480\mathrm{e}^2 H^4}{\lambda} \sum_{(s,a,h)} \sum_{t=1}^{\mathfrak{t}-1} d_h^{\pi^{\mathrm{mix},t}}(s,a) \frac{\beta^{\mathrm{KL}}(\delta, \overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a) \vee 1} \,.$$

Notice that $\beta^{\mathrm{KL}}(\delta, \cdot)$ is monotone and maximizes at $\mathfrak{t}-1$, and $d_h^{\pi^{\mathrm{mix},t+1}}(s,a) = \overline{n}_h^{t+1}(s,a) - \overline{n}_h^t(s,a)$. Thus, applying Lemma D.6, we have

$$\varepsilon(\mathfrak{t}-1) < \frac{1920\mathrm{e}^2 H^5 SA}{\lambda} \beta^{\mathrm{KL}}(\delta, \mathfrak{t}-1) \log(\mathfrak{t}) \,.$$

Then by definition of β^{KL}

$$\varepsilon(\mathfrak{t}-1) \leqslant \frac{1920\mathrm{e}^4 H^5 SA}{\lambda} \cdot (\log(2SAH/\delta) + S\log(\mathrm{e}\mathfrak{t})) \cdot \log(\mathfrak{t}) \,.$$

Step 3. Solving the recurrence. Define $A = 1920\mathrm{e}^2 H^5 SA/(\lambda\varepsilon)$ and $B = \log(2SAH/\delta) + S$. Our goal is to provide an upper bound for solutions to the following inequality

$$\mathfrak{t} \leqslant 1 + A(S\log(\mathfrak{t}) + B) \cdot \log(\mathfrak{t}) \,.$$

First, we obtain a loose solution by using inequality $\log(\mathfrak{t}) \leqslant \mathfrak{t}^{\beta}/\beta$ that holds for any $\mathfrak{t} \geqslant 1$. Taking $\beta = 1/3$ we have

$$\mathfrak{t} \leqslant 1 + 3A(3S \cdot \mathfrak{t}^{1/3} + B) \cdot \mathfrak{t}^{1/3} .$$

Also we may assume that $\mathfrak{t} \geqslant 2$, thus $1 \leqslant \mathfrak{t}/2$ and we achieve

$$\mathfrak{t}^{2/3} \leqslant 6A(3S\mathfrak{t}^{1/3} + B) .$$

Solving this quadratic inequality in $\mathfrak{t}^{1/3}$, we have

$$\mathfrak{t} \leqslant \left(\frac{18AS + \sqrt{(18AS)^2 + 24AB}}{2} \right)^3 \leqslant \left(18AS + \sqrt{24AB} \right)^3 .$$

Define $L = 3\log\left(54AS + \sqrt{18AB}\right)$. Then, we can easily upper bound the initial inequality as follows

$$\mathfrak{t} \leqslant 1 + A(B + SL)L .$$

□

B.5 Best Policy Identification in Regularized Linear MDPs

In this appendix, we first state some useful properties of regularized linear MDPs and describe the LSVI-UCB-Ent algorithm.

B.5.1 General Properties of Linear MDPs

Let us start with a description of how to generalize the techniques of regularized MDPs to the setup of linear function approximation (Jin et al. 2020b). Again, we consider the case of KL-regularized MDPs with respect to a reference policy $\widetilde{\pi}$. In this setting, the Q- and V-values could be defined through regularized Bellman equations

$$\begin{aligned} Q^{\pi}_{\widetilde{\pi},\lambda,h}(s,a) &= r_h(s,a) + p_h V^{\pi}_{\widetilde{\pi},\lambda,h+1}(s,a), \\ V^{\pi}_{\widetilde{\pi},\lambda,h}(s) &= \pi_h Q^{\pi}_{\widetilde{\pi},\lambda,h}(s) - \lambda \operatorname{KL}(\pi_h(s) \| \widetilde{\pi}_h(s)) \,. \end{aligned}$$

Moreover, for optimal Q- and V-functions we have

$$\begin{aligned} Q^{\star}_{\widetilde{\pi},\lambda,h}(s,a) &= r_h(s,a) + p_h V^{\star}_{\widetilde{\pi},\lambda,h+1}(s,a), \\ V^{\star}_{\widetilde{\pi},\lambda,h}(s) &= \max_{\pi \in \Delta_{\mathcal{A}}} \left\{ \pi Q^{\star}_{\widetilde{\pi},\lambda,h}(s) - \lambda \operatorname{KL}(\pi \| \widetilde{\pi}_h(s)) \right\}. \end{aligned}$$

Note that the value of a policy could be arbitrarily negative, however, we know a priori that the optimal policy has non-negative value $V^{\star}_{\widetilde{\pi},\lambda,h} \in [0, H]$, since the policy $\widetilde{\pi}$ itself has non-negative value.

In particular, under this assumption, we have the following simple proposition

Proposition B.22. *For a linear MDP, for any policy π such that $V^{\pi}_{\widetilde{\pi},\lambda,h}(s,a) \geqslant 0$ for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ there exists weights $\{w^{\pi}_h\}_{h \in [H]}$ such that for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ we have $Q^{\pi}_{\widetilde{\pi},\lambda,h}(s,a) = \psi(s,a)^{\intercal} w^{\pi}_h$. Moreover, for any $h \in [H]$ it holds $\|w^{\pi}_h\|_2 \leqslant 2H\sqrt{d}$.*

Proof. By Bellman equations

$$\begin{aligned} Q^{\pi}_{\widetilde{\pi},\lambda,h}(s,a) = r_h(s,a) + p_h V^{\pi}_{\widetilde{\pi},\lambda,h}(s,a) &= \psi(s,a)^{\intercal}\theta_h + \int_{\mathcal{S}} V^{\pi}_{\widetilde{\pi},\lambda,h}(s') \cdot \sum_{i=1}^{d} \psi(s,a)_i \mu_{h,i}(\mathrm{d}s') \\ &= \langle \psi(s,a), \theta_h + \int_{\mathcal{S}} V^{\pi}_{\widetilde{\pi},\lambda,h}(s') \mu_h(\mathrm{d}s') \rangle \,. \end{aligned}$$

To show the second part, we use Definition 3.2. First, we note that $\|\theta_h\| \leqslant \sqrt{d}$ and, at the same time

$$\int_{\mathcal{S}} V^{\pi}_{\widetilde{\pi},\lambda,h}(s') \mu_h(\mathrm{d}s') \leqslant H\sqrt{d}$$

since the value is bounded by H.

□

B.5.2 Algorithm Description

In this appendix, we describe the LSVI-UCB-Ent algorithm for regularized BPI in linear MDPs. LSVI-UCB-Ent is characterized by the following rules.

Sampling rule As for the UCBVI-Ent+ algorithm, we start with regularized optimistic planning under the linear function approximation

$$\begin{aligned}
\overline{Q}_h^t(s,a) &= \psi_h(s,a)^\intercal \overline{w}_h^t + b_h^t(s,a)\,, \\
\overline{V}_h^t(s) &= \mathrm{clip}\left(\max_{\pi\in\Delta(\mathcal{A})}\left\{\pi\overline{Q}_h^t(s) - \lambda\,\mathrm{KL}(\pi,\widetilde{\pi}_h(s))\right\}, 0, H\right), \\
\bar{\pi}_h^{t+1}(s) &= \arg\max_{\pi\in\Delta(\mathcal{A})}\left\{\pi\overline{Q}_h^t(s) - \lambda\,\mathrm{KL}(\pi,\widetilde{\pi}_h(s))\right\},
\end{aligned} \tag{B.12}$$

where b^t is some bonus defined as follows

$$b_h^t = \mathcal{B}\cdot\sqrt{[\psi(s,a)]^\intercal[\Lambda_h^t]^{-1}\psi(s,a)}$$

for $\mathcal{B}>0$ a bonus scaling factor, and the parameter w_h^t is obtained by least-square value iteration with Tikhonov regularization parameter α (Jin et al. 2020b),

$$\overline{w}_h^t = \arg\min_{w\in\mathbb{R}^d}\sum_{k=1}^{t}\left[r_h(s_h^k,a_h^k) + \overline{V}_{h+1}^t(s_{h+1}^k) - \psi(s_h^k,a_h^k)^\intercal w\right]^2 + \alpha\|w\|_2^2\,.$$

We notice that there is a closed-form solution to this problem given by

$$\overline{w}_h^t = \left[\Lambda_h^t\right]^{-1}\left[\sum_{\tau=1}^{t}\psi_h^\tau\left[r_h^\tau + \overline{V}_{h+1}^t(s_{h+1}^\tau)\right]\right],$$

where $\psi_h^\tau = \psi(s_h^\tau,a_h^\tau)$ and $\Lambda_h^t = \sum_{\tau=1}^t \psi_h^\tau[\psi_h^\tau]^\intercal + \alpha I$.

Then, we also define a family of exploratory policies by, for all $h'\in[H]$,

$$\pi^{t,(h')}(a|s) = \begin{cases} \pi^{t,(h')}(a|s) = \bar{\pi}_h^t(a|s) & \text{if } h\neq h' \\ \pi^{t,(h')}(a|s) = \mathbb{1}\left\{a\in\arg\max_{a'\in\mathcal{A}}(\overline{Q}_h^t(s,a') - \underline{Q}_h^t(s,a'))\right\} & \text{if } h = h' \end{cases}, \tag{B.13}$$

where $\underline{Q}^t$ is some lower bound on the optimal regularized Q-value defined as follows

$$\begin{aligned}
\underline{w}_h^t &= \left[\Lambda_h^t\right]^{-1}\left[\sum_{\tau=1}^{t}\psi_h^\tau\left[r_h^\tau + \underline{V}_{h+1}^t(s_{h+1}^\tau)\right]\right], \\
\underline{Q}_h^t(s,a) &= [\psi(s,a)]^\intercal\underline{w}_h^t - \mathcal{B}\cdot\sqrt{[\psi(s,a)]^\intercal[\Lambda_h^t]^{-1}\psi(s,a)}\,, \\
\underline{V}_h^t(s) &= \mathrm{clip}\left(\max_{\pi\in\Delta(\mathcal{A})}\left\{\pi\underline{Q}_h^t(s) - \lambda\,\mathrm{KL}(\pi\|\widetilde{\pi}_h(s))\right\}, 0, H\right).
\end{aligned} \tag{B.14}$$

The sampling rule is then obtained by picking uniformly at random a policy among the exploratory policies, $\pi^t = \pi^{t,(h')}$ for $h'\sim\mathcal{U}\mathrm{nif}[H]$. Notice that it is equivalent to using a non-Markovian mixture policy $\pi^{\mathrm{mix},t}$ over all $h\in[H]$. Additionally, it would be valuable to mention that computation of this policy could be done on-flight since we can compute $\overline{Q}$ and $\underline{Q}$.

Stopping and decision rule In the linear setting, we use a simple deterministic stopping rule $\tau = T$ for a fixed parameter T. In the finite setting, we could define an adaptive stopping rule by leveraging a certain Bernstein-like inequality on the gaps (see Lemma B.18). However, how to adapt such inequality to the linear setting remains unclear.

As decision rule `LSVI-UCB-Ent` returns the non-Markovian policy $\widehat{\pi}$, the uniform mixture over the optimistic policies $\{\bar{\boldsymbol{\pi}}^t\}_{t\in[T]}$ The complete procedure is described in Algorithm 9.

Algorithm 9 LSVI-UCB-Ent

Input: Number of episodes T, bonus function b^t, Tikhonov regularization parameter α.
for $t \in [T]$ **do**
  Compute $\bar{\boldsymbol{\pi}}^t$ by regularized optimistic planning with (B.12).
  Sample $h' \sim \mathcal{U}\mathrm{nif}\{1, \dots, H\}$ and set $\pi^t = \pi^{t,(h')}$
  for $h \in [H]$ **do**
    Play $a_h^t \sim \pi_h^t(s_h^t)$
    Observe $s_{h+1}^t \sim p_h(s_h^t, a_h^t)$
  end for
end for
Output $\widehat{\pi}$ the uniform mixture over $\{\pi^t\}_{t\in[T]}$.

B.5.3 Concentration Events

In this section, the required concentration events for a proof of sample complexity for LSVI-UCB-Ent will be described. First, we define several important objects. Let $\psi_h^\tau = \psi(s_h^\tau, a_h^\tau)$ for any $\tau \in \mathbb{N}, h \in [H]$ and define

$$\Lambda_h^t = \alpha I_d + \sum_{\tau=1}^{t} \psi_h^\tau [\psi_h^\tau]^\intercal \,, \quad \overline{\Lambda}_h^t = \alpha I_d + \sum_{\tau=1}^{t} \mathbb{E}_{\pi^{\mathrm{mix},\tau}}[\psi(s_h, a_h)[\psi(s_h, a_h)]^\intercal | s_1] \,,$$

where $\widetilde{\pi}^t$ is a uniform mixture policy of $\pi^{t,(h')}$ defined in (B.13) over all $h' \in \{0, \dots, H\}$.

Let $\beta^{\mathrm{conc}} \colon (0,1) \times \mathbb{N} \times \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ and $\beta^{\mathrm{cnt}} \colon (0,1) \times \mathbb{N} \to \mathbb{R}_+$ be some functions defined later on in Lemma B.23. We define the following favorable events for any fixed values of bonus scaling $\mathcal{B} > 0$ and Ridge coefficient $\alpha \geqslant 1$ that will be specified later.

$$\mathcal{E}^{\mathrm{conc}}(\delta, \mathcal{B}) \triangleq \Bigg\{ \forall t \in \mathbb{N}, \forall h \in [H] : \\ \left\| \sum_{\tau=1}^{t} \psi_h^\tau \left\{ \overline{V}_{h+1}^t(s_{h+1}^\tau) - p_h \overline{V}_{h+1}^t(s_h^\tau, a_h^\tau) \right\} \right\|_{[\Lambda_h^t]^{-1}} \leqslant 2dH\sqrt{\beta^{\mathrm{conc}}(\delta, t, \mathcal{B})} \\ \left\| \sum_{\tau=1}^{t} \psi_h^\tau \left\{ \underline{V}_{h+1}^t(s_{h+1}^\tau) - p_h \underline{V}_{h+1}^t(s_h^\tau, a_h^\tau) \right\} \right\|_{[\Lambda_h^t]^{-1}} \leqslant 2dH\sqrt{\beta^{\mathrm{conc}}(\delta, t, \mathcal{B})} \Bigg\} \,,$$

$$\mathcal{E}^{\mathrm{cnt}}(\delta) \triangleq \left\{ \forall t \in \mathbb{N}, \forall h \in [H] : \quad \Lambda_h^t \succcurlyeq \frac{1}{2}\overline{\Lambda}_h^t - \beta^{\mathrm{cnt}}(\delta, t) I_d \right\} .$$

We also introduce an intersection of these events of interest, $\mathcal{G}(\delta, \mathcal{B}) \triangleq \mathcal{E}^{\mathrm{conc}}(\delta, \mathcal{B}) \cap \mathcal{E}^{\mathrm{cnt}}(\delta)$. We prove that for the right choice of the functions $\beta^{\mathrm{conc}}, \beta^{\mathrm{cnt}}$ the above events hold with high probability.

Lemma B.23. *Let $\mathcal{B}, \alpha \geqslant 1$ be fixed. For any $\delta \in (0,1)$ and for the following choices of functions β,*

$$\beta^{\mathrm{conc}}(\delta, t, \mathcal{B}) \triangleq 2\log\left(\frac{H(1+t^2)}{\delta}\right) + 5 + \log\left(1 + 8d^{1/2}t^2 \cdot \left(\frac{\mathcal{B}}{Hd}\right)^2\right) ,$$
$$\beta^{\mathrm{cnt}}(\delta, t) \triangleq 4\log(8\mathrm{e}H(2t+1)/\delta) + 4d\log(3t) + 3 \,,$$

for any fixed $\alpha \geqslant 1$ it holds that

$$\mathbb{P}[\mathcal{E}^{\mathrm{conc}}(\delta, \mathcal{B})] \geqslant 1 - \delta/2 \,, \quad \mathbb{P}[\mathcal{E}^{\mathrm{cnt}}(\delta)] \geqslant 1 - \delta/2 \,,$$

In particular, $\mathbb{P}[\mathcal{G}(\delta, \mathcal{B})] \geqslant 1 - \delta$.

Proof. Let us fix $h \in [H]$. Then for all $t \in \mathbb{N}$ by Lemma B.24 we have $\|w_h^t\|_2 \leqslant 2H\sqrt{dt/\alpha}$ and by a construction of Λ_h^t we have $\lambda_{\min}(\Lambda_h^t) \geqslant \alpha$. Therefore, combination of Lemmas D.16 and D.17 for any fixed $\varepsilon > 0$ we have with probability at least $1 - \delta/H$

$$\left\|\sum_{\tau=1}^{t} \psi_h^\tau \left\{\overline{V}_{h+1}^t(s_{h+1}^\tau) - p_h\overline{V}_{h+1}^t(s_h^\tau, a_h^\tau)\right\}\right\|_{[\Lambda_h^t]^{-1}}^2 \leqslant 4H^2\left[\frac{d}{2}\log\left(\frac{H(t+\alpha)}{\alpha\delta}\right) + d\log\left(1 + \frac{8Hd^{1/2}t^{1/2}}{\varepsilon\alpha^{1/2}}\right) + d^2\log\left(1 + \frac{8d^{1/2}\mathcal{B}^2}{\alpha\varepsilon^2}\right)\right] + \frac{8t^2\varepsilon^2}{\alpha}\,.$$

Next we take $\varepsilon = Hd/t$ and obtain by using $\alpha \geqslant 1$ and $d \geqslant 1$

$$\left\|\sum_{\tau=1}^{t} \psi_h^\tau \left\{\overline{V}_{h+1}^t(s_{h+1}^\tau) - p_h\overline{V}_{h+1}^t(s_h^\tau, a_h^\tau)\right\}\right\|_{[\Lambda_h^t]^{-1}}^2 \leqslant 4H^2d^2\left[2\log\left(\frac{H(1+t^2)}{\delta}\right) + 5 + \log\left(1 + 8d^{1/2}t^2 \cdot \left(\frac{\mathcal{B}}{Hd}\right)^2\right)\right].$$

Taking the square root, we conclude the first half of the statement; the second half is exactly the same since Lemma D.16 gives a bound uniformly over all value functions.

By Theorem D.18 and union bound over $h \in [H]$, $\mathbb{P}[\mathcal{E}^{\mathrm{cnt}}(\delta)] \geqslant 1 - \delta/2$. The union bound over two prescribed events concludes $\mathbb{P}[\mathcal{G}(\delta, \mathcal{B})] \geqslant 1 - \delta$. □

The proof of the following lemma remains exactly the same as in Jin et al. (2020b).

Lemma B.24. *[Lemma B.2 by Jin et al. 2020b] For any $(t, h) \in \mathbb{N} \times [H]$ the weights w_h^t generated by* `LSVI-UCB-Ent` *satisfies*

$$\|w_h^t\|_2 \leqslant 2H\sqrt{dt/\alpha}\,.$$

B.5.4 Confidence Intervals

In this section, we provide the confidence intervals on the optimal Q-function that is required for the proof of sample complexity of `LSVI-UCB-Ent`.

We start by specifying the required values of α and $\mathcal{B}$.

$$\alpha \triangleq 2(\beta^{\mathrm{cnt}}(\delta, T) + 1), \qquad \mathcal{B} = 32dH\sqrt{\log\left(\frac{24edHT}{\delta}\right)}\,, \tag{B.15}$$

where β^{cnt} is defined in Lemma B.23.

Proposition B.25. *Let α and $\mathcal{B}$ satisfy (B.15). Then on the event $\mathcal{G}(\delta)$ defined in Lemma B.23 we have*

$$\langle\psi(s,a), \overline{w}_h^t\rangle - Q^\star_{\widetilde{\pi},\lambda,h}(s,a) = p_h[\overline{V}_{h+1}^t - V^\star_{\widetilde{\pi},\lambda,h+1}](s,a) + \overline{\Delta}_h^t,$$
$$\langle\psi(s,a), \underline{w}_h^t\rangle - Q^\star_{\widetilde{\pi},\lambda,h}(s,a) = p_h[\underline{V}_{h+1}^t - V^\star_{\widetilde{\pi},\lambda,h+1}](s,a) + \underline{\Delta}_h^t\,,$$

where $\overline{\Delta}_h^t(s,a)$ and $\underline{\Delta}_h^t(s,a)$ satisfies

$$\max\{|\overline{\Delta}_h^t(s,a)|, |\underline{\Delta}_h^t(s,a)|\} \leqslant \mathcal{B}\sqrt{\langle\psi(s,a), [\Lambda_h^t]^{-1}\psi(s,a)\rangle}\,.$$

Proof. We provide the proof only for the first equation since proof of one statement completely reassembles the other. By Proposition B.22 and Bellman equations we have

$$Q^\star_{\widetilde{\pi},\lambda,h}(s,a) = \langle \psi(s,a), w^\star_h \rangle = r_h(s,a) + p_h V^\star_{\widetilde{\pi},\lambda,h+1}(s,a)\,,$$

therefore

$$\begin{aligned}
\overline{w}^t_h - w^\star_h &= \left[\Lambda^t_h\right]^{-1}\left[\sum_{\tau=1}^t \psi^\tau_h[r^\tau_h + \overline{V}^t_{h+1}(s^\tau_{h+1})] - \sum_{\tau=1}^t \psi^\tau_h \langle \psi(s^\tau_h, a^\tau_h), w^\star_h\rangle - \alpha w^\star_h\right] \\
&= \underbrace{-\alpha\left[\Lambda^t_h\right]^{-1} w^\star_h}_{\xi_1} + \underbrace{\left[\Lambda^t_h\right]^{-1}\sum_{\tau=1}^t \psi^\tau_h\left[\overline{V}^t_{h+1}(s^\tau_{h+1}) - p_h\overline{V}^t_{h+1}(s^\tau_h, a^\tau_h)\right]}_{\xi_2} \\
&\quad + \underbrace{\left[\Lambda^t_h\right]^{-1}\sum_{\tau=1}^t \psi^\tau_h p_h\left[\overline{V}^t_{h+1} - V^\star_{\widetilde{\pi},\lambda,h+1}\right](s^\tau_h, a^\tau_h)}_{\xi_3}\,.
\end{aligned}$$

Next, we analyze the last term ξ_3. By Definition 3.2 we have

$$p_h\left[\overline{V}^t_{h+1} - V^\star_{\widetilde{\pi},\lambda,h+1}\right](s^\tau_h, a^\tau_h) = \left\langle \psi^\tau_h, \int_{\mathcal{S}} [\overline{V}^t_{h+1} - V^\star_{\widetilde{\pi},\lambda,h+1}](s')\mu_h(\mathrm{d}s')\right\rangle,$$

thus

$$\begin{aligned}
\xi_3 &= \left[\Lambda^t_h\right]^{-1}\left(\sum_{\tau=1}^t \psi^\tau_h[\psi^\tau_h]^\intercal + \alpha I_d - \alpha I_d\right)\left[\int_{\mathcal{S}} [\overline{V}^t_{h+1} - V^\star_{\widetilde{\pi},\lambda,h+1}](s')\mu_h(\mathrm{d}s')\right] \\
&= \int_{\mathcal{S}} [\overline{V}^t_{h+1} - V^\star_{\widetilde{\pi},\lambda,h+1}](s')\mu_h(\mathrm{d}s') \underbrace{-\alpha\left[\Lambda^t_h\right]^{-1}\int_{\mathcal{S}} [\overline{V}^t_{h+1} - V^\star_{\widetilde{\pi},\lambda,h+1}](s')\mu_h(\mathrm{d}s')}_{\xi_4}\,.
\end{aligned}$$

As a result, moving to Q-values directly, we have

$$\begin{aligned}
\langle \psi(s,a), \overline{w}^t_h\rangle - Q^\pi_h(s,a) &= \langle \psi(s,a), \overline{w}^t_h - w^\star_h\rangle \\
&= p_h[\overline{V}^t_{h+1} - V^\star_{\widetilde{\pi},\lambda,h+1}](s,a) + \underbrace{\langle \psi(s,a), \xi_1 + \xi_2 + \xi_4\rangle}_{\overline{\Delta}^t_h(s,a)}\,.
\end{aligned}$$

Next, we compute an upper bound for $\overline{\Delta}^t_h(s,a)$. We start from the first term, where we apply Cauchy-Schwartz inequality and the second statement of Proposition B.22

$$\begin{aligned}
|\langle \psi(s,a), \xi_1\rangle| = |\langle \psi(s,a), \alpha w^\star_h\rangle_{[\Lambda^t_h]^{-1}}| &\leqslant \alpha\|\psi(s,a)\|_{[\Lambda^t_h]^{-1}} \cdot \|w^\star_h\|_{[\Lambda^t_h]^{-1}} \\
&\leqslant 2H\sqrt{d\alpha}\|\psi(s,a)\|_{[\Lambda^t_h]^{-1}}\,,
\end{aligned}$$

where we have used that $\|[\Lambda^t_h]^{-1}\|_2 \leqslant 1/\alpha$. We apply the same construction for the third term and obtain the same upper bound. For the second term, we also apply Cauchy-Schwartz inequality and Lemma B.23

$$\begin{aligned}
|\langle \psi(s,a), \xi_1\rangle| &\leqslant \|\psi(s,a)\|_{[\Lambda^t_h]^{-1}}\left\|\sum_{\tau=1}^t \psi^\tau_h\left[\overline{V}^t_{h+1}(s^\tau_{h+1}) - p_h\overline{V}^t_{h+1}(s^\tau_h, a^\tau_h)\right]\right\|_{[\Lambda^t_h]^{-1}} \\
&\leqslant 2dH\sqrt{\beta^{\mathrm{conc}}(\delta, t, \mathcal{B})}\|\psi(s,a)\|_{[\Lambda^t_h]^{-1}}\,.
\end{aligned}$$

Thus, we have

$$|\overline{\Delta}_h^t(s,a)| \leqslant \left[2dH\sqrt{\beta^{\mathrm{conc}}(\delta,T)} + 4H\sqrt{2d(\beta^{\mathrm{cnt}}(\delta,T)+1)}\right]\sqrt{[\psi(s,a)]^\intercal[\Lambda_h^t]^{-1}\psi(s,a)}\,.$$

The only part is to show that for our particular choice of $\mathcal{B}$ it holds

$$2dH\sqrt{\beta^{\mathrm{conc}}(\delta,T,\mathcal{B})} + 4H\sqrt{2d(\beta^{\mathrm{cnt}}(\delta,T)+1)} \leqslant \mathcal{B}\,.$$

First, we notice that

$$4H\sqrt{2d(\beta^{\mathrm{cnt}}(\delta,T)+1)} \leqslant 16Hd\sqrt{\log\left(\frac{24\mathrm{e}HT}{\delta}\right)} \leqslant \mathcal{B}/2\,.$$

Thus, it is enough to show

$$\beta^{\mathrm{conc}}(\delta,T,\mathcal{B}) = 2\log\left(\frac{H(1+T^2)}{\delta}\right) + 5 + \log\left(1 + 8d^{1/2}T^2\left(\frac{\mathcal{B}}{dH}\right)^2\right) \leqslant \frac{1}{16}\left(\frac{\mathcal{B}}{dH}\right)^2.$$

First, we notice that since $T \geqslant 1$ and $\delta \in (0,1)$ then

$$2\log\left(\frac{H(1+T^2)}{\delta}\right) + 5 \leqslant 4\log\left(\frac{2TH\mathrm{e}^2}{\delta}\right),$$

and also, using the inequality $\log(1+x) \leqslant x$ for any $x \geqslant 0$

$$\begin{aligned}\log\left(1 + 8d^{1/2}T^2\left(\frac{\mathcal{B}}{dH}\right)^2\right) &\leqslant \log\left(1 + \frac{1}{32}\left(\frac{\mathcal{B}}{dH}\right)^2\right) + \log\left(32\cdot 8\cdot d^{1/2}T^2\right)\\ &\leqslant \frac{1}{32}\left(\frac{\mathcal{B}}{dH}\right)^2 + 2\log(16\cdot dT)\,.\end{aligned}$$

Thus, it is enough for $\mathcal{B}$ to satisfy the following inequalities

$$\log\left(\frac{24\mathrm{e}HT}{\delta}\right) \leqslant \left(\frac{\mathcal{B}}{32dH}\right)^2, \quad 4\log\left(\frac{2TH\mathrm{e}^2}{\delta}\right) \leqslant \left(\frac{\mathcal{B}}{8dH}\right)^2, \quad 2\log(16dT) \leqslant \left(\frac{\mathcal{B}}{8dH}\right)^2.$$

It is clear that the choice $\mathcal{B}$ defined in (B.15) satisfies all required inequalities. □

Corollary B.26 (Confidence intervals validity)**.** *Let constant α and $\mathcal{B}$ defined in (B.15). Then on the event $\mathcal{G}(\delta,\mathcal{B})$ we have $\overline{Q}_h^t(s,a) \geqslant Q^\star_{\widetilde{\pi},\lambda,h}(s,a) \geqslant \underline{Q}_h^t(s,a)$ and $\overline{V}_h^t(s) \geqslant V^\star_{\widetilde{\pi},\lambda,h}(s) \geqslant \underline{V}_h^t(s)$ for any $t \in [T], h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$.*

Proof. Let us prove using backward induction over $h \in [H]$. For $h = H+1$ this statement is trivially true. Let us assume that the statement holds for any $h' > h$. Thus by Proposition B.25

$$\overline{Q}_h^t(s,a) - Q^\star_{\widetilde{\pi},\lambda,h}(s,a) = \overline{\Delta}_h^t(s,a) + \mathcal{B}\sqrt{[\psi(s,a)]^\intercal[\Lambda_h^t]^{-1}\psi(s,a)} + p_h\left[\overline{V}_{h+1}^t - V^\star_{h+1}\right](s,a)\,.$$

Notice that $\overline{\Delta}_h^t(s,a) + \mathcal{B}\sqrt{[\psi(s,a)]^\intercal[\Lambda_h^t]^{-1}\psi(s,a)} \geqslant 0$ and by induction hypothesis $\overline{V}_{h+1}^t - V^\star_{h+1} \geqslant 0$ for any s'. Thus, we have proven the required statement for Q-values. To show it for V-values, we notice that if upper clipping in the definition $\overline{V}$ in (B.12) occurs, then the statement trivially holds. Otherwise, we have

$$\overline{V}_h^t(s) - V^\star_{\widetilde{\pi},\lambda,h}(s) \geqslant F_{\widetilde{\pi}_h(s),\lambda,h}(\overline{Q}_h^t(s)) - F_{\widetilde{\pi}_h(s),\lambda,h}(Q^\star_{\widetilde{\pi},\lambda,h}(s))\,.$$

However, the function $F_{\widetilde{\pi}_h(s),\lambda,h}$ is monotone since its gradients lie in a probability simplex. Thus, $\overline{V}_h^t(s) - V^\star_{\widetilde{\pi},\lambda,h}(s) \geqslant 0$. Using exactly the same reasoning, we may show the lower confidence bound. □

B.5.5 Sample Complexity Bounds

In this section, we provide the sample complexity result of the LSVI-UCB-Ent algorithm. We start with a general result that does not depend on the properties of linear MDPs but depends on the algorithms' properties.

Lemma B.27. *Let constants $\alpha, \mathcal{B}$ be defined by (B.15). Then on event $\mathcal{G}(\delta, \mathcal{B})$ defined in Lemma B.23 for any $t \in \mathbb{N}$, $s \in \mathcal{S}, h \in [H]$*

$$V^{\star}_{\widetilde{\pi},\lambda,h}(s) - V^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h}(s) \leqslant \frac{1}{2\lambda}\mathbb{E}_{\boldsymbol{\pi}^{t+1}}\left[\sum_{h=1}^{H} \max_{a\in\mathcal{A}}\left(\overline{Q}^t_h - \underline{Q}^t_h\right)^2(s_h, a)\right].$$

Proof. Let us proceed by induction. For $h = H + 1$ the statement is trivial. Assume that for any $h' > h$ the statement holds. Also, assume that $G^t_h(s) < H$; otherwise, the inequality on the policy error holds trivially. In particular, it holds that $n^t_h(s, a) > 0$ for all $a \in \mathcal{A}$.

We can start analysis from understanding the policy error by applying the smoothness of $F_{\widetilde{\pi}_h(s),\lambda,h}$.

$$\begin{aligned}
V^{\star}_{\widetilde{\pi},\lambda,h}(s) - V^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h}(s) &= F_{\widetilde{\pi}_h(s),\lambda,h}(Q^{\star}_{\widetilde{\pi},\lambda,h}(s,\cdot)) - \left(\bar{\pi}^{t+1}_h Q^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h}(s,\cdot) - \lambda\,\mathrm{KL}(\bar{\pi}^{t+1}_h(s)\|\widetilde{\pi}_h(s))\right)\\
&\leqslant F_{\widetilde{\pi}_h(s),\lambda,h}(\overline{Q}^t_h(s,\cdot)) + \langle \nabla F_{\widetilde{\pi}_h(s),\lambda,h}(\overline{Q}^t_h(s,\cdot)), Q^{\star}_{\widetilde{\pi},\lambda,h}(s,\cdot) - \overline{Q}^t_h(s,\cdot)\rangle\\
&\quad + \frac{1}{2\lambda}\|\overline{Q}^t_h - Q^{\star}_{\widetilde{\pi},\lambda,h}\|^2_{\infty}(s) - \left(\bar{\pi}^{t+1}_h Q^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h}(s,\cdot) - \lambda\,\mathrm{KL}(\bar{\pi}^{t+1}_h(s)\|\widetilde{\pi}_h(s))\right).
\end{aligned}$$

Next we recall that

$$\bar{\pi}^{t+1}_h(s) = \nabla F(\overline{Q}^t_h(s,\cdot)), \quad F(\overline{Q}^t_h)(s) = \bar{\pi}^{t+1}_h\overline{Q}^t_h(s) - \lambda\,\mathrm{KL}(\bar{\pi}^{t+1}_h(s)\|\widetilde{\pi}_h(s)),$$

thus we have

$$F(\overline{Q}^t_h)(s) - \left(\bar{\pi}^{t+1}_h Q^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h}(s,\cdot) - \lambda\,\mathrm{KL}(\bar{\pi}^{t+1}_h(s)\|\widetilde{\pi}_h(s))\right) = \bar{\pi}^{t+1}_h[\overline{Q}^t_h - Q^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h}](s)$$

and, by Bellman equations

$$\begin{aligned}
V^{\star}_{\widetilde{\pi},\lambda,h}(s) - V^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h}(s) &\leqslant \bar{\pi}^{t+1}_h\left[Q^{\star}_{\widetilde{\pi},\lambda,h} - Q^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h}\right](s) + \frac{1}{2\lambda}\|\overline{Q}^t_h - Q^{\star}_{\widetilde{\pi},\lambda,h}\|^2_{\infty}(s)\\
&\leqslant \bar{\pi}^{t+1}_h p_h\left[V^{\star}_{\widetilde{\pi},\lambda,h+1} - V^{\boldsymbol{\pi}^{t+1}}_{\widetilde{\pi},\lambda,h+1}\right](s) + \frac{1}{2\lambda}\|\overline{Q}^t_h - Q^{\star}_{\widetilde{\pi},\lambda,h}\|^2_{\infty}(s).
\end{aligned}$$

Next, we start by changing the norm and using the properties of $\overline{Q}$ and $\underline{Q}$ (see Corollary B.26)

$$\|\overline{Q}^t_h - Q^{\star}_{\widetilde{\pi},\lambda,h}\|^2_{\infty}(s) = \max_{a\in\mathcal{A}}\left(\overline{Q}^t_h(s,a) - Q^{\star}_{\widetilde{\pi},\lambda,h}(s,a)\right)^2 \leqslant \max_{a\in\mathcal{A}}\left(\overline{Q}^t_h(s,a) - \underline{Q}^t_h(s,a)\right)^2.$$

□

Theorem B.28. *Let $\varepsilon > 0, \delta \in (0, 1)$. Then LSVI-UCB-Ent algorithm with a choice of parameters described in (B.15) is (ε, δ)-PAC for the best policy identification in regularized MDPs after*

$$T = \widetilde{\mathcal{O}}\left(\frac{H^5 d^2}{\lambda\varepsilon}\right)$$

iterates.

Proof. First, we notice that the definition of the output policy $\widehat{\pi}$ as a mixture policy over $\{\bar{\boldsymbol{\pi}}^t\}_{t\in[T]}$ allows us to define the following

$$V^{\star}_{\widetilde{\pi},\lambda,1}(s_1) - V^{\widehat{\pi}}_{\widetilde{\pi},\lambda,1}(s_1) = \frac{1}{T}\sum_{i=1}^{T}\{V^{\star}_{\widetilde{\pi},\lambda,1}(s_1) - V^{\bar{\pi}^t}_{\widetilde{\pi},\lambda,1}(s_1)\}\,.$$

Therefore it is enough to compute only the average regret of the presented procedure. In the sequel we assume the event $\mathcal{G}(\delta,\mathcal{B})$ for $\mathcal{B}$ defined in (B.15).

Step 1. Study of sub-optimality gap. Let us fix $t\in[T]$, then by Lemma B.27 we have

$$V^{\star}_{\widetilde{\pi},\lambda,1}(s_1) - V_1^{\boldsymbol{\pi}^{t+1}}(s_1) \leqslant \frac{1}{2\lambda}\sum_{h=1}^{H}\mathbb{E}_{\boldsymbol{\pi}^{t+1}}\Big[\max_{a\in\mathcal{A}}\Big(\overline{Q}^t_h - \underline{Q}^t_h\Big)^2(s_h,a)\Big]\,.$$

Next, we analyze each term separately, starting from the difference between Q-values inside. Let us fix $h\in[H]$, then by Proposition B.25

$$\begin{aligned}\overline{Q}^t_h(s,a) - \underline{Q}^t_h(s,a) &= [\overline{Q}^t_h - Q^{\star}_{\widetilde{\pi},\lambda,h}](s,a) - [\underline{Q}^t_h - Q^{\star}_{\widetilde{\pi},\lambda,h}](s,a)\\ &\leqslant p_h\Big[\overline{V}^t_{h+1} - \underline{V}^t_{h+1}\Big](s,a) + 4\mathcal{B}\|\psi(s,a)\|_{[\Lambda^t_h]^{-1}}\,.\end{aligned} \tag{B.16}$$

Next, we have for any $h'\in[H]$ and any $s'\in\mathcal{S}$

$$\Big[\overline{V}^t_{h'} - \underline{V}^t_{h'}\Big](s') \leqslant \bar{\pi}^{t+1}_h\Big[\overline{Q}^t_{h'} - \underline{Q}^t_{h'}\Big](s')\,,$$

therefore we can roll-out the equation (B.16) and obtain

$$\overline{Q}^t_h(s,a) - \underline{Q}^t_h(s,a) \leqslant 4\mathcal{B}\mathbb{E}_{\boldsymbol{\pi}^{t+1}}\left[\sum_{h'=h}^{H}\|\psi(s_{h'},a_{h'})\|_{[\Lambda^t_{h'}]^{-1}}\middle|(s_h,a_h)=(s,a)\right].$$

Next, we apply Lemma D.12 for any fixed $h\in[H]$

$$\mathbb{E}_{\boldsymbol{\pi}^{t+1}}\Big[\max_{a\in\mathcal{A}}\Big(\overline{Q}^t_h - \underline{Q}^t_h\Big)^2(s_h,a)|s_1\Big] = \mathbb{E}_{\pi^{t+1,(h)}}\Big[\Big(\overline{Q}^t_h - \underline{Q}^t_h\Big)^2(s_h,a_h)|s_1\Big]$$

and for any $h'\geqslant h$

$$\mathbb{E}_{\boldsymbol{\pi}^{t+1}}\left[\|\psi(s_{h'},a_{h'})\|_{[\Lambda^t_{h'}]^{-1}}\middle|(s_h,a_h)=(s,a)\right] = \mathbb{E}_{\pi^{t+1,(h)}}\left[\|\psi(s_{h'},a_{h'})\|_{[\Lambda^t_{h'}]^{-1}}\middle|(s_h,a_h)=(s,a)\right].$$

Therefore, applying Jensen's inequality to conditional measure and the tower property of conditional expectation

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\pi}^{t+1}}\Big[\max_{a\in\mathcal{A}}\Big(\overline{Q}^t_h - \underline{Q}^t_h\Big)^2(s_h,a)|s_1\Big] &\leqslant 16\mathcal{B}^2\mathbb{E}_{\pi^{t+1,(h)}}\left[\left(\sum_{h'=h}^{H}\|\psi(s_{h'},a_{h'})\|_{[\Lambda^t_{h'}]^{-1}}\right)^2|s_1\right]\\ &\leqslant 16\mathcal{B}^2H\sum_{h'=h}^{H}\mathbb{E}_{\pi^{t+1,(h)}}\Big[\|\psi(s_{h'},a_{h'})\|^2_{[\Lambda^t_{h'}]^{-1}}|s_1\Big]\\ &\leqslant 16\mathcal{B}^2H\sum_{h'=1}^{H}\mathbb{E}_{\pi^{t+1,(h)}}\Big[\|\psi(s_{h'},a_{h'})\|^2_{[\Lambda^t_{h'}]^{-1}}|s_1\Big]\,.\end{aligned}$$

Summing over all h and recalling $\widetilde{\pi}^{t+1}$ as a mixture policy of all $\pi^{t+1,(h)}$

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\pi}^{t+1}}\Big[\max_{a\in\mathcal{A}}\Big(\overline{Q}^t_h - \underline{Q}^t_h\Big)^2(s_h,a)|s_1\Big] &\leqslant 16\mathcal{B}^2H\sum_{h'=1}^{H}\sum_{h=1}^{H}\mathbb{E}_{\pi^{t+1,(h)}}\Big[\|\psi(s_{h'},a_{h'})\|^2_{[\Lambda^t_{h'}]^{-1}}|s_1\Big]\\ &= 16\mathcal{B}^2H^2\sum_{h'=1}^{H}\mathbb{E}_{\pi^{\mathrm{mix},t+1}}\Big[\|\psi(s_{h'},a_{h'})\|^2_{[\Lambda^t_{h'}]^{-1}}|s_1\Big]\,.\end{aligned}$$

Step 2. Summing sub-optimality gaps. Next, we sum all sub-optimality gaps and obtain

$$\sum_{t=1}^{T}\{V^{\star}_{\widetilde{\pi},\lambda,1}(s_1) - V^{\bar{\boldsymbol{\pi}}^t}_{\widetilde{\pi},\lambda,1}(s_1)\} \leqslant H + \sum_{t=1}^{T-1}\{V^{\star}_{\widetilde{\pi},\lambda,1}(s_1) - V_1^{\boldsymbol{\pi}^{t+1}}(s_1)\}$$
$$\leqslant H + \frac{16\mathcal{B}^2H^2}{\lambda}\sum_{h'=1}^{H}\sum_{t=1}^{T-1}\mathbb{E}_{\pi^{\mathrm{mix},t+1}}\Big[\|\psi(s_{h'},a_{h'})\|^2_{[\Lambda^t_{h'}]^{-1}}|s_1\Big].$$

Next, we apply the definition of event $\mathcal{E}^{\mathrm{cnt}}(\delta)$

$$\Lambda^t_h \succcurlyeq \frac{1}{2}\overline{\Lambda}^t_h - \beta^{\mathrm{cnt}}(\delta,T)I_d \Rightarrow [\Lambda^t_h]^{-1} \preccurlyeq 2\Big[\overline{\Lambda}^t_h - 2\beta^{\mathrm{cnt}}(\delta,T)I_d\Big]^{-1}.$$

Notice that by the choice of $\alpha = 2(\beta^{\mathrm{cnt}}(\delta,t)+1)$ we have

$$\widetilde{\Lambda}^t_h \triangleq \overline{\Lambda}^t_h - 2\beta^{\mathrm{cnt}}(\delta,T)I_d = 2I_d + \sum_{\tau=1}^{t}\mathbb{E}_{\pi^{\mathrm{mix},\tau}}[\psi(s_h,a_h)[\psi(s_h,a_h)]^{\intercal}].$$

Thus, we may apply Lemma D.8

$$\sum_{t=1}^{T-1}\mathbb{E}_{\widetilde{\pi}^{t+1}}\Big[\|\psi(s_{h'},a_{h'})\|^2_{[\Lambda^t_{h'}]^{-1}}|s_1\Big] \leqslant 2\sum_{t=1}^{T-1}\mathbb{E}_{\pi^{\mathrm{mix},t+1}}\Big[\|\psi(s_{h'},a_{h'})\|^2_{[\widetilde{\Lambda}^t_{h'}]^{-1}}|s_1\Big]$$
$$\leqslant 4\log\det\Big(\widetilde{\Lambda}^T_{h'}\Big).$$

To upper bound the determinant, we upper bound the operator norm using the triangle inequality

$$\|\widetilde{\Lambda}^T_h\|_2 = \Big\|2I_d + \sum_{\tau=1}^{T-1}\mathbb{E}_{\widetilde{\pi}^{\tau}}[\psi(s_h,a_h)[\psi(s_h,a_h)]^{\intercal}]\Big\|_2 \leqslant 2 + (T-1),$$

therefore, combining with a definition of $\mathcal{B}$ given in (B.15) we have

$$\sum_{t=1}^{T}\{V^{\star}_{\widetilde{\pi},\lambda,1}(s_1) - V^{\bar{\boldsymbol{\pi}}^t}_{\widetilde{\pi},\lambda,1}(s_1)\} \leqslant H + \frac{64\cdot 32^2d^2H^5\cdot\log(T+1)}{\lambda}\cdot\log\Big(\frac{24\mathrm{e}dHT}{\delta}\Big),$$

yielding

$$V^{\star}_{\widetilde{\pi},\lambda,1}(s_1) - V^{\widehat{\pi}}_{\widetilde{\pi},\lambda,1}(s_1) \leqslant \frac{H}{T} + \frac{64\cdot 32^2d^2H^5\cdot\log(T+1)}{\lambda T}\cdot\log\Big(\frac{24\mathrm{e}dHT}{\delta}\Big).$$

In particular, it implies that $T = \widetilde{\mathcal{O}}\Big(\frac{H^5d^2}{\lambda\varepsilon}\Big)$ is enough to achieve ε-accurate policy. □

B.6 Demonstration-Regularized Preference-Based Learning

B.6.1 Maximum Likelihood Estimation for Reward Model

In this section, we discuss the maximum likelihood estimation problem for the reward estimation, following Zhan et al. (2023). Let $\mathcal{G}$ be a function class of reward functions that satisfies the following assumption

Assumption 6. For a function class $\mathcal{G}$, we assume that the true reward belongs to it: $r^\star \in \mathcal{G}$. Additionally, for the following function family $\mathcal{Q} = \{q_r(\tau_0, \tau_1) = \sigma(r(\tau_1) - r(\tau_0)) : r \in \mathcal{G}\}$ equipped with an ℓ_∞-norm has a finite *bracketing dimension*, that means there is a $d_\mathcal{G} > 0$ and $R_\mathcal{G} > 0$ such that

$$\forall \varepsilon \in (0,1) : \log \mathcal{N}_{[]}(\varepsilon, \mathcal{Q}, \|\cdot\|_\infty) \leqslant d_\mathcal{G} \log(R_\mathcal{G}/\varepsilon)\,.$$

The bracketing numbers are commonly used in statistics for MLE, M-estimation, and, more generally, in the empirical processes theory, see Geer (2000). Related to our setting, this assumption is satisfied in the setting of tabular MDPs with a dimension $d_\mathcal{G} = SAH$ and linear MDPs with dimensions $d_\mathcal{G} = dH$, see Lemmas B.33-B.34.

For each $r \in \mathcal{G}$ and a pair of trajectories (τ_0, τ_1) define the induced preference model as follows

$$q_r(\tau_0, \tau_1) \triangleq \sigma(r(\tau_1) - r(\tau_0))\,.$$

To measure the complexity of the reward class $\mathcal{G}$, we will use the bracketing numbers of the function class $\mathcal{Q} = \{q_r \colon \mathcal{T} \times \mathcal{T} \to [0,1] : r \in \mathcal{G}\}$, where $\mathcal{T} = (\mathcal{S} \times \mathcal{A})^H$ is a space of all trajectories. See Definition B.2 for the definition of bracketing numbers.

Given the dataset of preferences $\mathcal{D}^{\mathrm{RM}} = \{(\tau_0^k, \tau_1^k, o^k)\}_{k=1}^{N^{\mathrm{RM}}}$, we define the maximum likelihood estimate of the reward model as follows

$$\widehat{r} = \arg\max_{r \in \mathcal{G}} \left\{ \sum_{k=1}^{N^{\mathrm{RM}}} o^k \log q_r(\tau_0^k, \tau_1^k) + (1 - o^k) \log(1 - q_r(\tau_0^k, \tau_1^k)) \right\}. \tag{B.17}$$

The following result is a standard result on MLE estimation and, generally, M-estimation, see Geer (2000). The proof heavily uses PAC-Bayes techniques by Zhang (2006), see also Agarwal et al. (2020a) for non-i.i.d. extension. We notice that a similar result could be extracted from Lemma 2 by Zhan et al. (2023); however, we did not find the proof of exactly this statement in their paper or references within.

Proposition B.29. *Let Assumptions 4-6 hold. Let $\mathcal{D}^{\mathrm{RM}}$ be a preference dataset and assume that trajectories τ_0^k and τ_1^k were generated i.i.d. by following the policy π. Then for any $\delta \in (0,1)$ with probability at least $1-\delta$ the following bound for the MLE reward estimate $\widehat{r}$ given by solution to B.17 holds*

$$\mathbb{E}_{\tau_0, \tau_1 \sim q^\pi}\left[\Big(q_\star(\tau_0, \tau_1) - q_{\widehat{r}}(\tau_0, \tau_1)\Big)^2\right] \leqslant \frac{2 + 2 d_\mathcal{G} \log(R_\mathcal{G} N^{\mathrm{RM}}) + \log(1/\delta)}{N^{\mathrm{RM}}}\,,$$

where q^π is a distribution over trajectories induced by policy π

Proof. In the sequel, we drop the superscript from $\mathcal{D}^{\mathrm{RM}}$ and N^{RM} to simplify the notation.

Let us consider a maximal set of ε-brackets B of size $\mathcal{N}_{[]}(\varepsilon, \mathcal{Q}, \|\cdot\|_\infty)$ and apply Lemma 2.1 by Zhang (2006) (or Lemma 21 by Agarwal et al. (2020a)), where consider brackets $[\ell, u]$ from B as parameters θ, prior π is a uniform over B, and the density $w_\mathcal{D}([\ell, u])$ is equal to a (properly weighted) Dirac measure on a bracket $[\ell^\star, u^\star]$ that contains the MLE estimate (ties are resolved arbitrary). It implies that for any function $\mathcal{L} \colon B \times \mathcal{D} \to \mathbb{R}$ it holds

$$\mathbb{E}_\mathcal{D}\Big[\exp\Big\{\mathcal{L}(\mathcal{D}, [\ell^\star, u^\star]) - \log \mathbb{E}_{\mathcal{D}'}\Big[\mathrm{e}^{\mathcal{L}(\mathcal{D}', [\ell^\star, u^\star])}\Big] - \log \mathcal{N}_{[]}(\varepsilon, \mathcal{Q}, \|\cdot\|_\infty)\Big\}\Big] \leqslant 1\,,$$

where $\mathcal{D}'$ is an independent copy of the dataset $\mathcal{D}$ and the KL-divergence between a Dirac measure on an MLE bracket and the uniform distribution is computed exactly. Since we can control the exponential moment, a simple Chernoff argument implies that with probability at last $1-\delta$

$$-\log \mathbb{E}_{\mathcal{D}'}\left[\mathrm{e}^{\mathcal{L}(\mathcal{D}',[l^\star,u^\star])}\right] \leqslant -\mathcal{L}(\mathcal{D},[l^\star,u^\star]) + \log \mathcal{N}_{[]}(\varepsilon,\mathcal{Q},\|\cdot\|_\infty) + \log(1/\delta)\,.$$

Next we choose $\mathcal{L}(\mathcal{D},[\ell,u])$ as log-likelihood ratio

$$\mathcal{L}(\mathcal{D},[\ell,u]) = -\frac{1}{2}\sum_{k=1}^N \left\{ \frac{o^k \log q_\star(\tau_0^k,\tau_1^k) + (1-o^k)\log(1-q_\star(\tau_0^k,\tau_1^k))}{o^k \log u(\tau_0^k,\tau_1^k) + (1-o^k)\log(1-\ell(\tau_0^k,\tau_1^k))} \right\}.$$

By the choice of a brackets $[\ell,u]$ it holds $\ell^\star(\tau_0,\tau_1) \leqslant q_{\widehat{r}}(\tau_0,\tau_1) \leqslant u^\star(\tau_0,\tau_1)$. Using the fact that $\widehat{r}$ is a solution to (B.17)

$$\begin{aligned}
-\mathcal{L}(\mathcal{D},[\ell^\star,u^\star]) &= \frac{1}{2}\sum_{k=1}^N \left\{ \frac{o^k \log q_\star(\tau_0^k,\tau_1^k) + (1-o^k)\log(1-q_\star(\tau_0^k,\tau_1^k))}{o^k \log u(\tau_0^k,\tau_1^k) + (1-o^k)\log(1-\ell(\tau_0^k,\tau_1^k))} \right\} \\
&\leqslant \frac{1}{2}\sum_{k=1}^N \left\{ \frac{o^k \log q_\star(\tau_0^k,\tau_1^k) + (1-o^k)\log(1-q_\star(\tau_0^k,\tau_1^k))}{o^k \log q_{\widehat{r}}(\tau_0^k,\tau_1^k) + (1-o^k)\log(1-q_{\widehat{r}}(\tau_0^k,\tau_1^k))} \right\} \leqslant 0\,.
\end{aligned}$$

At the same time

$$-\log \mathbb{E}_{\mathcal{D}'}\left[\mathrm{e}^{\mathcal{L}(\mathcal{D}',[l^\star,u^\star])}\right] = -N \log \mathbb{E}\left[\exp\left\{ -\frac{1}{2}\frac{o\log q_\star(\tau_0,\tau_1) + (1-o)\log(1-q_\star(\tau_0,\tau_1))}{o\log u^\star(\tau_0,\tau_1) + (1-o)\log(1-\ell^\star(\tau_0,\tau_1))} \right\}\right],$$

where in the last expectation $\tau_0,\tau_1 \sim q^\pi, o \sim \mathcal{B}\mathrm{er}(q_\star(\tau_0,\tau_1))$.

By Fubini's theorem, we have

$$\begin{aligned}
-\frac{1}{N}\log \mathbb{E}_{\mathcal{D}'}\left[\mathrm{e}^{\mathcal{L}(\mathcal{D}',[l^\star,u^\star])}\right] = -\log\, \mathbb{E}_{\tau_0,\tau_1}\Bigg[&\sqrt{q_\star(\tau_0,\tau_1)u(\tau_0,\tau_1)} \\
&+ \sqrt{(1-q_\star(\tau_0,\tau_1))(1-\ell(\tau_0,\tau_1))}\Bigg]\,.
\end{aligned}$$

Next, we study the expression under the square root. By the definition of the ε-bracket we have $\ell^\star(\tau_0,\tau_1) \leqslant q_{\widehat{r}}(\tau_0,\tau_1) \leqslant u^\star(\tau_0,\tau_1)$ and $u^\star(\tau_0,\tau_1) - \ell^\star(\tau_0,\tau_1) \leqslant \varepsilon$, therefore

$$\sqrt{q_\star(\tau_0,\tau_1)u(\tau_0,\tau_1)} \leqslant \sqrt{q_\star(\tau_0,\tau_1)(q_{\widehat{r}}(\tau_0,\tau_1)+\varepsilon)} \leqslant \sqrt{q_\star(\tau_0,\tau_1)q_{\widehat{r}}(\tau_0,\tau_1)} + \sqrt{\varepsilon}\,.$$

and the similar bound for the second term. Applying inequality $-\log(x) \geqslant 1-x$ we have

$$\begin{aligned}
-\frac{1}{N}\log \mathbb{E}_{\mathcal{D}'}\left[\mathrm{e}^{\mathcal{L}(\mathcal{D}',[l^\star,u^\star])}\right] \geqslant 1 - \mathbb{E}_{\tau_0,\tau_1}\Bigg[&\sqrt{q_\star(\tau_0,\tau_1)q_{\widehat{r}}(\tau_0,\tau_1)} \\
&+ \sqrt{(1-q_\star(\tau_0,\tau_1))(1-q_{\widehat{r}}(\tau_0,\tau_1))}\Bigg] - 2\sqrt{\varepsilon}\,.
\end{aligned}$$

By the properties of the Hellinger distance $d_{\mathcal{H}}$ (see Section 2.4 and Lemma 2.3 by Tsybakov 2008) we have

$$\begin{aligned}
&1 - \mathbb{E}_{\tau_0,\tau_1}\left[\sqrt{q_\star(\tau_0,\tau_1)q_{\widehat{r}}(\tau_0,\tau_1)} + \sqrt{(1-q_\star(\tau_0,\tau_1))(1-q_{\widehat{r}}(\tau_0,\tau_1))}\right] \\
&= \frac{1}{2}\mathbb{E}_{\tau_0,\tau_1}\left[d^2_{\mathcal{H}}(\mathcal{B}\mathrm{er}(q_\star(\tau_0,\tau_1)),\mathcal{B}\mathrm{er}(q_{\widehat{r}}(\tau_0,\tau_1)))\right] \geqslant \mathbb{E}_{\tau_0,\tau_1}\left[(q_\star(\tau_0,\tau_1) - q_{\widehat{r}}(\tau_0,\tau_1))^2\right].
\end{aligned}$$

Overall, we obtain

$$\mathbb{E}_{\tau_0,\tau_1}\left[\left(q_\star(\tau_0,\tau_1) - q_{\widehat{r}}(\tau_0,\tau_1)\right)^2\right] \leqslant 2\sqrt{\varepsilon} + \frac{\log \mathcal{N}_{[]}(\varepsilon, \mathcal{Q}, \|\cdot\|_\infty) + \log(1/\delta)}{N}\,.$$

Taking $\varepsilon = 1/N^2$ and applying the upper bound on the bracketing number by Assumption 6 we conclude the statement. □

We also have a simple corollary of this result that shows convergence of the reward models.

Theorem B.30. *Let Assumptions 4-6 hold. Let $\mathcal{D}^{\mathrm{RM}}$ be a preference dataset and assume that trajectories τ_0^k and τ_1^k were generated i.i.d. by following the policy π. Then for any $\delta \in (0,1)$ with probability at least $1-\delta$ the following bound holds*

$$\mathbb{E}_{\tau_0,\tau_1\sim q^\pi}\left[\left(r^\star(\tau_1) - r^\star(\tau_0) - \widehat{r}(\tau_1) + \widehat{r}(\tau_1)\right)^2\right] \leqslant \frac{2\zeta^2 d_{\mathcal{G}} \log(R_{\mathcal{G}}/N^{\mathrm{RM}}) + \zeta^2 \log(\mathrm{e}^2/\delta)}{N^{\mathrm{RM}}}\,,$$

where $\zeta = 1/(\inf_{x\in[-H,H]} \sigma'(x))$ the non-linearity measure of the link function σ defined in Assumption 4.

Remark B.31. We additionally notice that since two trajectories are i.i.d., we have

$$\mathbb{E}_{\tau_0,\tau_1\sim q^\pi}\left[\left([r^\star(\tau_1) - r^\star(\tau_0)] - [\widehat{r}(\tau_1) - \widehat{r}(\tau_0)]\right)^2\right] = 2\mathrm{Var}_{q^\pi}[r^\star(\tau_1) - \widehat{r}(\tau_1)]\,.$$

This means, in particular, that if the reward function is estimated up to a constant shift, then the MLE estimation error is zero.

Remark B.32. In the setting of a sigmoid link function $\sigma(x) = 1/(1+\exp(-x))$ we have $\zeta = \exp\{\Theta(H)\}$, yields the exponential dependence on the reward scaling. However, for the general distribution of trajectories, the exponential dependence is unavoidable even in the linear setting (Hazan et al. 2014).

Proof. Follows from the application of Proposition B.29 and the following application of mean-value theorem for any two fixed τ_0, τ_1

$$\begin{aligned}[r^\star(\tau_1) - r^\star(\tau_0)] - [\widehat{r}(\tau_1) - \widehat{r}(\tau_0)] &= \sigma^{-1}(q_\star(\tau_0,\tau_1)) - \sigma^{-1}(q_{\widehat{r}}(\tau_0,\tau_1)) \\ &= (\sigma^{-1})'(\xi)[q_\star(\tau_0,\tau_1) - q_{\widehat{r}}(\tau_0,\tau_1)]\,,\end{aligned}$$

where ξ is a point between $q_\star(\tau_0,\tau_1)$ and $q_{\widehat{r}}(\tau_0,\tau_1)$. The observation $(\sigma^{-1})'(\xi) = 1/\sigma'(\sigma^{-1}(\xi))$ yields the statement. □

Next, we compute the required quantities $d_{\mathcal{G}}$ for the case of finite and linear MDPs for a choice of $\sigma = 1/(1+\exp(-x)$ as a sigmoid function.

Lemma B.33. *Let a reward function $\{r_h(s,a)\}_{h\in[H]}$ be an arbitrary function $r_h \colon \mathcal{S}\times\mathcal{A} \to [0,1]$. Let us define $\mathcal{G} = \{r(\tau) = \sum_{h=1}^H r_h(s_h,a_h)\}$. Then Assumption 6 holds with constants $d_{\mathcal{G}} = HSA$ and $R_{\mathcal{G}} = 3H/2$.*

Proof. Let us define a functional class of interest $\mathcal{Q} = \{q_r(\tau_1,\tau_2) = \sigma(r(\tau_1) - r(\tau_2)) \mid r \in \mathcal{G}\}$. Since σ is a monotonically increasing function that satisfies $\sigma'(x) \leqslant 1/4$. Thus, by combination of Lemma B.35 and Lemma B.36 we have

$$\mathcal{N}_{[]}(\varepsilon, \mathcal{Q}, \|\cdot\|_\infty) \leqslant \mathcal{N}_{[]}(2\varepsilon, \mathcal{G}, \|\cdot\|_\infty)\,.$$

Next we define a function classes $\mathcal{F}_h = \{r_h \colon \mathcal{S} \times \mathcal{A} \to [0,1]\}$ of one-step rewards. By Lemma B.37 it holds

$$\mathcal{N}_{[]}(2\varepsilon, \mathcal{G}, \|\cdot\|_\infty) \leqslant \prod_{h=1}^{H} \mathcal{N}_{[]}(2\varepsilon/H, \mathcal{F}_h, \|\cdot\|_\infty)\,.$$

Then we can associate a function space $\mathcal{F}_h$ with a parameters $\Theta_h = [0,1]^{SA}$ and by Lemma B.38 and a standard results in bounding of covering numbers of balls in normed spaces, see Handel (2016),

$$\mathcal{N}_{[]}(\varepsilon, \mathcal{F}_h, \|\cdot\|_\infty) \leqslant \mathcal{N}(\varepsilon/2, [0,1]^{SA}, \|\cdot\|_\infty) \leqslant (3/\varepsilon)^{SA}\,.$$

As a result, we have

$$\log \mathcal{N}_{[]}(\varepsilon, \mathcal{Q}, \|\cdot\|_\infty) \leqslant HSA \log(3H/(2\varepsilon))\,.$$

□

Lemma B.34. *Let a reward function $\{r_h(s,a)\}_{h\in[H]}$ be parametrized as $r_h(s,a) = \psi(s,a)^\intercal \theta_h$ for $\psi \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ that satisfies $\|\psi(s,a)\|_2 \leqslant 1$, and $\theta_h \in \Theta_h$ for $\Theta_h = \{\theta_h \in \mathbb{R}^d \mid \|\theta_h\| \leqslant \sqrt{d}\}$. Let us define $\mathcal{G} = \{r(\tau) = \sum_{h=1}^{H} r_h(s_h, a_h)\}$. Then Assumption 6 holds with constants $d_{\mathcal{G}} = dH$ and $R_{\mathcal{G}} = 3H\sqrt{d}/2$.*

Proof. Let us define one-step rewards as follows $\mathcal{F}_h = \{r_h(s,a) = \psi(s,a)^\intercal \theta_h \mid \theta_h \in \mathbb{R}^d, \|\theta_h\| \leqslant \sqrt{d}\}$. Then, following exactly the same reasoning as in Lemma B.33

$$\mathcal{N}_{[]}(\varepsilon, \mathcal{Q}, \|\cdot\|_\infty) \leqslant \prod_{h=1}^{H} \mathcal{N}_{[]}(2\varepsilon/H, \mathcal{F}_h, \|\cdot\|_\infty)\,.$$

Next we notice that $\mathcal{F}_h$ is Lipschtiz in ℓ_2-norm with respect to parameters θ_h with a constant 1:

$$|r_h(s,a) - r'_h(s,a)| = |\psi(s,a)^\intercal(\theta_h - \theta'_h)| \leqslant \|\theta_h - \theta'_h\|_2\,.$$

Therefore, since Θ_h is a ball of radius $\sqrt{d}$ we obtain

$$\mathcal{N}_{[]}(\varepsilon, \mathcal{F}_h, \|\cdot\|_\infty) \leqslant \mathcal{N}(\varepsilon/2, \Theta_h, \|\cdot\|_2) \leqslant (3\sqrt{d}/\varepsilon)^d\,.$$

As a result, we have

$$\log \mathcal{N}_{[]}(\varepsilon, \mathcal{Q}, \|\cdot\|_\infty) \leqslant dH \log(3H\sqrt{d}/(2\varepsilon))\,.$$

□

B.6.2 Properties of Bracketing numbers

In this section, we provide a list of elementary properties of bracketing numbers for completeness. See Section 7 of Dudley (2014) for additional information.

Lemma B.35. *Let $(\mathcal{G}, \|\cdot\|_\infty)$ be a normed space of functions over a set $\mathcal{X}$ that takes values in the interval I and let $\sigma \colon I \to [0,1]$ be a monotonically increasing link function that satisfies $\sup_{x\in I} \sigma'(x) \leqslant C$ for $C > 0$. Then for $\mathcal{F} = \sigma \circ \mathcal{G} = \{f = \sigma \circ g \mid g \in \mathcal{G}\}$ it holds for any $\varepsilon > 0$*

$$\mathcal{N}_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \leqslant \mathcal{N}_{[]}(\varepsilon/C, \mathcal{G}, \|\cdot\|_\infty)\,.$$

Proof. Let G be a minimal set of ε brackets that covers $\mathcal{G}$. Then let us take a set of brackets $F = \{[\sigma \circ \ell, \sigma \circ u] : [l,u] \in G\}$. The set of F consists of brackets since σ is monotone and covers all the space $\mathcal{F}$. Additionally, we have

$$|\sigma \circ u(x) - \sigma \circ \ell(x)| = \sigma'(\xi)|(u(x) - \ell(x))| \leqslant C\varepsilon\,,$$

therefore the set F consists of $C\varepsilon$-brackets. By rescaling, we conclude the statement. □

Lemma B.36. *Let $(\mathcal{G}, \|\cdot\|_\infty)$ be a normed space of functions over a set $\mathcal{X}$ and let us define a set of functions over $\mathcal{X} \times \mathcal{X}$ as $\mathcal{F} = \{f(x,x') = g(x) - g(x') \mid g \in \mathcal{G}\}$. Then it holds*

$$\mathcal{N}_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \leqslant \mathcal{N}_{[]}(\varepsilon/2, \mathcal{G}, \|\cdot\|_\infty)\,.$$

Proof. Let G be a minimal set of ε brackets that covers $\mathcal{G}$. Let us define the following set of brackets

$$F = \{[f_\ell(x,x') = \ell(x) - u(x'), f_u(x,x') = u(x) - \ell(x')] \mid [\ell, u] \in G\}\,.$$

We may check that elements cover all the set $\mathcal{F}$. Let us take $f \in \mathcal{F}$ and the corresponding $g \in \mathcal{G}$. Then let us take a bracket $[\ell, u]$ that covers g. In this case, we have

$$f_\ell(x,x') = \ell(x) - u(x') \leqslant g(x) - g(x') \leqslant u(x) - \ell(x') = f_u(x,x')\,.$$

At the same time, we have

$$\|f_\ell - f_u\|_\infty \leqslant 2\|u - \ell\|_\infty \leqslant 2\varepsilon\,.$$

Thus, F is a set of 2ε-brackets that covers $\mathcal{F}$. □

Lemma B.37. *Let $\{\mathcal{G}_k\}_{k=1}^K$ be a sequence of spaces of functions over a set $\mathcal{X}$ equipped with a norm $\|\cdot\|_\infty$ and let us define a set $\mathcal{F}$ of functions over $\mathcal{X}^K$ as follows*

$$\mathcal{F} = \left\{ f(x_1, \ldots, x_k) = \sum_{k=1}^K g_k(x_k) \mid g_k \in \mathcal{G} \right\}.$$

Then the following bound holds

$$\mathcal{N}_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \leqslant \prod_{k=1}^K \mathcal{N}_{[]}(\varepsilon/K, \mathcal{G}, \|\cdot\|_\infty)\,.$$

Proof. For any $k \in [K]$ let G_k be a minimal set of ε brackets that covers $\mathcal{G}_k$. Then we construct the set F as follows

$$F = \left\{ \left[f_\ell(x) = \sum_{k=1}^K \ell_k(x_k), f_u(x) = \sum_{k=1}^k u_k(x_k) \right] \mid \forall k \in [K] : [\ell_k, u_k] \in \mathcal{G}_k \right\}.$$

It holds that $|F| \leqslant \prod_{k=1}^K |G_k|$ and also it is clear that F consists of brackets and covers all the set $\mathcal{F}$. Additionally, we notice that

$$\|f_\ell - f_u\|_\infty \leqslant \sum_{k=1}^K \|\ell_k - u_k\|_\infty \leqslant K\varepsilon\,.$$

By rescaling, we conclude the statement. □

Lemma B.38. *Let Θ be a set of parameters and let $\mathcal{F} = \{f_\theta(x) \mid \theta \in \Theta\}$ be a set of functions over $\mathcal{X}$. Assume that $f_\theta(x)$ is L-Lipschitz in θ with respect to an arbitrary norm $\|\cdot\|$:*

$$\forall x \in \mathcal{X} : |f_\theta(x) - f_{\theta'}(x)| \leqslant L\|\theta - \theta'\|\,.$$

Then we have

$$\mathcal{N}_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \leqslant \mathcal{N}(\varepsilon/(2L), \Theta, \|\cdot\|)\,.$$

Proof. Let X be a ε-covering of Θ and let us define a set $F = \{[\ell = f_\theta - \varepsilon L, u = f_\theta + \varepsilon L] \mid \theta \in X\}$. Let us show that F consists of brackets. Let $f_\theta \in \mathcal{F}$, then there is $\widehat{\theta} \in X$. By Lipchitzness

$$\forall x \in \mathcal{X} : |f_\theta(x) - f_{\widehat{\theta}}(x)| \leqslant L\|\theta - \widehat{\theta}\| = L\varepsilon\,,$$

therefore a bracket that corresponding ℓ and u indeed satisfy $\ell(x) \leqslant f_\theta(x) \leqslant u(x)$ for any $x \in \mathcal{X}$. Also, we notice that $\|\ell - u\|_\infty = 2\varepsilon L$; therefore, we conclude the statement by rescaling. □

B.6.3 Proof for Demonstration-regularized RLHF

During this section, we assume that the MDP is finite, i.e., $|\mathcal{S}| < +\infty$, to simplify the manipulations with the trajectory space. However, the state space could be arbitrarily large.

In this section, we provide the proof for the demonstration-regularized RLHF pipeline defined in Algorithm 4. We start from the general oracle version of this inequality. For a reward function r let the value with respect to this value be defined as follows

$$V_h^\pi(s;r) = \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \mid s_h = s\right], \quad V_{\widetilde{\pi},\lambda,h}^{\pi}(s;r) = V_h^\pi(s;r) - \lambda \operatorname{KL}_{\mathrm{traj}}(\widetilde{\pi}\|\pi)\,.$$

A similar definition holds for Q-values.

Theorem B.39. *Let us assume that there is an underlying reward function $r^\star$ such that*

1. *There is an expert policy π^{E} such that $V_1^\star(s_1;r^\star) - V_1^{\pi^{\mathrm{E}}}(s_1;r^\star) \leqslant \varepsilon_{\mathrm{E}}$;*
2. *There is a behavior cloning policy π^{BC} that satisfies $\sqrt{\operatorname{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}}\|\pi^{\mathrm{BC}})} \leqslant \varepsilon_{\mathrm{KL}}$;*
3. *There is an estimate of reward function $\widehat{r}$ that satisfies $\sqrt{\operatorname{Var}_{q^{\pi^{\mathrm{BC}}}}[r^\star(\tau) - \widehat{r}(\tau)]} \leqslant \varepsilon_{\mathrm{RM}}$;*

Let π^{RL} be a $\varepsilon_{\mathrm{RL}}$-optimal policy in the λ-regularized MDP with π^{BC} and using $\widehat{r}$ as rewards:

$$V_{\pi^{\mathrm{BC}},\lambda,1}^{\star}(s_1;\widehat{r}) - V_{\pi^{\mathrm{BC}},\lambda,1}^{\pi^{\mathrm{RL}}}(s_1;\widehat{r}) \leqslant \varepsilon_{\mathrm{RL}}\,.$$

Then, for any $\lambda \geqslant H$ we have the following optimality guarantees for π^{RL} in the unregularized MDP equipped with the true reward function

$$V_1^\star(s_1;r^\star) - V_1^{\pi^{\mathrm{RL}}}(s_1;r^\star) \leqslant 3(\varepsilon_{\mathrm{E}} + \varepsilon_{\mathrm{RL}}) + 9/H \cdot \varepsilon_{\mathrm{RM}}^2 + (\lambda + 4H)\varepsilon_{\mathrm{KL}}^2\,.$$

Proof. We start from the following decomposition, using the assumption on the expert policy

$$V_1^\star(s_1;r^\star) - V_1^{\pi^{\mathrm{RL}}}(s_1;r^\star) \leqslant V_1^{\pi^{\mathrm{E}}}(s_1;r^\star) - V_1^{\pi^{\mathrm{RL}}}(s_1;r^\star) + \varepsilon_{\mathrm{E}}\,.$$

Next, we change the optimal reward function to its reward-shaped version $r^\star$ and apply the following decomposition

$$V_1^{\pi^{\mathrm{E}}}(s_1;r^\star) - V_1^{\pi^{\mathrm{RL}}}(s_1;r^\star) = \underbrace{V_1^{\pi^{\mathrm{E}}}(s_1;\widehat{r}) - V_1^{\pi^{\mathrm{RL}}}(s_1;\widehat{r})}_{(\mathbf{A})} + \underbrace{V_1^{\pi^{\mathrm{E}}}(s_1;r^\star - \widehat{r}) - V_1^{\pi^{\mathrm{RL}}}(s_1;r^\star - \widehat{r})}_{(\mathbf{B})}\,.$$

We start from the analysis of the term (**A**). By properties of the behavior cloning policy and the BPI policy π^{RL}

$$\begin{aligned}(\mathbf{A}) &= V_{\pi^{\mathrm{BC}},\lambda,1}^{\pi^{\mathrm{E}}}(s_1;\widehat{r}) + \lambda \operatorname{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}}\|\pi^{\mathrm{BC}}) - V_{\pi^{\mathrm{BC}},\lambda,1}^{\pi^{\mathrm{RL}}}(s_1;\widehat{r}) - \lambda \operatorname{KL}_{\mathrm{traj}}(\pi^{\mathrm{RL}}\|\pi^{\mathrm{BC}}) \\ &\leqslant V_{\pi^{\mathrm{BC}},\lambda,1}^{\star}(s_1;\widehat{r}) - V_{\pi^{\mathrm{BC}},\lambda,1}^{\pi^{\mathrm{RL}}}(s_1;\widehat{r}) + \lambda\varepsilon_{\mathrm{KL}}^2 \leqslant \varepsilon_{\mathrm{RL}} + \lambda\varepsilon_{\mathrm{KL}}^2\,.\end{aligned}$$

Next, we have to analyze the second term (**B**). We decompose it as follows

$$(\mathbf{B}) = \underbrace{V_1^{\pi^{\mathrm{E}}}(s_1;r^\star - \widehat{r}) - V_1^{\pi^{\mathrm{BC}}}(s_1;r^\star - \widehat{r})}_{(\mathbf{C})} + \underbrace{V_1^{\pi^{\mathrm{BC}}}(s_1;r^\star - \widehat{r}) - V_1^{\pi^{\mathrm{RL}}}(s_1;r^\star - \widehat{r})}_{(\mathbf{D})}\,.$$

We start from the analysis of (**D**). We can apply Lemma D.9 since the space of the trajectories is finite

$$\begin{aligned}(\mathbf{D}) &= \mathbb{E}_{q^{\pi^{\mathrm{BC}}}}[r^\star(\tau) - \widehat{r}(\tau)] - \mathbb{E}_{q^{\pi^{\mathrm{RL}}}}[r^\star(\tau) - \widehat{r}(\tau)] \\ &\leqslant \sqrt{2\operatorname{Var}_{q^{\pi^{\mathrm{BC}}}}[r^\star(\tau) - \widehat{r}(\tau)] \cdot \operatorname{KL}_{\mathrm{traj}}(\pi^{\mathrm{RL}}\|\pi^{\mathrm{BC}})} + \frac{H}{3}\operatorname{KL}_{\mathrm{traj}}(\pi^{\mathrm{RL}}\|\pi^{\mathrm{BC}})\,.\end{aligned}$$

Next, we notice that by an assumption on the reward estimate, we have

$$(\mathbf{D}) \leqslant \sqrt{2\varepsilon_{\mathrm{RM}}^2 \cdot \mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{RL}} \| \pi^{\mathrm{BC}})} + \frac{H}{3} \mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{RL}} \| \pi^{\mathrm{BC}}) \,.$$

Next, we have to estimate the trajectory KL-divergence between π^{RL} and π^{BC}.
Let us use the ε-optimality of the policy π^{RL} with respect to reward $\widehat{r}$ in the regularized MDP

$$V_1^{\pi^{\mathrm{E}}}(s_1; \widehat{r}) - \lambda \mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}} \| \pi^{\mathrm{BC}}) \leqslant V_1^{\pi^{\mathrm{RL}}}(s_1; \widehat{r}) + \varepsilon_{\mathrm{RL}} - \lambda \mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{RL}} \| \pi^{\mathrm{BC}}) \,.$$

By rerranging the terms we have

$$\begin{aligned}
\mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{RL}} \| \pi^{\mathrm{BC}}) &\leqslant \frac{1}{\lambda}\Big(V_1^{\pi^{\mathrm{RL}}}(s_1; \widehat{r}) - V_1^{\pi^{\mathrm{E}}}(s_1; \widehat{r}) + \varepsilon_{\mathrm{RL}}\Big) + \varepsilon_{\mathrm{KL}}^2 \\
&\leqslant \frac{\varepsilon_{\mathrm{E}} + \varepsilon_{\mathrm{RL}}}{\lambda} + \frac{1}{\lambda}\Big(V_1^{\pi^{\mathrm{RL}}}(s_1; \widehat{r} - r^\star) - V_1^{\pi^{\mathrm{E}}}(s_1; \widehat{r} - r^\star)\Big) + \varepsilon_{\mathrm{KL}}^2 \\
&= \frac{1}{\lambda}\Big(V_1^{\pi^{\mathrm{E}}}(s_1; r^\star - \widehat{r}) - V_1^{\pi^{\mathrm{RL}}}(s_1; r^\star - \widehat{r})\Big) + \varepsilon_{\mathrm{KL}}^2 + \frac{\varepsilon_{\mathrm{E}} + \varepsilon_{\mathrm{RL}}}{\lambda} \,.
\end{aligned}$$

Recalling that $(\mathbf{B}) \triangleq V_1^{\pi^{\mathrm{E}}}(s_1; r^\star - \widehat{r}) - V_1^{\pi^{\mathrm{RL}}}(s_1; r^\star - \widehat{r})$, we obtain the following recursion

$$(\mathbf{B}) \leqslant (\mathbf{C}) + \sqrt{\frac{2\varepsilon_{\mathrm{RM}}^2}{\lambda} \cdot \Big((\mathbf{B}) + \varepsilon_{\mathrm{E}} + \varepsilon_{\mathrm{RL}}\Big) + 2\varepsilon_{\mathrm{RM}}^2 \cdot \varepsilon_{\mathrm{KL}}^2} + \frac{H\varepsilon_{\mathrm{KL}}^2}{3} + \frac{H}{3\lambda}\Big((\mathbf{B}) + \varepsilon_{\mathrm{E}} + \varepsilon_{\mathrm{RL}}\Big) \,.$$

Next, we apply the bound $2\sqrt{ab} \leqslant a + b$ for $a, b > 0$

$$\sqrt{2\varepsilon_{\mathrm{RM}}^2\left(\frac{1}{\lambda} \cdot \Big((\mathbf{B}) + \varepsilon_{\mathrm{E}} + \varepsilon_{\mathrm{RL}}\Big) + \varepsilon_{\mathrm{KL}}^2\right)} \leqslant \frac{H}{3\lambda}\Big((\mathbf{B}) + \varepsilon_{\mathrm{E}} + \varepsilon_{\mathrm{RL}}\Big) + \frac{H}{3}\varepsilon_{\mathrm{KL}}^2 + \frac{3\varepsilon_{\mathrm{RM}}^2}{2H} \,.$$

Overall, we have

$$\left(1 - \frac{2H}{3\lambda}\right)(\mathbf{B}) \leqslant (\mathbf{C}) + \frac{3\varepsilon_{\mathrm{RM}}^2}{2H} + \frac{2H\varepsilon_{\mathrm{KL}}^2}{3} + \frac{2H}{3\lambda}(\varepsilon_{\mathrm{E}} + \varepsilon_{\mathrm{RL}}) \,.$$

Next, using the assumption that $\lambda \geqslant H$ we get $1 - 2H/(3\lambda) \geqslant 1/3$ and thus

$$(\mathbf{B}) \leqslant 3(\mathbf{C}) + \frac{9\varepsilon_{\mathrm{RM}}^2}{2H} + 2H\varepsilon_{\mathrm{KL}}^2 + 2(\varepsilon_{\mathrm{E}} + \varepsilon_{\mathrm{RL}}) \,.$$

Next, we analyze the term $(\mathbf{C})$. By Lemma D.9 we have

$$\begin{aligned}
(\mathbf{C}) &= \mathbb{E}_{q^{\pi^{\mathrm{E}}}}[r^\star(\tau) - \widehat{r}(\tau)] - \mathbb{E}_{q^{\pi^{\mathrm{BC}}}}[r^\star(\tau) - \widehat{r}(\tau)] \\
&\leqslant \sqrt{2\mathrm{Var}_{q^{\pi^{\mathrm{BC}}}}(r^\star - \widehat{r}) \cdot \mathrm{KL}_{\mathrm{traj}}(\pi^{\mathrm{E}} \| \pi^{\mathrm{BC}})} + \frac{H}{3}\mathrm{KL}_{\mathrm{traj}}\Big(\pi^{\mathrm{E}} \| \pi^{\mathrm{BC}}\Big) \leqslant \frac{3\varepsilon_{\mathrm{RM}}^2}{2H} + \frac{2H}{3}\varepsilon_{\mathrm{KL}}^2 \,,
\end{aligned}$$

where for the last inequality we applied $2\sqrt{ab} \leqslant a + b$. Overall, we have the final rates for any $\lambda \geqslant H$

$$V_1^\star(s_1; r^\star) - V_1^{\pi^{\mathrm{RL}}}(s_1; r^\star) \leqslant 3(\varepsilon_{\mathrm{E}} + \varepsilon_{\mathrm{RL}}) + \lambda\varepsilon_{\mathrm{KL}}^2 + 9/H \cdot \varepsilon_{\mathrm{RM}}^2 + 4H\varepsilon_{\mathrm{KL}}^2 \,.$$

□

Corollary (Restatement of Corollary 3.17)**.** *Let Assumption 4 hold. For $\varepsilon > 0$ and $\delta \in (0, 1)$, assume that an expert policy ε_{E} is $\varepsilon/15$-optimal and satisfies Assumption 3 in the linear case. Let π^{BC} be the behavioral cloning policy obtained using function sets described in Section 3.3 and let the set $\mathcal{G}$ be defined in Lemma B.33 for finite and in Lemma B.34 for linear setting, respectively.*

If the following two conditions hold

$$(1)\ N^{\mathrm{RM}} \geqslant \widetilde{\Omega}\Big(\zeta \widetilde{D}/\varepsilon\Big); \qquad (2)\ N^{\mathrm{E}} \geqslant \widetilde{\Omega}\Big(H^2 \widetilde{D}/\varepsilon\Big)$$

for $\widetilde{D} = SA$ / d in finite / linear MDPs, then demonstration-regularized RLHF based on `UCBVI-Ent+` */* `LSVI-UCB-Ent` *with parameters $\varepsilon_{\mathrm{RL}} = \varepsilon/15$, $\delta_{\mathrm{RL}} = \delta/3$ and $\lambda = \lambda^\star = \widetilde{\mathcal{O}}\Big(N^{\mathrm{E}}\varepsilon/(SAH)\Big)$ / $\widetilde{\mathcal{O}}\Big(N^{\mathrm{E}}\varepsilon/(dH)\Big)$ is (ε, δ)-PAC for BPI with demonstration in finite / linear MDPs with sample complexity*

$$\mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta) = \widetilde{\mathcal{O}}\left(\frac{H^6 S^3 A^2}{N^{\mathrm{E}}\varepsilon^2}\right) \textit{ (finite)} \qquad \mathcal{C}(\varepsilon, N^{\mathrm{E}}, \delta) = \widetilde{\mathcal{O}}\left(\frac{H^6 d^3}{N^{\mathrm{E}}\varepsilon^2}\right) \textit{ (linear)}.$$

Proof. By symmetry of the distribution $q^{\pi^{\mathrm{BC}}} \otimes q^{\pi^{\mathrm{BC}}}$, we have

$$\mathbb{E}_{\tau^0, \tau^1 \sim \pi^{\mathrm{BC}}}\left[\Big(r^\star(\tau_0) - r^\star(\tau_1) - (r(\tau_0) - r(\tau_1))\Big)^2\right] = 2\mathrm{Var}_{q^{\pi^{\mathrm{BC}}}}[r^\star - \widehat{r}]\,. \tag{B.18}$$

First of all, notice that under the assumption $N^{\mathrm{RM}} \geqslant \widetilde{\Omega}\Big(\zeta \widetilde{D}/\varepsilon\Big)$, Theorem B.30 combined with (B.18) imply $\varepsilon_{\mathrm{RM}}^2 \leqslant H\varepsilon/45$ with probability at least $1 - \delta/3$.

Next, notice that under the assumption $N^{\mathrm{E}} \geqslant \widetilde{\Omega}\Big(H^2 \widetilde{D}/\varepsilon\Big)$, behavior cloning guarantees imply $4H\varepsilon_{\mathrm{KL}}^2 \leqslant \varepsilon/5$ with probability at least $1-\delta/3$ (see Appendix B.2.2 and Appendix B.2.3). Then, the optimal choice $\lambda^\star$ for demonstration-regularized RL (see Theorem 3.10) implies $\lambda^\star = \varepsilon/(5\varepsilon_{\mathrm{KL}}^2) \geqslant H$, thus Theorem B.39 is applicable. By a union bound, we have with probability at least $1 - \delta$

$$V_1^\star(s_1; r^\star) - V_1^{\pi^{\mathrm{RL}}}(s_1; r^\star) \leqslant 3\Big(\underbrace{\varepsilon_{\mathrm{E}}}_{\leqslant \varepsilon/15} + \underbrace{\varepsilon_{\mathrm{RL}}}_{\leqslant \varepsilon/15}\Big) + \underbrace{\lambda^\star \varepsilon_{\mathrm{KL}}^2}_{\leqslant \varepsilon/5} + \underbrace{9/H \cdot \varepsilon_{\mathrm{RM}}^2}_{\leqslant \varepsilon/5} + \underbrace{4H\varepsilon_{\mathrm{KL}}^2}_{\leqslant \varepsilon/5} \leqslant \varepsilon\,.$$

□

Appendix C

Complements on Chapter 4

Contents

C.1 Detailed Algorithm Description

In this section, we first provide a fully-detailed algorithmic description of the strategy introduced in Section 4.3.1 (which had been summarized in an algorithm sketch titled Algorithm 5), and we then write closed-form expressions of the policy constructions 4.3 based on the strategies of Examples 4.7 and 4.8, as well as AdaHedge.

Closed-form expressions of the policy constructions. We first recall the statement (4.3) for the construction of policies: for all $t \geqslant 1$, all $h \in [H]$, and $s \in \mathcal{S}$,

$$\pi_{t,h}(\,\cdot \mid s) = \varphi_t\left(\left(\widehat{A}_{\tau,h}(s,\,\cdot\,)\right)_{\tau \in \mathcal{E}_{e_t} \cap [t-1]}\right).$$

We now illustrate this definition with the strategies of Examples 4.7 and 4.8, as well as with `AdaHedge`. A key observation to do so will be that, by definition of advantage functions and since $\widehat{A}_{\tau,h}(s,\,\cdot\,)$ is based on the policy $\pi_{\tau,h}$,

$$\forall \tau \in [T], \quad \forall s \in \mathcal{S}, \qquad \sum_{a \in \mathcal{A}} \pi_{\tau,h}(a \mid s)\, \widehat{A}_{\tau,h}(s,a) = 0\,.$$

Polynomial potential (Example 4.7). We denote by $(x)_+ = \max\{x,\, 0\}$ the non-negative part of $x \in \mathbb{R}$. We have $\varphi_1 \equiv (1/A,\, \ldots,\, 1/A)$ and for $t \geqslant 2$, whenever $\mathcal{E}_{e_t} \cap [t-1]$ contains at least one

element,

$$\pi_{t,h}(a \mid s) = \frac{\left(\sum_{\tau \in \mathcal{E}_{e_t} \cap [t-1]} \widehat{A}_{\tau,h}(s,a) - \sum_{a'' \in \mathcal{A}} \pi_{\tau,h}(a'' \mid s)\, \widehat{A}_{\tau,h}(s,a'')\right)_+^{2\ln(A)}}{\sum_{a' \in \mathcal{A}} \left(\sum_{\tau \in \mathcal{E}_{e_t} \cap [t-1]} \widehat{A}_{\tau,h}(s,a') - \sum_{a'' \in \mathcal{A}} \pi_{\tau,h}(a'' \mid s)\, \widehat{A}_{\tau,h}(s,a'')\right)_+^{2\ln(A)}}$$
$$= \frac{\left(\sum_{\tau \in \mathcal{E}_{e_t} \cap [t-1]} \widehat{A}_{\tau,h}(s,a)\right)_+^{2\ln(A)}}{\sum_{a' \in \mathcal{A}} \left(\sum_{\tau \in \mathcal{E}_{e_t} \cap [t-1]} \widehat{A}_{\tau,h}(s,a')\right)_+^{2\ln(A)}}.$$

Exponential potential (Example 4.8). Similarly to above, we have $\varphi_1 \equiv (1/A, \ldots, 1/A)$ and for $t \geqslant 2$, whenever $\mathcal{E}_{e_t} \cap [t-1]$ contains at least one element,

$$\pi_{t,h}(a \mid s) = \frac{\exp\left(\eta_t \sum_{\tau \in \mathcal{E}_{e_t} \cap [t-1]} \widehat{A}_{\tau,h}(s,a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\eta_t \sum_{\tau \in \mathcal{E}_{e_t} \cap [t-1]} \widehat{A}_{\tau,h}(s,a')\right)} \qquad \text{where} \qquad \eta_t = \frac{1}{H+1}\sqrt{\frac{\ln(A)}{\left|\mathcal{E}_{e_t} \cap [t-1]\right|}}$$

are time-varying learning rates, based on the cardinality $\left|\mathcal{E}_{e_t} \cap [t-1]\right|$ of $\mathcal{E}_{e_t} \cap [t-1]$.

Adaptive versions of exponential-potential-based strategies. The literature proposed many ways of setting the learning rates for exponential potentials based on past information—a series of work initiated by Auer et al. (2002), whose learning rates were used in the paragraph above. One may cite, among (many) others, Cesa-Bianchi et al. (2007), Erven et al. (2011), Rooij et al. (2014), Orabona and Pál (2015); sometimes, the resulting strategy is called `AdaHedge`. For instance, Orabona (2019, Section 7.6) summarizes this literature by the following learning rates:

$$\eta_t = \frac{\max\{4, 2^{-1/4}\sqrt{\ln(A)}\}}{\sqrt{\sum_{\tau \in \mathcal{E}_{e_t} \cap [t-1]} \max_{a' \in \mathcal{A}} \left(\widehat{A}_{\tau,h}(s,a')\right)^2}}.$$

These updates correspond to an OLO strategy satisfying the bound of Definition 4.6 with a performance bound $B_{T,K} = 4\sqrt{T\ln(K)}$.

Space and computational complexity. We begin by outlining the space and computational complexity of the policy optimization strategy discussed above. In both examples (Example 4.7-4.8), it is clear that the computation of the policy relies solely on the sum of advantage estimates over the current epoch. These can be stored with a space complexity of $\mathcal{O}(HSA)$, and the computational complexity of updating the policy is $\mathcal{O}(HSA)$ per step, as it involves two key operations: (1) updating the sum of advantage estimates, and (2) updating the policy via a closed-form formula.

Now, the overall computations for one iteration of the final algorithm, outlined in Algorithm 10, consists of two main phases: (1) the dynamic programming update, which requires $\mathcal{O}(S^2AH)$

Algorithm 10 Adversarial Policy Optimization based on Monotonic Value Propagation (`APO-MVP`)

Input: Number of rounds T, number of states, actions and horizon S, A, H, confidence level δ, online linear optimization strategy $\varphi = (\varphi_t)_{t \geqslant 1}$.
Initialize kernels $\widehat{P}_h(s'|s,a) = 1/S$ for all $(s', s, a, h) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A} \times [H-1]$;
Initialize counters $n_h(s,a,s') = n_h(s,a) = 0$ for all $(s', s, a, h) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A} \times [H-1]$;
Initialize histories $\mathcal{H}_{h,s} = \emptyset$ for all $h \in [H]$ and $s \in \mathcal{S}$;
Initialize $\pi_{1,h}(\cdot \mid s) = \varphi_1(\emptyset)$ for all $h \in [H]$ and $s \in \mathcal{S}$;
Select initial state $s_1 \in \mathcal{S}$;
for rounds $t = 1, \ldots, T$ **do**
  `// Interaction`
  Set $s_{t,1} = s_1$;
  for $h = 1, \ldots, H$ **do**
    Play action $a_{t,h} \sim \pi_{t,h}(\cdot \mid s_{t,h})$;
    Receive next state $s_{t,h+1} \sim P_h(\cdot \mid s_{t,h}, a_{t,h})$;
    Update counters $n_h(s_{t,h}, a_{t,h})$ += 1 and $n_h(s_{t,h}, a_{t,h}, s_{t,h+1})$ += 1;
    `// Trigger, update the model`
    if $n_h(s_{t,h}, a_{t,h}) = 2^{\ell-1}$ for some $\ell \geqslant 1$ **then**
      $\widehat{P}_h(s'|s_{t,h}, a_{t,h}) = \frac{n_h(s_{t,h}, a_{t,h}, s')}{2^{\ell-1}}$ for all $s' \in \mathcal{S}$;
      $b_h(s_{t,h}, a_{t,h}) = \sqrt{\frac{2H^2 \ln(2SATH \log_2(2T)/\delta)}{2^{\ell-1}}} \wedge H$;
      Activate `trigger`;
    end if
  end for
  Receive reward function $\boldsymbol{r}_t = (r_{t,h})_{h \in [H]}$;
  if `trigger` **then**
    Drop all histories, i.e., set $\mathcal{H}_{h,s} = \emptyset$ for all $h \in [H]$ and $s \in \mathcal{S}$;
    Set $\pi_{t+1,h}(\cdot \mid s) = \varphi_t(\emptyset)$ for all $h \in [H]$ and $s \in \mathcal{S}$;
    Deactivate `trigger`;
  else
    `// Compute first the advantage functions via Bellman's equations`
    Let $\widehat{Q}_H(s,a) = r_{t,H}(s,a)$ and $\widehat{V}_h(s) = \pi_{t,H}\widehat{Q}_H(s)$ for each $(s,a) \in \mathcal{S} \times \mathcal{A}$;
    for $h = H-1, H-2, \ldots, 1$ **do**
      Let $\widehat{Q}_h(s,a) = r_{t,h}(s,a) + b_h(s,a) + \widehat{P}_h\widehat{V}_{h+1}(s,a)$ and $\widehat{V}_h(s) = \pi_{t,h}\widehat{Q}_h(s)$ for each $(s,a) \in \mathcal{S} \times \mathcal{A}$;
    end for
    Let $\widehat{A}_h(s,a) = \widehat{Q}_h(s,a) - \widehat{V}_h(s)$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$;
    Add $\left(\widehat{A}_h(s,a)\right)_{a \in \mathcal{A}}$ to the history $\mathcal{H}_{h,s}$ for each $(h,s) \in [H] \times \mathcal{S}$;
    `// Next, obtain` $\boldsymbol{\pi}_{t+1}$ `via the learning strategy`
    Let $\pi_{t+1,h}(\cdot \mid s) = \varphi_t(\mathcal{H}_{h,s})$ for all $(h,s) \in [H] \times \mathcal{S}$;
  end if
end for
Output: Sequence of policies $\boldsymbol{\pi}^t = (\pi_{t,h})_{h \in [H]}$, for $t \in [T]$.

calculations in the worst case, and (2) the policy optimization update, which requires $\mathcal{O}(SAH)$ calculations, as previously described. The space complexity of the algorithm is standard for model-based reinforcement learning algorithms and is equal to $\mathcal{O}(S^2AH)$.

In summary, the algorithm introduces no additional computational and space overhead beyond

that of standard dynamic programming methods.

C.2 Term (B)

It only remains to prove Lemma 4.9, which we restate below. For the sake of self-completeness, we copy the proofs by Jonckheere et al. (2025). As noted by Jonckheere et al. (2025, Remark 1), the proof below may be adapted to strategies run with Q–functions rather than with advantage functions: (C.1) is then replaced by

$$2M(H-h+1)\,B_{T,A} \geqslant \max_{a\in\mathcal{A}} \sum_{t=1}^{T} Q_h^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s,a) - \sum_{t=1}^{T} \underbrace{\sum_{a\in\mathcal{A}} \pi'_{t,h}(a|s)\, Q_h^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s,a)}_{=V_h^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s)}$$

$$= \max_{a\in\mathcal{A}} \sum_{t=1}^{T} A_h^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s,a)\,,$$

and the rest of the proof is unchanged. However, advantage functions are preferred in practice.

Lemma 4.9 (Jonckheere et al. 2025)**.** *If the learning strategy satisfies the conditions of Definition 4.6, then the sequence of policies defined right above is such that, for all fixed policies* $\boldsymbol{\pi} = (\pi_h)_{h\in[H]}$*, for all* $T \geqslant 1$*,*

$$\sum_{t=1}^{T}\Big(V_1^{\boldsymbol{\pi},\boldsymbol{r}'_t,\boldsymbol{P}'}(s_1) - V_1^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s_1)\Big) \leqslant 2MH^2\,B_{T,A}\,.$$

Proof. As the reward function takes values in $[0,M]$, we have that $\left|A_{\tau,h}^{\boldsymbol{\pi}'_\tau}(s,a)\right| \leqslant M(H-h+1)$. By the definition of advantage functions (for the equality to 0) and by Definition 4.6 (for the upper bound), we have, for all $s\in\mathcal{S}$,

$$\max_{a\in\mathcal{A}} \sum_{t=1}^{T} A_h^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s,a) - \sum_{t=1}^{T} \underbrace{\sum_{a\in\mathcal{A}} \pi'_{t,h}(a|s)\, A_h^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s,a)}_{=0} \leqslant 2M(H-h+1)\,B_{T,A}\,. \tag{C.1}$$

Now, the so-called performance difference lemma (see, e.g., Kakade and Langford (2002) for the result in the discounted setting) shows that

$$V_1^{\boldsymbol{\pi},\boldsymbol{r}'_t,\boldsymbol{P}'}(s_1) - V_1^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s_1) = \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} \mu_h^{\boldsymbol{\pi},\boldsymbol{P}',s_1}(s) \sum_{a\in\mathcal{A}} \pi_h(a|s)\, A_h^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s,a)$$

where $\mu_h^{\boldsymbol{\pi},\boldsymbol{P}',s_1}$ is the distribution of $s_{t,h}$ induced in the h–th episode by $\boldsymbol{\pi}$ given the state transitions $\boldsymbol{P}'$ and the initial state s_1. Summing this equality over t and rearranging, we get

$$\begin{aligned}
\sum_{t=1}^{T}\Big(V_1^{\boldsymbol{\pi},\boldsymbol{r}'_t,\boldsymbol{P}'}(s_1) - V_1^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s_1)\Big) &= \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} \mu_h^{\boldsymbol{\pi},\boldsymbol{P}',s_1}(s) \sum_{a\in\mathcal{A}} \pi_h(a|s) \sum_{t=1}^{T} A_h^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s,a) \\
&\leqslant \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} \mu_h^{\boldsymbol{\pi},\boldsymbol{P}',s_1}(s) \underbrace{\max_{a\in\mathcal{A}} \sum_{t=1}^{T} A_h^{\boldsymbol{\pi}'_t,\boldsymbol{r}'_t,\boldsymbol{P}'}(s,a)}_{\leqslant 2M(H-h+1)\,B_{T,A}} \\
&\leqslant 2M \sum_{h=1}^{H} (H-h+1)\,B_{T,A} = MH(H+1)\,B_{T,A} \\
&\leqslant 2MH^2\,B_{T,A}\,,
\end{aligned}$$

where we substituted (C.1). Here, we crucially used that the convex combination with weights $\mu_h^{\boldsymbol{\pi},\boldsymbol{P}',s_1}(s)$ is independent of t and only depends on the fixed benchmark policy $\boldsymbol{\pi}$, on the state transitions $\boldsymbol{P}'$, and on the initial states s_1 (identical for all t). □

C.3 Term (A)

We start with the following consequence of Hoeffding's inequality.

Lemma C.1. *For each $t \in [T]$, for each $h \in [H-1]$, for each $(s,a) \in \mathcal{S} \times \mathcal{A}$, for each $\ell \in \left[\lceil \log_2(T) \rceil\right]$,*

$$\mathbb{P}\left\{ \left| \frac{1}{2^{\ell-1}} \sum_{j=1}^{2^{\ell-1}} V_{h+1}^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t,\boldsymbol{P}}(\sigma_{h,s,a,j}) - P_h V_{h+1}^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t,\boldsymbol{P}}(s,a) \right| > \sqrt{\frac{2H^2 \ln(2SATH\log_2(2T)/\delta)}{2^{\ell-1}}} \wedge H \right\} \leqslant \frac{\delta}{SATH\log_2(2T)},$$

where $(\sigma_{h,s,a,j})_{1 \leqslant j \leqslant 2^{\ell-1}}$ is a sequence of i.i.d. variables with distribution $P_h(\cdot \mid s,a)$.

Proof. The policy $\boldsymbol{\pi}^\star$ is fixed, as it only depends on the $\boldsymbol{r}_t$ and $\boldsymbol{P}$, which are all fixed beforehand. The function $g = V_{h+1}^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t,\boldsymbol{P}}$ is therefore a fixed deterministic function. The expectation of $g(\sigma_{h,s,a,j})$ is indeed, given our notation, $P_h g$. By the boundedness of rewards in $[0,1]$, and thus the boundedness of values in the range $[0,H]$, we may therefore apply Hoeffding's inequality: we do so for each $t \in [T]$, each $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H-1]$, and each $\ell \in \left[\lceil \log_2(T) \rceil\right]$, and get that for all $\delta' \in (0,1)$, with probability at least $1-\delta'$,

$$\left| \frac{1}{2^{\ell-1}} \sum_{j=1}^{2^{\ell-1}} V_{h+1}^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t,\boldsymbol{P}}(\sigma_{h,s,a,j}) - P_h V_{h+1}^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t,\boldsymbol{P}}(s,a) \right| \leqslant \sqrt{\frac{2H^2 \ln(2/\delta')}{2^{\ell-1}}}.$$

The proof is concluded by keeping in mind that the left-hand side necessarily belongs to $[0,H]$ by boundedness of values in $[0,H]$. □

We are now ready to prove Lemma 4.13, which we restate below; we do so by mimicking the proof of Azar et al. (2017, Lemma 18).

Lemma 4.13. *With probability at least $1-\delta$, for all $t \in [T]$ and all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$,*

$$Q_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t,\boldsymbol{P}}(s,a) \leqslant Q_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s,a), \qquad V_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t,\boldsymbol{P}}(s) \leqslant V_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s).$$

Specifically, with probability $\geqslant 1-\delta$, we have $(\mathbf{A}) \leqslant 0$.

Proof. We proceed by backward induction, for each given $t \in [T]$; more precisely, we consider, for $h \in [H]$, the induction hypothesis

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}, \qquad Q_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t,\boldsymbol{P}}(s,a) \leqslant Q_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s,a) \qquad (\mathcal{H}_h)$$
$$\text{and} \qquad V_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t,\boldsymbol{P}}(s) \leqslant V_h^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s).$$

For $h = H$, we note that for all (s,a),

$$Q_H^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t,\boldsymbol{P}}(s,a) = r_{t,H}(s,a) = r_{t,H}(s,a) + b_{t,H}(s,a) = Q_H^{\boldsymbol{\pi}^\star,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s,a),$$

so that $(\mathcal{H}_H)$ is trivially satisfied. For $h \in [H-1]$, by Bellman equations,

$$Q_h^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s,a) - Q_h^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) = b_{t,h}(s,a) + \widehat{P}_{t,h} V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s,a) - P_h V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a)\,,$$

where by the induction hypothesis $(\mathcal{H}_{h+1})$, we have $V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}^t}(s') \geqslant V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s')$ for any s'. Thus,

$$Q_h^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s,a) - Q_h^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) \geqslant b_{t,h}(s,a) + \Big(\widehat{P}_{t,h} - P_h\Big) V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a)\,, \tag{C.2}$$

and a similar inequality for values, as the latter are obtained as convex combinations of Q–values, where the convex weights are determined solely by the common policy $\boldsymbol{\pi}^\star$ used. Therefore, $(\mathcal{H}_h)$ holds at least on the event

$$\mathcal{G}_{t,h} \triangleq \left\{ \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \quad \left| \Big(\widehat{P}_{t,h} - P_h\Big) V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) \right| \leqslant b_{t,h}(s,a) \right\}.$$

All in all, the inequalities required in the statement of the lemma thus hold on the intersection of the events $\mathcal{G}_{t,h}$ over $t \in [T]$ and $h \in [H-1]$.

To conclude the proof, it suffices to show that this intersection is of probability at least $1-\delta$. By considering the complements and by a union bound, it suffices to show that any event

$$\overline{\mathcal{G}}_{t,h,s,a} \triangleq \left\{ \left| \Big(\widehat{P}_{t,h} - P_h\Big) V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) \right| > b_{t,h}(s,a) \right\}$$

is of probability at most $\delta/(SATH)$. We partition the probability space based on the value of $\ell_{t-1,h}(s,a)$, resort to optional skipping and Corollary 4.12, with the deterministic function $g = V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}$ (see the proof of Lemma C.1), to get the first inequality below, and to Lemma C.1 for the second inequality below. We also use the definition (4.5) of $b_{t,h}$:

$$\begin{aligned}
\mathbb{P}\big(\overline{\mathcal{G}}_{t,h,s,a}\big) &= \sum_{\ell=0}^{\lceil \log_2(T) \rceil} \mathbb{P}\Big(\overline{\mathcal{G}}_{t,h,s,a} \cap \big\{\ell_{t-1,h}(s,a) = \ell\big\}\Big) \\
&= \sum_{\ell=1}^{\lceil \log_2(T) \rceil} \mathbb{P}\left\{ \left| (\widehat{P}_{t,h} - P_h) V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) \right| > \sqrt{\frac{2H^2 \ln(2SATH \log_2(2T)/\delta)}{2^{\ell-1}}} \wedge H \quad \text{and} \quad \ell_{t-1,h}(s,a) = \ell \right\} \\
&\leqslant \sum_{\ell=1}^{\lceil \log_2(T) \rceil} \mathbb{P}\left\{ \left| \frac{1}{2^{\ell-1}} \sum_{j=1}^{2^{\ell-1}} V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(\sigma_{h,s,a,j}) - P_h V_{h+1}^{\boldsymbol{\pi}^\star, \boldsymbol{r}_t, \boldsymbol{P}}(s,a) \right| > \sqrt{\frac{2H^2 \ln(2SATH \log_2(2T)/\delta)}{2^{\ell-1}}} \wedge H \right\} \\
&\leqslant \lceil \log_2(T) \rceil \, \frac{\delta}{SATH \log_2(2T)} \leqslant \frac{\delta}{SATH}\,,
\end{aligned}$$

where we used the fact that $b_{t,h}(s,a) = H$ is a trivial upper bound on the difference of values at hand in the case $\ell = 0$, which is why the element $\ell = 0$ gets dropped in the summation in the second equality. □

C.4 Term (D)

We first restate and then prove Lemma 4.14.

Lemma 4.14. *With probability at least $1-\delta$, we have*

$$(\mathbf{D}) \leqslant 7\sqrt{H^4 SAT \ln(2SATH \log_2(2T)/\delta)} + H^2 SA\,.$$

Proof. Since $b_{t,h}(s,a) \in [0,H]$, the Hoeffding-Azuma inequality implies that with probability at least $1-\delta$,

$$\sum_{t=1}^{T} V_1^{\boldsymbol{\pi}_t, \boldsymbol{b}_t, \boldsymbol{P}}(s_1) - \sum_{t=1}^{T} \sum_{h\in[H]} b_{t,h}(s_{t,h}, a_{t,h}) \leqslant \sqrt{\frac{(H^2)^2\, T \ln(1/\delta)}{2}}\,; \tag{C.3}$$

we crucially use here that the policies $\boldsymbol{\pi}_t$ only depend on information gathered during previous episodes $\tau \leqslant t-1$ and that the stochastic environment $\boldsymbol{P}$ considered in the definition of (**D**) is the true underlying environment.

We fix $h \in [H-1]$ (recall that $b_{t,H} \equiv 0$) and a pair $(s',a') \in \mathcal{S}\times\mathcal{A}$, and show that

$$\sum_{t=1}^{T} b_{t,h}(s_{t,h},a_{t,h})\, \mathbb{1}\Big\{(s_{t,h},,a_{t,h}) = (s',a')\Big\} \leqslant H + \sqrt{2H^2 \ln(2SATH\log_2(2T)/\delta)\, n_{T,h}(s',a')}\,. \tag{C.4}$$

Indeed, there can only be at most one t such $(s_{t,h},,a_{t,h}) = (s',a')$ and $n_{t,h}(s',a') = 1$; for this t, we use the upper bound $b_{t,h}(s_{t,h},a_{t,h}) \leqslant H$. For t such that $n_{t,h}(s',a') \geqslant 2$, we have, by the definitions in Section 4.3.1, that $2^{\ell_{t,h}(s',a')} \geqslant n_{t,h}(s',a')$ and $\ell_{t-1,h}(s',a')+1 \geqslant \ell_{t,h}(s',a') \geqslant \ell_{t-1,h}(s',a')$. Therefore, $2^{\ell_{t-1,h}(s',a')-1} \geqslant 2^{\ell_{t,h}(s',a')-2} \geqslant n_{t,h}(s',a')/4$. Substituting this inequality in the definition (4.5) of $b_{t,h}$, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T} b_{t,h}(s_{t,h},a_{t,h})\, \mathbb{1}\Big\{(s_{t,h},,a_{t,h}) = (s',a')\Big\} \\
&\qquad \leqslant H + \sum_{t=1}^{T} \sqrt{\frac{8H^2 \ln(2SATH\log_2(2T)/\delta)}{n_{t,h}(s',a')}}\, \mathbb{1}\Big\{(s_{t,h},,a_{t,h}) = (s',a')\Big\}\, \mathbb{1}\Big\{n_{t,h}(s',a') \geqslant 2\Big\}\,,
\end{aligned}$$

where, using that the counters $n_{t,h}(s',a')$ vary (by $+1$) if and only if $(s_{t,h},,a_{t,h}) = (s',a')$, we also have

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n_{t,h}(s',a')}}\, \mathbb{1}\Big\{(s_{t,h},,a_{t,h}) = (s',a')\Big\}\, \mathbb{1}\Big\{n_{t,h}(s',a') \geqslant 2\Big\} = \sum_{n=2}^{n_{T,h}(s',a')} \frac{1}{\sqrt{n}} \leqslant 2\sqrt{n_{T,h}(s',a')}\,.$$

We conclude the proof by noting first that for each $(s',a') \in \mathcal{S}\times\mathcal{A}$, by concavity of the root,

$$\sum_{(s',a')} \sqrt{n_{T,h}(s',a')} \leqslant \sqrt{SAT}\,,$$

so that summing (C.4) over $h \in [H-1]$ and $(s',a') \in \mathcal{S}\times\mathcal{A}$ yields

$$\sum_{t=1}^{T} \sum_{h\in[H]} b_{t,h}(s_{t,h},a_{t,h}) \leqslant H^2 SA + 4H\sqrt{2H^2 SAT \ln(2SATH\log_2(2T)/\delta)}\,.$$

We combine this inequality with (C.3) and note that

$$\sqrt{\frac{H^4\, T \ln(1/\delta)}{2}} \leqslant \sqrt{H^4 SAT \ln(2SATH\log_2(2T)/\delta)}$$

to get the claimed bound. □

C.5 Term (C)

This section is devoted to the analysis of the term (**C**), which is the most involved part of the proof. We leverage the recently developed techniques of Zhang et al. (2024). We start by restating the claimed bound.

Lemma 4.15. *With probability at least* $1-\delta$,

$$(\mathbf{C}) \leqslant 2\sqrt{H^7SAT\Big(\log_2(2T)\Big)^3} + 2\sqrt{2H^6\,T\log_2(2T)\ln(4H/\delta)} + SAH^3\,.$$

The proof starts with an application of the performance-difference lemma in case of different transition kernels (see, e.g., Russo 2019, Lemma 3):

$$V_1^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s_1) - V_1^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\boldsymbol{P}}(s_1) = \mathbb{E}_{\boldsymbol{\pi}_t,\boldsymbol{P}}\left[\sum_{h=1}^{H-1}\Big(\widehat{P}_{t,h}-P_h\Big)V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s'_h,a'_h)\right],$$

where the piece of notation $\mathbb{E}_{\boldsymbol{\pi}_t,\boldsymbol{P}}$ indicates (as in Section 4.2) that the expectation is taken over trajectories $(s'_1,a'_1,\,\ldots,\,s'_H,a'_H)$ started at $s'_1=s_1$ and induced by the policies $\boldsymbol{\pi}_t$ and the transition kernels $\boldsymbol{P}$. Actually, one such trajectory is exactly $(s_{t,1},a_{t,1},\,\ldots,\,s_{t,H},a_{t,H})$ and we could rewrite the considered expectation as a conditional expectation:

$$\mathbb{E}_{\boldsymbol{\pi}_t,\boldsymbol{P}}\left[\sum_{h=1}^{H-1}\Big(\widehat{P}_{t,h}-P_h\Big)V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s'_h,a'_h)\right] = \mathbb{E}\left[\sum_{h=1}^{H-1}\Big(\widehat{P}_{t,h}-P_h\Big)V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s_{t,h},a_{t,h})\,\middle|\,\boldsymbol{\pi}_t,\,\boldsymbol{b}_t,\,\widehat{\boldsymbol{P}}_t\right].$$

Next, we apply the Hoeffding–Azuma inequality, by resorting to a lexicographic order on pairs (t,h) and by noting that the random variables at hand satisfy

$$\Big(\widehat{P}_{t,h}-P_h\Big)V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s_{t,h},a_{t,h}) \in [-H^2,H^2]\,;$$

indeed, the sums $r_{t,h}+b_{t,h}$ lies in $[0,H+1]$ and the value functions are weighted sums of at most $H-1$ such terms. We obtain that with probability at least $1-\delta/2$,

$$\begin{aligned}(\mathbf{C}) &= \sum_{t=1}^{T} V_1^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s_1) - V_1^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\boldsymbol{P}}(s_1)\\ &\leqslant \sum_{t=1}^{T}\sum_{h=1}^{H-1}\Big(\widehat{P}_{t,h}-P_h\Big)V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s_{t,h},a_{t,h}) + \sqrt{2H^5T\ln(2/\delta)}\,. \qquad \text{(C.5)}\end{aligned}$$

We fix $h\in[H-1]$ and use the decomposition

$$\begin{aligned}&\sum_{t=1}^{T}\Big(\widehat{P}_{t,h}-P_h\Big)V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s_{t,h},a_{t,h})\\ &\qquad\leqslant SAH^2 + \sum_{t=1}^{T}\sum_{(s,a)}\sum_{\ell=1}^{\lceil\log_2 T\rceil}\sum_{j=1}^{2^{\ell-1}}\Big(\widehat{P}_{t,h}-P_h\Big)V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s,a)\,\mathbb{1}\Big\{(s_{t,h},a_{t,h})=(s,a)\Big\}\\ &\qquad\qquad\times\mathbb{1}\Big\{n_{t,h}(s,a)=2^{\ell-1}+j\Big\}\,; \qquad \text{(C.6)}\end{aligned}$$

the term SAH^2 comes from the fact that for each pair (s,a), there is at most once round t when $(s_{t,h},a_{t,h})=(s,a)$ and $n_{t,h}(s,a)=1$. We prove below the following lemma.

Lemma C.2. *For each pair* (ℓ,j), *where* $\ell\in\Big[\lceil\log_2 T\rceil\Big]$ *and* $j\in[2^{\ell-1}]$, *with probability at least* $1-\delta/(4TH)$,

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{(s,a)}\Big(\widehat{P}_{t,h}-P_h\Big)V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s,a)\,\mathbb{1}\Big\{(s_{t,h},a_{t,h})=(s,a)\Big\}\,\mathbb{1}\Big\{n_{t,h}(s,a)=2^{\ell-1}+j\Big\}\\ &\qquad\leqslant\sqrt{2H^4\,\frac{1}{2^{\ell-1}}\sum_{(s,a)}\mathbb{1}\Big\{n_{T,h}(s,a)\geqslant 2^{\ell-1}+j\Big\}\ln\left(\frac{4H(T+1)^{1+SAH\lceil\log_2(T)\rceil}}{\delta}\right)}\,.\end{aligned}$$

Remark C.3. Lemma C.2 is established for a fixed h, with the derived bound scaling as $\sqrt{H^5}$ in terms of the episode length. Consequently, the final bound on term-(**C**) scales as $H\sqrt{H^5} = \sqrt{H^7}$ by summing the bound of Lemma C.2 over $h \in [H]$.

A potential refinement of the bound on term (**C**) involves incorporating the summation over h appearing in (C.5) and proving a strengthened version of Lemma C.2. In particular, taking the sum over h into account (instead of a fixed value h) would replace $\sqrt{H^5}$ in Lemma C.2 by $\sqrt{H^6}$. Since no more summation over h is needed, the final bound would also scale as $\sqrt{H^6}$.

For the sake of simplicity and because the control of term (**B**) leads anyway to a factor $\sqrt{H^7}$ in the final regret bound, we do not pursue this direction.

We conclude the proof of Lemma 4.15 based on Lemma C.2. There are at most $2T$ different pairs (ℓ, j) considered, so that all events considered in Lemma C.2 hold simultaneously with probability at least $1 - \delta/(2H)$. In addition, a first application of Jensen's inequality guarantees that for each $1 \leqslant \ell \leqslant \lceil \log_2 T \rceil$,

$$\sum_{j=1}^{2^{\ell-1}} \sqrt{\frac{1}{2^{\ell-1}} \sum_{(s,a)} \mathbb{1}\Big\{n_{T,h}(s,a) \geqslant 2^{\ell-1} + j\Big\}} \leqslant \sqrt{\sum_{(s,a)} \sum_{j=1}^{2^{\ell-1}} \mathbb{1}\Big\{n_{T,h}(s,a) \geqslant 2^{\ell-1} + j\Big\}},$$

and a second application yields

$$\begin{aligned}
&\sum_{\ell=1}^{\lceil \log_2 T \rceil} \sum_{j=1}^{2^{\ell-1}} \sqrt{\frac{1}{2^{\ell-1}} \sum_{(s,a)} \mathbb{1}\Big\{n_{T,h}(s,a) \geqslant 2^{\ell-1} + j\Big\}} \\
&\qquad \leqslant \sqrt{\lceil \log_2 T \rceil \sum_{(s,a)} \underbrace{\sum_{\ell=1}^{\lceil \log_2 T \rceil} \sum_{j=1}^{2^{\ell-1}} \mathbb{1}\Big\{n_{T,h}(s,a) \geqslant 2^{\ell-1} + j\Big\}}_{= n_{T,h}(s,a) - 1}} \leqslant \sqrt{T \lceil \log_2 T \rceil}\,.
\end{aligned}$$

Therefore, substituting the bound above (together with Lemma C.2) into (C.6), we proved so far that with probability at least $1 - \delta/2$,

$$\begin{aligned}
&\sum_{t=1}^{T} \Big(\widehat{P}_{t,h} - P_h\Big) V_{h+1}^{\boldsymbol{\pi}_t, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s_{t,h}, a_{t,h}) \\
&\qquad \leqslant SAH^2 + \sqrt{2H^4\, T \lceil \log_2 T \rceil \ln\left(\frac{4H(T+1)^{1+SAH\lceil \log_2(T) \rceil}}{\delta}\right)} \\
&\qquad \leqslant SAH^2 + 2\sqrt{H^5 SA\, T \Big(\log_2(2T)\Big)^3} + \sqrt{2H^4\, T \lceil \log_2 T \rceil \ln(4H/\delta)}\,.
\end{aligned}$$

Summing this bound over $h \in [H-1]$ and combining the outcome with (C.5) leads to

$$(\mathbf{C}) \leqslant SAH^3 + 2\sqrt{H^7 SA\, T \Big(\log_2(2T)\Big)^3} + \sqrt{2H^6\, T \lceil \log_2 T \rceil \ln(4H/\delta)} + \sqrt{2H^5 T \ln(2/\delta)}\,,$$

and thus to the upper bound claimed in Lemma 4.15.

It therefore only remains to prove Lemma C.2.

Proof. We denote by

$$\tau_{\ell,j,h}(s,a) \triangleq \begin{cases} t & \text{if } (s_{t,h}, a_{t,h}) = (s,a) \text{ and } n_{t,h}(s,a) = 2^{\ell-1} + j\,, \\ +\infty & \text{if } n_{T,h}(s,a) \leqslant 2^{\ell-1} + j - 1\,, \end{cases}$$

the stopping time whether and when (s,a) was reached in stage h for the $(2^{\ell-1}+j)$–th time, with the convention $\tau_{\ell,j,h}(s,a)=+\infty$ if (s,a) was reached fewer times than that. To apply optional skipping, we will partition the underlying probability space according to the values of all the $\ell_{t',h'}(s',a')$ as t',h',s',a' vary and of the $\tau_{\ell,j,h}(s',a')$ as s',a' only vary.

Part 1: Hoeffding-Azuma inequality. We fix consistent sequences $k_{t',h'}(s',a')\in[T]$ and $\kappa_{\ell,j,h}(s',a')$ of values for the $\ell_{t',h'}(s',a')$ and the $\tau_{\ell,j,h}(s',a')$; in particular, $k_{\kappa_{\ell,j,h}(s,a)-1,h}(s,a)=\ell$. The notation in the display below is heavy but the high-level idea is simple to grasp: only rounds $t=\kappa_{\ell,j,h}(s,a)$ matter, and we know to which global epoch each of these rounds belongs and, in particular, we know which averages are in the components $\widehat{P}_{t,h}(\,\cdot\mid s',a')$ of $\widehat{P}_{t,h}$.

We rewrite the quantity at hand on the event associated with the sequences fixed:

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{(s,a)}\Big(\widehat{P}_{t,h}-P_h\Big)V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}(s,a)\,\mathbb{1}\Big\{(s_{t,h},a_{t,h})=(s,a)\Big\}\,\mathbb{1}\Big\{n_{t,h}(s,a)=2^{\ell-1}+j\Big\}\\
&\times\prod_{(s',a')\in\mathcal{S}\times\mathcal{A}}\left(\mathbb{1}\Big\{\tau_{\ell,j,h}(s',a')=\kappa_{\ell,j,h}(s',a')\Big\}\prod_{t'=1}^{T}\prod_{h'\in[H-1]}\mathbb{1}\Big\{\ell_{t',h'}(s',a')=k_{t',h'}(s',a')\Big\}\right)\\
\leqslant&\sum_{(s,a):\kappa_{\ell,j,h}(s,a)\leqslant T}\Big(\widehat{P}_{\kappa_{\ell,j,h}(s,a),h}-P_h\Big)\widehat{V}_{\kappa_{\ell,j,h}(s,a),h+1}(s,a)\\
&\qquad\times\prod_{(s',a')\in\mathcal{S}\times\mathcal{A}}\prod_{h'\in[H-1]}\mathbb{1}\Big\{\ell_{\kappa_{\ell,j,h}(s,a)-1,h'}(s',a')=k_{\kappa_{\ell,j,h}(s,a)-1,h'}(s',a')\Big\}\,,
\end{aligned}\tag{C.7}$$

where we used the short-hand notation $\widehat{V}_{t,h+1}\triangleq V_{h+1}^{\boldsymbol{\pi}_t,\boldsymbol{r}_t+\boldsymbol{b}_t,\widehat{\boldsymbol{P}}_t}$.

We are now ready to apply optional skipping—a concept recalled in Section 4.4.2, whose notation we use again here. On the events

$$\mathcal{C}'\triangleq\bigcap_{(s',a')\in\mathcal{S}\times\mathcal{A}}\bigcap_{h'\in[H-1]}\Big\{\ell_{\kappa_{\ell,j,h}(s,a)-1,h'}(s',a')=k_{\kappa_{\ell,j,h}(s,a)-1,h'}(s',a')\Big\}$$

considered, the empirical averages $\widehat{P}_{\kappa_{\ell,j,h}(s,a),h'}(\,\cdot\mid s',a')$ have the same distributions as the empirical frequency vectors associated with the i.i.d. random variables

$$\sigma_{h',s',a',j}\,,\qquad j\in\Big[2^{k_{\kappa_{\ell,j,h}(s,a)-1,h'}-1}\Big]\,,\qquad\text{with distribution}\quad P_{h'}(\,\cdot\mid s',a'),$$

and are independent from each other as h',s',a' vary. In particular, $\widehat{P}_{\kappa_{\ell,j,h}(s,a),h}(\,\cdot\mid s,a)$ is distributed as the empirical frequency vector of $2^{\ell-1}$ i.i.d. random variables $\sigma_{h,s,a,j}$, with $j\in[2^{\ell-1}]$. In addition, Bellman's equations (see the beginning of Section 4.3.1) show that on the events $\mathcal{C}'$ considered to apply optional skipping, $\widehat{V}_{\kappa_{\ell,j,h}(s,a),h+1}$ only depends on the $\pi_{\kappa_{\ell,j,h}(s,a),h'}$ with $h'\geqslant h+1$, on the $\widehat{P}_{\kappa_{\ell,j,h}(s,a),h'}(\,\cdot\mid s',a')$ with $h'\geqslant h+1$, and on state-action pairs relative to stages $h'\geqslant h+1$. Given the form of the adversarial learning strategy used, we conclude that on the events $\mathcal{C}'$ considered to apply optional skipping, all the $\widehat{V}_{\kappa_{\ell,j,h}(s,a),h+1}$, as s,a vary, only depend on state-action pairs of stages $h'\geqslant h+1$ and are therefore independent from all the $\widehat{P}_{\kappa_{\ell,j,h}(s',a'),h}$, as s',a' vary.

Put differently, optional skipping entails here that for all $\varepsilon>0$,

$$\begin{aligned}
&\mathbb{P}\left(\left\{\sum_{(s,a):\kappa_{\ell,j,h}(s,a)\leqslant T}\Big(\widehat{P}_{\kappa_{\ell,j,h}(s,a),h}-P_h\Big)\widehat{V}_{\kappa_{\ell,j,h}(s,a),h+1}(s,a)>\varepsilon\right\}\cap\mathcal{C}'\right)\\
&\leqslant\mathbb{P}\left\{\sum_{(s,a):\kappa_{\ell,j,h}(s,a)\leqslant T}\frac{1}{2^{\ell-1}}\sum_{j\in[2^{\ell-1}]}\Big(\widetilde{V}_{s,a,h+1}(\sigma_{h,s,a,j})-P_h\widetilde{V}_{s,a,h+1}(s,a)\Big)>\varepsilon\right\},
\end{aligned}\tag{C.8}$$

for some $\widetilde{V}_{s,a,h+1}$ independent from all the $\sigma_{h,s,a,j}$ as s,a,j vary (and h is fixed). We recall that the $\sigma_{h,s,a,j}$ are independent from each other as s,a,j vary (and h is fixed).

By the independencies noted above, and by boundedness of the values functions in $[0,(H-h)(H+1)] \subseteq [0,H^2]$, the Hoeffding–Azuma inequality guarantees that for all $\delta' \in (0,1)$,

$$\mathbb{P}\Bigg\{ \sum_{(s,a):\kappa_{\ell,j,h}(s,a)\leqslant T} \frac{1}{2^{\ell-1}} \sum_{j\in[2^{\ell-1}]} \Big(\widetilde{V}_{s,a,h+1}(\sigma_{h,s,a,j}) - P_h \widetilde{V}_{s,a,h+1}(s,a)\Big) \\ > \sqrt{\frac{1}{2\times 2^{\ell-1}} \sum_{(s,a):\kappa_{\ell,j,h}(s,a)\leqslant T} H^4 \ln\left(\frac{1}{\delta'}\right)} \Bigg\} = \delta' \,. \quad \text{(C.9)}$$

We summarize what we proved so far. Denoting by

$$\Delta_{T,\ell,j} \triangleq \sum_{t=1}^{T} \sum_{(s,a)} \Big(\widehat{P}_{t,h} - P_h\Big) V_{h+1}^{\boldsymbol{\pi}_t, \boldsymbol{r}_t + \boldsymbol{b}_t, \widehat{\boldsymbol{P}}_t}(s,a)\, \mathbb{1}\Big\{(s_{t,h},a_{t,h}) = (s,a)\Big\}\, \mathbb{1}\Big\{n_{t,h}(s,a) = 2^{\ell-1}+j\Big\}$$

the target quantity, and by

$$\mathcal{C} \triangleq \bigcap_{(s',a')\in\mathcal{S}\times\mathcal{A}} \left(\Big\{\tau_{\ell,j,h}(s',a') = \kappa_{\ell,j,h}(s',a')\Big\} \bigcap_{t'=\in[T]} \bigcap_{h'\in[H-1]} \Big\{\ell_{t',h'}(s',a') = k_{t',h'}(s',a')\Big\} \right)$$

the event associated with the values fixed, the bounds (C.7)–(C.8)–(C.9) show that for all $\delta' \in (0,1)$,

$$\mathbb{P}\left(\left\{ \Delta_{T,\ell,j} > \sqrt{2H^4 \frac{1}{2^{\ell-1}} \sum_{(s,a)} \mathbb{1}\Big\{n_{T,h}(s,a) \geqslant 2^{\ell-1}+j\Big\} \ln\left(\frac{1}{\delta'}\right)} \right\} \cap \mathcal{C} \right) \\ = \mathbb{P}\left(\left\{ \Delta_{T,\ell,j} > \sqrt{\frac{1}{2^{\ell}} \sum_{(s,a):\kappa_{\ell,j,h}(s,a)\leqslant T} H^2 \ln\left(\frac{1}{\delta'}\right)} \right\} \cap \mathcal{C} \right) \leqslant \delta' \,. \quad \text{(C.10)}$$

Part 2: Union bound and counting the sequences. The proof is concluded by counting how many different sets $\mathcal{C}$ may be obtained. We do so in a rough way, that will be sufficient for our purposes. First, we need to count the profile values (4.8). There are $T(H-1)$ functions $\ell_{t,h} \colon \mathcal{S}\times\mathcal{A} \to \Big[\lceil \log_2(T)\rceil\Big]^*$, satisfying some monotonicity constraints, as well as some other constraints which we ignore. The monotonicity constraints imply that for each (s,a) and $h \in [H-1]$, it is sufficient to determine the at most $\lceil \log_2(T)\rceil$ time steps t among $[T]$ when $\ell_{t,h}$ increases by 1. Thus, there are at most

$$(T+1)^{\lceil \log_2(T)\rceil}$$

possible sequences of values for the $\ell_{t,h}(s,a)$ as t varies and h,s,a are fixed. All in all, the profile part in the number of different sets $\mathcal{C}$ is smaller than

$$\Big((T+1)^{\lceil \log_2(T)\rceil}\Big)^{SA(H-1)} .$$

For stopping times, we need to determine, for each (s,a), a single value, in a set included in $[T]\cup\{+\infty\}$. We neglect other constraints and see that there are therefore at most $(T+1)^{SA}$ such choices.

As a conclusion, there are at most

$$M = (T+1)^{SAH\lceil \log_2(T)\rceil}$$

different possible values for the sets $\mathcal{C}$. The proof of Lemma C.2 is concluded by a union bound over the events (C.10), with $\delta' = \delta/(4HTM)$. □

Appendix D

Technical Lemmas

Contents

D.1 Entropy Properties

Lemma D.1. *For any reward-free MDP* $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \{p_h\}_{h\in[H]}, s_1)$ *the following holds*

$$\begin{aligned}\max_{d\in\mathcal{K}_p} \mathcal{H}_{visit}(d) &= \max_{d\in\mathcal{K}_p} \min_{\bar{d}\in\mathcal{K}} \sum_{(h,s,a)} d_h(s,a) \log \frac{1}{\bar{d}_h(s,a)} \\ &= \min_{\bar{d}\in\mathcal{K}} \max_{d\in\mathcal{K}_p} \sum_{(h,s,a)} d_h(s,a) \log \frac{1}{\bar{d}_h(s,a)}\end{aligned}$$

Proof. The first equality is due to

$$\begin{aligned}\mathcal{H}_{\text{visit}}(d) &= \sum_{h=1}^{H} \mathcal{H}(d_h) = \sum_{h=1}^{H} \mathcal{H}(d_h) + \mathrm{KL}(d_h, d_h) \\ &= \min_{\bar{d}\in\mathcal{K}} \sum_{h=1}^{H} (\mathcal{H}(d_h) + \mathrm{KL}(d_h, \bar{d}_h)) = \min_{\bar{d}\in\mathcal{K}} \sum_{(h,s,a)} d_h(s,a) \log \frac{1}{\bar{d}_h(s,a)}\end{aligned}$$

and the second equality uses Sion's theorem (Sion 1958) with the fact that $\mathcal{K}, \mathcal{K}_p$ are compact convex sets and $(d, \bar{d}) \mapsto -\sum_{h,s,a} d_h(s,a) \log \bar{d}_h(s,a)$ is concave-convex. □

Lemma D.2. *The trajectory entropy $\mathcal{H}_{traj}(\pi)$ has the following representation*

$$\mathcal{H}_{traj}(\pi) = \sum_{m\in\mathcal{T}} q^\pi(m)\log\frac{1}{q^\pi(m)} = \mathbb{E}_\pi\left[\sum_{h=1}^H \mathcal{H}(p_h(s_h,a_h)) + \mathcal{H}(\pi_h(s_h))|s_1\right],$$

with a convention $\mathcal{H}(p_H(s,a)) = 0$.

Proof. By a definition of $q^\pi(m)$:

$$q^\pi(m) = \pi_1(a_1|s_1)\prod_{h=2}^H p_{h-1}(s_h|s_{h-1},a_{h-1})\pi_h(a_h|s_h),$$

thus

$$\sum_{m\in\mathcal{T}} q^\pi(m)\log\frac{1}{q^\pi(m)} = -\sum_{m\in\mathcal{T}} q^\pi(m)\Bigg(\log(\pi_1(a_1|s_1)) + \sum_{h=2}^H \log(p_{h-1}(s_h|s_{h-1},a_{h-1})) + \log(\pi_h(a_h|s_h))\Bigg).$$

By rearranging the terms we have

$$\mathcal{H}_{\text{traj}}(\pi) = \sum_{h=1}^H\sum_{m\in\mathcal{T}} q_h^\pi(m)\left(\log\frac{1}{p_h(s_{h+1}|s_h,a_h)} + \log\frac{1}{\pi_h(a_h|s_h)}\right),$$

where under convention p_H is a deterministic transition to s_1. Marginalizing $q_h^\pi(m)$ over s_h, a_h, s_{h+1} we get

$$\begin{aligned}\mathcal{H}_{\text{traj}}(\pi) &= \sum_{h=1}^H\sum_{s,a,s'} d^\pi(s,a)p_h(s'|s,a)\left(\log\frac{1}{p_h(s'|s,a)} + \log\frac{1}{\pi_h(a|s)}\right)\\ &= \sum_{s,a} d_h^\pi(s,a)\sum_{h=1}^H \mathcal{H}(p_h(s,a)) + \sum_s d_h^\pi(s)\mathcal{H}(\pi_h(s)).\end{aligned}$$

□

Lemma D.3. *For any reward-free MDP $\mathcal{M} = (\mathcal{S},\mathcal{A},\{p_h\}_{h\in[H]},s_1)$ and any policy π it holds*

$$\mathcal{H}_{traj}(q^\pi) \leqslant \mathcal{H}_{visit}(d^\pi) \leqslant H\mathcal{H}_{traj}(q^\pi).$$

Proof. The first inequality is a result of the non-negativity of mutual information, specifically the difference $\mathcal{H}_{\text{visit}}(d^\pi) - \mathcal{H}_{\text{traj}}(q^\pi)$ is equal to the Kullback-Leibler divergence between q^π and the product distribution $\otimes_{h=1}^H d_h^\pi$,

$$\begin{aligned}\mathcal{H}_{\text{visit}}(d^\pi) &= \sum_{s,a,h} d_h^\pi(s,a)\log\frac{1}{d_h^\pi(s,a)}\\ &= \sum_{s,a,h}\sum_{m=(s_1,a_1,\ldots,s_H,a_H)\in\mathcal{T}} \mathbb{1}\{(s,a)=(s_h,a_h)\}q^\pi(m)\log\frac{1}{d_h^\pi(s,a)}\\ &= \sum_{h=1}^H\sum_{m=(s_1,a_1,\ldots,s_H,a_H)\in\mathcal{T}} q^\pi(m)\log\frac{1}{d_h^\pi(s_h,a_h)}\\ &= \sum_{m=(s_1,a_1,\ldots,s_H,a_H)\in\mathcal{T}} q^\pi(m)\log\frac{1}{\prod_{h=1}^H d_h^\pi(s_h,a_h)}\\ &= \mathrm{KL}(q^\pi,\otimes_{h=1}^H d_h^\pi) + \mathcal{H}_{\text{traj}}(q^\pi)\\ &\geqslant \mathcal{H}_{\text{traj}}(q^\pi).\end{aligned}$$

The second inequality is simply a consequence of the monotonicity of logarithm

$$\begin{aligned}
\mathcal{H}_{\text{traj}}(q^\pi) &= \sum_{m\in\mathcal{T}} q^\pi(m)\log\frac{1}{q^\pi(m)} \\
&= \sum_{s,a,h}\frac{1}{H}\sum_{m=(s_1,a_1,\dots,s_H,a_H)\in\mathcal{T}} \mathbb{1}\{(s,a)=(s_h,a_h)\}q^\pi(m)\log\left(\frac{1}{q^\pi(m)}\right) \\
&\geqslant \sum_{s,a,h}\frac{1}{H}\sum_{m=(s_1,a_1,\dots,s_H,a_H)\in\mathcal{T}} \mathbb{1}\{(s,a)=(s_h,a_h)\}q^\pi(m) \\
&\qquad\qquad \times\log\left(\frac{1}{\sum_{m'=(s'_1,a'_1,\dots,s'_H,a'_H)\in\mathcal{T}}\mathbb{1}\{(s,a)=(s'_h,a'_h)\}q^\pi(m')}\right) \\
&= \frac{1}{H}\mathcal{H}_{\text{visit}}(d^\pi),
\end{aligned}$$

where we used that $\log(1/x) = -\log(x)$ is monotonically decreasing, and

$$\sum_{m'=(s'_1,a'_1,\dots,s'_H,a'_H)\in\mathcal{T}} \mathbb{1}\{(s,a)=(s'_h,a'_h)\}q^\pi(m') \geqslant q^\pi(m)$$

for any $m=(s_1,a_1,\dots,s_H,a_H)$ such that $(s_h,a_h)=(s,a)$. □

Lemma D.4. *In a deterministic MDP, i.e. the transition probability distributions are deterministic, the maximum trajectory entropy policy is the uniform policy over actions*

$$\pi_h^{\star,TE}(s,a) = 1/A\,.$$

Proof. The proof is straightforward by using the Bellman equations for MTEE. By induction over the step h we prove that $\pi_h^{\star,\text{TE}}(s,a)=1/A$ and $V_h^\star(s)=(H+1-h)\log(A)$. Assume the induction hypothesis at step $h+1$. Then we get $Q_h^\star(s,a)=0+p_hV_{h+1}^\star(s,a)=(H-h)\log(A)$ which yields $\pi_h^{\star,\text{TE}}(s,a)=1/A$ and $V_h^\star(s)=\log\left(\sum_{a\in\mathcal{A}}A^{H-h}\right)=(H+1-h)\log(A)$. The base case $h=H$ is immediate. □

D.2 Counts to pseudo-counts

Here we state Lemma 8 and Lemma 9 by Ménard et al. (2021).

Lemma D.5. *On event $\mathcal{E}^{\text{cnt}}$, for any $\beta(\delta,\cdot)$ such that $x\mapsto\beta(\delta,x)/x$ is non-increasing for $x\geqslant 1$, $x\mapsto\beta(\delta,x)$ is non-decreasing $\forall h\in[H],(s,a)\in\mathcal{S}\times\mathcal{A}$,*

$$\forall t\in\mathbb{N}^\star,\ \frac{\beta(\delta,n_h^t(s,a))}{n_h^t(s,a)}\wedge 1 \leqslant 4\frac{\beta(\delta,\bar{n}_h^t(s,a))}{\bar{n}_h^t(s,a)\vee 1}\,.$$

Lemma D.6. *For $T\in\mathbb{N}^\star$ and $(u_t)_{t\in\mathbb{N}^\star}$, for a sequence where $u_t\in[0,1]$ and $U_t\triangleq\sum_{l=1}^t u_\ell$, we get*

$$\sum_{t=0}^{T}\frac{u_{t+1}}{U_t\vee 1}\leqslant 4\log(U_{T+1}+1)\,.$$

D.3 Counts to pseudo-counts in linear MDPs

Let $\{X_t\}_{t=1}^\infty$ be a sequence of random vectors of dimension d adapted to a filtration $\{\mathcal{F}_t\}_{t=1}^\infty$ such that $\|X_t\|_2\leqslant 1$ a.s.. Define a sequence of positive semi-definite matrices $A_t=\mathbb{E}[X_tX_t^\intercal|\mathcal{F}_{t-1}]$. Notice that $\|A_t\|_2=\sigma_{\max}(A_t)\leqslant 1$. Also define $\Lambda_t=\lambda I_d+\sum_{j=1}^t X_jX_j^\intercal$ and $\overline{\Lambda}_t=\lambda I_d+\sum_{j=1}^t A_j$.

Lemma D.7. *Let $A \succcurlyeq 0$ be a positive semi-definite matrix such that $\|A\|_2 \leqslant 1$. Then*

$$\log\det(I+A) \leqslant \mathrm{Tr}(A) \leqslant 2\log\det(I+A)\,.$$

Proof. Follows from eigendecomposition for A and numeric inequality $\log(1+x) \leqslant x \leqslant 2\log(1+x)$ for all $x \in [0,1]$. □

The next result generalized Lemma D.2 by Jin et al. (2020b); see also Abbasi-Yadkori et al. (2011). Also, it could be treated as a generalization of Lemma D.6 for linear MDPs.

Lemma D.8. *Let $\|A_t\|_2 \leqslant 1$ for all $t \geqslant 1$. Then for any $T \geqslant 1$*

$$\log\frac{\det(\overline{\Lambda}_T)}{\det(\overline{\Lambda}_0)} \leqslant \sum_{t=1}^{T} \mathbb{E}\Big[X_t^\intercal[\overline{\Lambda}_{t-1}]^{-1}X_t|\mathcal{F}_{t-1}\Big] \leqslant 2\log\frac{\det(\overline{\Lambda}_T)}{\det(\overline{\Lambda}_0)}\,.$$

Proof. First, we notice that $\overline{\Lambda}_{t-1}$ is $\mathcal{F}_{t-1}$-measurable, thus

$$\begin{aligned}\mathbb{E}\Big[X_t^\intercal[\overline{\Lambda}_{t-1}]^{-1}X_t|\mathcal{F}_{t-1}\Big] &= \mathbb{E}\Big[\mathrm{Tr}\Big([\overline{\Lambda}_{t-1}]^{-1}X_tX_t^\intercal\Big)|\mathcal{F}_{t-1}\Big] = \mathrm{Tr}\Big([\overline{\Lambda}_{t-1}]^{-1}\mathbb{E}[X_tX_t^\intercal|\mathcal{F}_{t-1}]\Big)\\ &= \mathrm{Tr}\Big([\overline{\Lambda}_{t-1}]^{-1}A_t\Big) = \mathrm{Tr}\Big([\overline{\Lambda}_{t-1}]^{-1/2}A_t[\overline{\Lambda}_{t-1}]^{-1/2}\Big)\,.\end{aligned}$$

Notice that $\Sigma_t = [\overline{\Lambda}_{t-1}]^{-1/2}A_t[\overline{\Lambda}_{t-1}]^{-1/2}$ is positive semi-definite matrix. Then by Lemma D.7

$$\log\det(I_d+\Sigma_t) \leqslant \mathbb{E}\Big[X_t^\intercal[\overline{\Lambda}_{t-1}]^{-1}X_t|\mathcal{F}_{t-1}\Big] \leqslant 2\log\det(I_d+\Sigma_t)\,.$$

At the same time, we have

$$\det(\overline{\Lambda}_t) = \det(\overline{\Lambda}_{t-1}+A_t) = \det(\overline{\Lambda}_{t-1})\cdot\det\Big(I_d + [\overline{\Lambda}_{t-1}]^{-1/2}A_t[\overline{\Lambda}_{t-1}]^{-1/2}\Big)\,.$$

Thus by telescoping property

$$\log\frac{\det(\overline{\Lambda}_T)}{\det(\overline{\Lambda}_0)} \leqslant \sum_{t=1}^{T} \mathbb{E}\Big[X_t^\intercal[\overline{\Lambda}_{t-1}]^{-1}X_t|\mathcal{F}_{t-1}\Big] \leqslant 2\log\frac{\det(\overline{\Lambda}_T)}{\det(\overline{\Lambda}_0)}\,.$$

□

D.4 On the Bernstein inequality

We restate here a Bernstein-type inequality by Talebi and Maillard (2018). Remark that the second part of Corollary 11 by Talebi and Maillard (2018) is not correct, i.e., there exist two measures p, q such that

$$qf - pf > \sqrt{2\mathrm{Var}_q(f)\,\mathrm{KL}(p\|q)}\,,$$

whereas the first inequality still holds. We provide a direct proof in Lemma D.9 for completeness.

Lemma D.9. *Let $p, q \in \Delta_S$, where Δ_S denotes the probability simplex of dimension S, and assume $p \ll q$. For all functions $f\colon \mathcal{S} \mapsto [0,b]$ defined on $\mathcal{S}$,*

$$\begin{aligned}pf - qf &\leqslant \sqrt{2\mathrm{Var}_q(f)\,\mathrm{KL}(p\|q)} + \frac{1}{3}b\,\mathrm{KL}(p\|q)\,,\\ qf - pf &\leqslant \sqrt{2\mathrm{Var}_q(f)\,\mathrm{KL}(p\|q)} + \frac{1}{3}b\,\mathrm{KL}(p\|q)\,,\end{aligned}$$

where use the expectation operator defined as $pf \triangleq \mathbb{E}_{s\sim p}f(s)$ and the variance operator defined as $\mathrm{Var}_p(f) \triangleq \mathbb{E}_{s\sim p}\Big(f(s) - \mathbb{E}_{s'\sim p}f(s')\Big)^2 = p(f-pf)^2$.

Proof. Notice that in the case $\mathrm{KL}(p\|q) = 0$ the inequality holds trivially since both sides of the inequality are equal to zero, thus we assume $\mathrm{KL}(p\|q) > 0$. Additionally, let us consider the case $\mathrm{Var}_q[f] = 0$, it implies that $q = \delta_{s^\star}$ for some $s^\star \in \mathcal{S}$ and thus $\mathrm{KL}(p\|q) < +\infty$ if and only if $p = \delta_{s^\star}$, thus the inequality also holds trivially.

By the Donsker and Varadahn's variational formula for KL-divergence (Donsker and Varadhan 1983) we have for any function $g\colon \mathcal{S} \to \mathbb{R}$

$$\log \mathbb{E}_{X\sim q}[\exp(g(X))] = \sup_p \{pg - \mathrm{KL}(p\|q)\}\,.$$

Taking $g(s) \triangleq \lambda \cdot f(s)$ for a parameter $\lambda > 0$, we have

$$\log \mathbb{E}_{X\sim q}[\exp(\lambda[f(X) - qf])] + \lambda qf \geqslant \lambda pf - \mathrm{KL}(p\|q)\,,$$

and, denoting $\psi(\lambda) \triangleq \log \mathbb{E}_{X\sim q}[\exp(\lambda[f(X) - qf])]$ after some rearranging, we have

$$pf - qf \leqslant \inf_{\lambda>0}\left\{\frac{\psi(\lambda) + \mathrm{KL}(p\|q)}{\lambda}\right\}.$$

Next, we notice that a random variable $Y = f(X) - qf$ is centered and bounded $|Y| \leqslant b$, thus the moment generating function is bounded by Bennett's inequality (Boucheron et al. 2013, Theorem 2.9) for any $\lambda \in (0, 3/b)$

$$\psi(\lambda) \leqslant \frac{\mathbb{E}[Y^2]}{b^2}\left(\mathrm{e}^{b\lambda} - b\lambda - 1\right) \leqslant \mathrm{Var}_q[f] \cdot \frac{\lambda^2}{2(1 - b\lambda/3)}\,,$$

where the second inequality holds by a simple bound for any $x < 3$

$$\mathrm{e}^x = 1 + x + x^2 \cdot \sum_{k=0}^{\infty} \frac{x^k}{(k+2)!} \leqslant 1 + x + \frac{x^2}{2} \cdot \sum_{k=0}^{\infty} \frac{x^k}{3^k} = 1 + x + \frac{x^2}{2(1 - x/3)}\,.$$

Notice that for $\lambda^\star = 1/(b/3 + \sqrt{v/y})$, where $v = \mathrm{Var}_q[f]/2$ and $y = \mathrm{KL}(p\|q)$, we have $\lambda^\star/(1 - b\lambda^\star/3) = \sqrt{y/v}$ and $1/\lambda^\star = b/3 + \sqrt{v/y}$, thus

$$\begin{aligned} pf - qf &\leqslant \inf_{\lambda\in(0,3/b)}\left(\frac{\mathrm{Var}_q[f]}{2} \cdot \frac{\lambda}{1 - b\lambda/3} + \frac{\mathrm{KL}(p\|q)}{\lambda}\right) \\ &\leqslant \sqrt{2\mathrm{Var}_q[f]\,\mathrm{KL}(p\|q)} + \frac{b}{3}\,\mathrm{KL}(p\|q)\,. \end{aligned}$$

The second inequality follows directly from the first one by applying it to a function $g \triangleq b - f$. □

Lemma D.10. *Let $p, q \in \Delta_S$ and a function $f : \mathcal{S} \mapsto [0, b]$, then*

$$\begin{aligned} \mathrm{Var}_q(f) &\leqslant 2\mathrm{Var}_p(f) + 4b^2\,\mathrm{KL}(p\|q)\,, \\ \mathrm{Var}_p(f) &\leqslant 2\mathrm{Var}_q(f) + 4b^2\,\mathrm{KL}(p\|q). \end{aligned}$$

Proof. Let $p \otimes p$ be the distribution of the pair of random variables (X, Y) where X, Y are i.i.d. according to the distribution p. Similarly, let $q \otimes q$ be the distribution of the pair of random variables (X, Y) where X, Y are i.i.d. according to distribution q. Since Kullback–Leibler divergence is additive for independent distributions, we know that

$$\mathrm{KL}(p \otimes p, q \otimes q) = 2\,\mathrm{KL}(p, q).$$

Using Lemma D.9 for the function $g(x,y) = (f(x) - f(y))^2$ defined on $\mathcal{S}^2$, such that $0 \leqslant g \leqslant b^2$, we get

$$\begin{aligned}|(p \otimes p)g - (q \otimes q)g| &\leqslant \sqrt{4\mathrm{Var}_{q\otimes q}(g)\,\mathrm{KL}(p,q)} + \frac{2}{3}b^2\,\mathrm{KL}(p,q) \\ &\leqslant \sqrt{4b^2\,\mathrm{KL}(p,q)(q\otimes q)g} + \frac{2}{3}b^2\,\mathrm{KL}(p,q) \\ &\leqslant \frac{1}{2}(q\otimes q)g + 3b^2\,\mathrm{KL}(p,q)\,,\end{aligned}$$

where in the last line we used $2\sqrt{xy} \leqslant x + y$ for $x, y \geqslant 0$. In particular, we obtain

$$\begin{aligned}(p\otimes p)g &\leqslant \frac{3}{2}(q\otimes q)g + 3b^2\,\mathrm{KL}(p,q)\,, \\ (q\otimes q)g &\leqslant 2(p\otimes p)g + 6b^2\,\mathrm{KL}(p,q)\,.\end{aligned}$$

To conclude, it remains to note that

$$(p\otimes p)g = 2\mathrm{Var}_p(f) \text{ and } (q\otimes q)g = 2\mathrm{Var}_q(f)\,.$$

□

Lemma D.11. *For $p, q \in \Delta_S$, for $f, g : \mathcal{S} \mapsto [0, b]$ two functions defined on $\mathcal{S}$, we have that*

$$\begin{aligned}\mathrm{Var}_p(f) &\leqslant 2\mathrm{Var}_p(g) + 2bp|f-g| \quad \textit{and} \\ \mathrm{Var}_q(f) &\leqslant \mathrm{Var}_p(f) + 3b^2\|p-q\|_1\,,\end{aligned}$$

where we denote the absolute operator by $|f|(s) = |f(s)|$ for all $s \in \mathcal{S}$.

Proof. For the first inequality, we have

$$\begin{aligned}\mathrm{Var}_p(f) &= p(f-pf)^2 = p(f-g+g-pg+pg-pf)^2 \\ &\leqslant 2p(f-g-p(f-g))^2 + 2p(g-pg)^2 = 2\mathrm{Var}_p(g) + 2\mathrm{Var}_p(f-g)\,,\end{aligned}$$

and the bound $\mathrm{Var}_p(f-g) \leqslant p(f-g)^2 \leqslant bp|f-g|$.

The second inequality follows from the following representation:

$$\mathrm{Var}_q(f) - \mathrm{Var}_p(f) = (q-p)f^2 + (pf-qf)(pf+qf) \leqslant 3b^2\|p-q\|_1\,.$$

□

D.5 Change of policy

Let π and π' be two Markovian policies and let $\pi^{(h')}$ for $h' \in [H]$ be a family of policies defined as follows

$$\pi_h^{(h')}(s) = \begin{cases}\pi_h(s) & h \neq h'\,, \\ \pi'_h(s) & h = h'\,.\end{cases}$$

Lemma D.12. *For any measurable function $f\colon \mathcal{S} \to \mathbb{R}$ and any $h \leqslant h'$ it holds*

$$\mathbb{E}_\pi[f(s_h)|s_1] = \mathbb{E}_{\pi^{(h')}}[f(s_h)|s_1]\,.$$

Moreover, for any measurable $g\colon \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ and any $h \geqslant h'$

$$\mathbb{E}_\pi[g(s_h,a_h)|(s_{h'},a_{h'}) = (s,a)] = \mathbb{E}_{\pi^{(h')}}[g(s_h,a_h)|(s_{h'},a_{h'}) = (s,a)]\,.$$

Proof. At first, we recall that by definition of a kernel p_h we have for any measurable f:

$$\mathbb{E}_\pi[f(s_{h+1})|s_h, a_h] = p_h f(s_h, a_h), \quad \mathbb{E}_\pi[g(s_h, a_h)|s_h] = \pi_h g(s_h)\,. \tag{D.1}$$

We show the first statement by induction over $h \leqslant h'$. For $h = 1$ the statement is trivial. Next, we assume that it holds for h and we have to show for $h + 1 \leqslant h'$. By the tower property of conditional expectation and (D.1)

$$\begin{aligned}\mathbb{E}_\pi[f(s_{h+1})|s_1] &= \mathbb{E}_\pi[\mathbb{E}_\pi[f(s_{h+1})|s_h, a_h]|s_1] = \mathbb{E}_\pi[p_h f(s_h, a_h)|s_1] \\ &= \mathbb{E}_\pi[\mathbb{E}_\pi[p_h f(s_h, a_h)|s_h]|s_1] = \mathbb{E}_\pi[\pi_h[p_h f](s_h)|s_1]\,.\end{aligned}$$

We notice that since $h < h'$ then $\pi_h = \pi_h^{(h')}$. Then we can apply the induction hypothesis to a function $\pi_h^{(h')}[p_h f](s_h)$

$$\mathbb{E}_\pi[f(s_{h+1})|s_1] = \mathbb{E}_\pi[\pi_h[p_h f](s_h)|s_1] = \mathbb{E}_\pi[\pi_h^{(h')}[p_h f](s_h)|s_1] = \mathbb{E}_{\pi^{(h')}}[f(s_{h+1})|s_1]\,.$$

We use induction over $h \geqslant h'$ to show the second statement. For $h = h'$ the statement is trivial. Next, we assume that it holds for $h \geqslant h'$, and we have to show it for $h + 1$. Again, using the tower property

$$\mathbb{E}_\pi[g(s_{h+1}, a_{h+1})|s_{h'}, a_{h'}] = \mathbb{E}_\pi[\mathbb{E}_\pi[g(s_{h+1}, a_{h+1})|s_{h+1}]|s_{h'}, a_{h'}] = \mathbb{E}_\pi[\pi_{h+1} g(s_{h+1})|s_{h'}, a_{h'}]\,.$$

We notice that $\pi_{h+1} = \pi_{h+1}^{(h')}$ since $h \geqslant h'$. Thus, again applying the tower property, (D.1), and induction hypothesis

$$\begin{aligned}\mathbb{E}_\pi[g(s_{h+1}, a_{h+1})|s_{h'}, a_{h'}] &= \mathbb{E}_\pi[\mathbb{E}_\pi[\pi_{h+1}^{(h')} g(s_{h+1})|s_h, a_h]|s_{h'}, a_{h'}] \\ &= \mathbb{E}_\pi[p_h[\pi_{h+1}^{(h')} g](s_h, a_h)|s_{h'}, a_{h'}] \\ &= \mathbb{E}_{\pi^{(h')}}[p_h[\pi_{h+1}^{(h')} g](s_h, a_h)|s_{h'}, a_{h'}] = \mathbb{E}_{\pi^{(h')}}[g(s_{h+1}, a_{h+1})|s_{h'}, a_{h'}]\,.\end{aligned}$$

□

D.6 Deviation Inequalities

D.6.1 Deviation inequality for categorical distributions

Next, we state the deviation inequality for categorical distributions by Jonsson et al. (2020, Proposition 1). Let $(X_t)_{t\in\mathbb{N}^\star}$ be i.i.d. samples from a distribution supported on $\{1, \dots, m\}$, of probabilities given by $p \in \Delta_{m-1}$, where Δ_{m-1} is the probability simplex of dimension $m - 1$. We denote by $\widehat{p}_n$ the empirical vector of probabilities, i.e., for all $k \in \{1, \dots, m\}$,

$$\widehat{p}_{n,k} \triangleq \frac{1}{n}\sum_{\ell=1}^{n} \mathbb{1}\{X_\ell = k\}\,.$$

Note that an element $p \in \Delta_{m-1}$ can be seen as an element of $\mathbb{R}^{m-1}$ since $p_m = 1 - \sum_{k=1}^{m-1} p_k$. This will be clear from the context.

Theorem D.13. *For all $p \in \Delta_{m-1}$ and for all $\delta \in [0, 1]$,*

$$\mathbb{P}(\exists n \in \mathbb{N}^\star,\, n\,\mathrm{KL}(\widehat{p}_n, p) > \log(1/\delta) + (m-1)\log(e(1 + n/(m-1)))) \leqslant \delta\,.$$

D.6.2 Deviation inequality for sequence of Bernoulli random variables

Below, we state the deviation inequality for Bernoulli distributions by Dann et al. (2017, Lemma F.4). Let $\mathcal{F}_t$ for $t \in \mathbb{N}$ be a filtration and $(X_t)_{t\in\mathbb{N}^\star}$ be a sequence of Bernoulli random variables with $\mathbb{P}(X_t = 1|\mathcal{F}_{t-1}) = P_t$ with P_t being $\mathcal{F}_{t-1}$-measurable and X_t being $\mathcal{F}_t$-measurable.

Theorem D.14. *For all $\delta > 0$,*

$$\mathbb{P}\left(\exists n : \sum_{t=1}^{n} X_t < \sum_{t=1}^{n} P_t/2 - \log\frac{1}{\delta}\right) \leqslant \delta\,.$$

D.6.3 Deviation inequality for bounded distributions

Below, we state the self-normalized Bernstein-type inequality by Domingues et al. (2021b). Let $(Y_t)_{t\in\mathbb{N}^\star}$, $(w_t)_{t\in\mathbb{N}^\star}$ be two sequences of random variables adapted to a filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$. We assume that the weights are in the unit interval $w_t \in [0,1]$ and predictable, i.e. $\mathcal{F}_{t-1}$ measurable. We also assume that the random variables Y_t are bounded $|Y_t| \leqslant b$ and centered $\mathbb{E}[Y_t|\mathcal{F}_{t-1}] = 0$. Consider the following quantities

$$S_t \triangleq \sum_{s=1}^{t} w_s Y_s, \quad V_t \triangleq \sum_{s=1}^{t} w_s^2 \cdot \mathbb{E}\Big[Y_s^2|\mathcal{F}_{s-1}\Big], \quad \text{and} \quad W_t \triangleq \sum_{s=1}^{t} w_s$$

and let $h(x) \triangleq (x+1)\log(x+1) - x$ be the Cramér transform of a Poisson distribution of parameter 1.

Theorem D.15 (Bernstein-type concentration inequality)**.** *For all $\delta > 0$,*

$$\mathbb{P}\left(\exists t \geqslant 1, (V_t/b^2 + 1)h\left(\frac{b|S_t|}{V_t + b^2}\right) \geqslant \log(1/\delta) + \log(4e(2t+1))\right) \leqslant \delta\,.$$

The previous inequality can be weakened to obtain a more explicit bound: if $b \geqslant 1$ with probability at least $1-\delta$, for all $t \geqslant 1$,

$$|S_t| \leqslant \sqrt{2V_t \log(4e(2t+1)/\delta)} + 3b\log(4e(2t+1)/\delta)\,.$$

D.6.4 Deviation inequality for vector-valued self-normalized processes

Next, we state Lemma D.4 by Jin et al. (2020b). For any symmetric positive definite matrix A we define $\|x\|_A = \sqrt{x^\intercal A x}$.

Lemma D.16 (Jin et al. 2020b)**.** *Let $\{s_\tau\}_{\tau=1}^{\infty}$ be a stochastic process on the state space $\mathcal{S}$ adapted to a filtration $\{\mathcal{F}_\tau\}_{\tau=0}^{\infty}$. Let $\{X_\tau\}_{\tau=0}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process where ψ_τ is $\mathcal{F}$-predictable (X_τ is $\mathcal{F}_{\tau-1}$ measurable) and $\|X_\tau\| \leqslant 1$. Let $\Lambda_t = \alpha I_d + \sum_{\tau=1}^{t} X_t X_t^\intercal$ and let $\mathcal{V}$ be a family of function over the state-space $\mathcal{S}$ such that $\forall V \in \mathcal{V}, \forall s \in \mathcal{S} : 0 \leqslant V(s) \leqslant H$. Then for any $\delta \in (0,1)$ with probability at least $1-\delta$*

$$\forall t \in \mathbb{N}, \forall V \in \mathcal{V} : \left\|\sum_{\tau=1}^{t} X_\tau\{V(s_\tau) - \mathbb{E}_{\tau-1}[V(s_\tau)]\}\right\|_{\Lambda_t^{-1}}^2 \leqslant 4H^2\left[\frac{d}{2}\log\left(\frac{t+\alpha}{\alpha}\cdot\frac{|\mathcal{N}_\varepsilon|}{\delta}\right)\right] + \frac{8t^2\varepsilon^2}{\alpha}\,,$$

where $\mathcal{N}_\varepsilon$ is a minimal ε-cover of $\mathcal{V}$ with respect to the distance $\rho(V,V') = \sup_{s\in\mathcal{S}}|V(s) - V'(s)|$.

Also, we state the result on the covering dimension of the class of bonus function, see Lemma D.6 by Jin et al. (2020b) for a similar result.

Lemma D.17. *Let $\mathcal{V}$ be a class of functions over $\mathcal{S}$ such that*

$$\forall s \in \mathcal{S} : V(s) = \operatorname{clip}\left(\max_{\pi \in \Delta(\mathcal{A})} \left\{\pi\left[w^\top \psi(s,\cdot) + \beta\sqrt{\psi(s,\cdot)^\top \Lambda^{-1} \psi(s,\cdot)}\right] - \lambda \Phi_{h,s}(\pi)\right\}, 0, H\right)$$

is parameterized by a tuple (w, β, Λ) such that $\|w\| \leqslant L, \beta \in [0, B]$, $\lambda_{\min}(\Lambda) \geqslant \alpha$. Assume $\|\psi(s,a)\| \leqslant 1$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. Then the covering number $|\mathcal{N}_\varepsilon|$ of the function space $\mathcal{V}$ with respect to the distance $\rho(V, V') = \sup_{s \in \mathcal{S}} |V(s) - V'(s)|$ satisfies

$$\log \mathcal{N}(\varepsilon, \mathcal{V}, \rho) \leqslant d \log\left(1 + \frac{4L}{\varepsilon}\right) + d^2 \log\left(1 + \frac{8 d^{1/2} B^2}{\alpha \varepsilon^2}\right).$$

Proof. Following the approach of Jin et al. (2020b), we reparametrize the following set by setting $A = \beta^2 \Lambda^{-1}$ and obtain

$$V(s) = \operatorname{clip}\left(\max_{\pi \in \Delta(\mathcal{A})} \left\{\pi\left[w^\top \psi(s,\cdot) + \sqrt{\psi(s,\cdot)^\top A \psi(s,\cdot)}\right] - \lambda \Phi_{h,s}(\pi)\right\}, 0, H\right),$$

for $\|w\|_2 \leqslant L, \|A\|_2 \leqslant B^2 \alpha^{-1}$. Next, let $V_1, V_2 \in \mathcal{V}$ be two functions that corresponds to parameters (w_1, A_1) and (w_2, A_2). Then, using non-expanding property of $\operatorname{clip}(\cdot, 0, H)$ and $\max_{\pi \in \Delta(\mathcal{A})}\{\cdot\}$ we have

$$\begin{aligned}
\rho(V_1, V_2) &\leqslant \sup_{s \in \mathcal{S}} \max_{\pi \in \Delta(\mathcal{A})} \pi\left[(w_1^\top \psi(s,\cdot) + \sqrt{\psi(s,\cdot)^\top A_1 \psi(s,\cdot)}) - (w_2^\top \psi(s,\cdot) + \sqrt{\psi(s,\cdot)^\top A_2 \psi(s,\cdot)})\right] \\
&\leqslant \sup_{s,a \in \mathcal{S} \times \mathcal{A}} \left[[w_1 - w_2]^\top \psi(s,a) + \sqrt{\psi(s,a)^\top A_1 \psi(s,a)}) - \sqrt{\psi(s,a)^\top A_2 \psi(s,a)}\right] \\
&\leqslant \sup_{\psi : \|\psi\| \leqslant 1} |\langle w_1 - w_2, \psi \rangle| + \sup_{\psi : \|\psi\| \leqslant 1} \sqrt{|\psi^\top (A_1 - A_2) \psi|} \\
&\leqslant \|w_1 - w_2\|_2 + \sqrt{\|A_1 - A_2\|_F}\,.
\end{aligned}$$

The rest of the proof follows Lemma D.6 by Jin et al. 2020b and uses the result on covering numbers of Euclidean balls in $\mathbb{R}^d$. □

D.6.5 Deviation inequality for sample covariance matrices

The following result generalizes Theorem D.14 in the case of linear MDPs and generalized counters.

Let $\{X_t\}_{t=1}^\infty$ be a sequence of random vectors of dimension d adapted to a filtration $\{\mathcal{F}_t\}_{t=1}^\infty$ such that $\|X_t\|_2 \leqslant 1$ a.s.. Define a sequence of positive semi-definite matrices $A_t = \mathbb{E}[X_t X_t^\top | \mathcal{F}_{t-1}]$. Notice that $\|A_t\|_2 = \sigma_{\max}(A_t) \leqslant 1$. Also define $\Lambda_t = \lambda I_d + \sum_{j=1}^t X_j X_j^\top$ and $\overline{\Lambda}_t = \lambda I_d + \sum_{j=1}^t A_j$.

Lemma D.18. *Let $\delta \in (0,1)$. Then, the following event*

$$\mathcal{E}^{\mathrm{cnt}'}(\delta) = \left\{\forall t \geqslant 1 : \Lambda_t \succcurlyeq \frac{1}{2}\overline{\Lambda}_t - \beta(\delta, t) I_d\right\}$$

under the choice $\beta(\delta, t) = 4 \log(4\mathrm{e}(2t+1)/\delta) + 4d \log(3t) + 3$ holds with probability at least $1 - \delta$.

Proof. Let us fix a vector v from a unit sphere $\|v\|_2 = 1$. Next, note that

$$v^\top \Lambda_t v - v^\top \overline{\Lambda}_t v = \sum_{j=1}^t \underbrace{\langle v, X_j \rangle^2 - \mathbb{E}\left[\langle v, X_j \rangle^2 | \mathcal{F}_{j-1}\right]}_{\Delta M_j}.$$

Notice that ΔM_j is a martingale-difference sequence that satisfied $|\Delta M_j| \leqslant 2$ and

$$\sum_{j=1}^{t} \mathbb{E}\Big[\Delta M_j^2 | \mathcal{F}_{j-1}\Big] = \sum_{j=1}^{t} \mathbb{E}\Big[\Delta M_j^2 | \mathcal{F}_{j-1}\Big] \leqslant \sum_{j=1}^{t} \mathbb{E}\Big[\langle v, X_j\rangle^4 | \mathcal{F}_{j-1}\Big] \leqslant v^\top \overline{\Lambda}_t v\,.$$

Therefore, applying self-normalized Bernstein inequality (Theorem D.15) we have that with probability at least $1-\delta$ for all $t \geqslant 1$

$$|v^\top \Lambda_t v - v^\top \overline{\Lambda}_t v| \leqslant \sqrt{2 v^\top \overline{\Lambda}_t v \cdot \log(4\mathrm{e}(2t+1)/\delta)} + 3\log(4\mathrm{e}(2t+1)/\delta)\,.$$

Next, inequality $2ab \leqslant a^2 + b^2$ implies that

$$|v^\top \Lambda_t v - v^\top \overline{\Lambda}_t v| \leqslant \frac{1}{2} v^\top \overline{\Lambda}_t v + 4\log(4\mathrm{e}(2t+1)/\delta)\,.$$

Let us denote by $\mathcal{N}_\varepsilon$ a ε-net over a unit sphere of dimension d. By union bound over this net we have with probability at least $1-\delta$

$$\forall \widehat{v} \in \mathcal{N}_\varepsilon : |\widehat{v}^\top \Lambda_t \widehat{v} - v^\top \overline{\Lambda}_t \widehat{v}| \leqslant \frac{1}{2} \widehat{v}^\top \overline{\Lambda}_t \widehat{v} + 4\log(4\mathrm{e}(2t+1)/\delta) + 4\log(|\mathcal{N}_\varepsilon|)\,.$$

Let v be an arbitrary vector on a unit sphere and let $\widehat{v} \in \mathcal{N}_\varepsilon$ be the closest vector to v in the ε-net. Then for any matrix $A \in \mathbb{R}^{d\times d}$ we have

$$|v^\top A v - \widehat{v}^\top A \widehat{v}| \leqslant |v^\top A (v - \widehat{v})| + |(v-\widehat{v})^\top A \widehat{v}| \leqslant 2\varepsilon \|A\|_2\,.$$

Next we notice that $\|\Lambda_t - \overline{\Lambda}_t\|_2 \leqslant 2t$ and $\|\overline{\Lambda}_t\| \leqslant t$. Then we have

$$\forall v \in \mathcal{S}^d : |v^\top \Lambda_t v - v \overline{\Lambda}_t v| \leqslant \frac{1}{2} v^\top \overline{\Lambda}_t v + 4\log(4\mathrm{e}(2t+1)/\delta) + 4\log(|\mathcal{N}_\varepsilon|) + 3t\varepsilon\,.$$

Finally, we have the upper bound on covering number $|\mathcal{N}_\varepsilon| \leqslant (3/\varepsilon)^d$ and, taking $\varepsilon = 1/t$ for all fixed $t \geqslant 1$ we have

$$\forall v \in \mathcal{S}^d : |v^\top \Lambda_t v - v \overline{\Lambda}_t v| \leqslant \frac{1}{2} v^\top \overline{\Lambda}_t v + 4\log(4\mathrm{e}(2t+1)/\delta) + 4d\log(3t) + 3\,.$$

Thus, with probability at least $1-\delta$ we have for all $t \geqslant 1$

$$\frac{3}{2}\overline{\Lambda}_t + \beta(\delta, t) I_d \succcurlyeq \Lambda_t \succcurlyeq \frac{1}{2}\overline{\Lambda}_t - \beta(\delta, t) I_d\,.$$

where $\beta(\delta, t) = 4\log(4\mathrm{e}(2t+1)/\delta) + 4d\log(3t) + 3$.

□

D.6.6 Deviation Inequalities for Expectation over Sampling Measure

In this section we describe two inequalities that shows the deviations of quantities of empirical transition kernels constructed by an independent samples from some base distribution over states. Let $\{z_k\}_{k\in[N]}$ be a set of independent trajectories using a fixed policy π in the original MDP: $\forall i \in [H] : a_i \sim \pi(s_i), s_{i+1} \sim p(s_i, a_i)$, and let $\mu_h(s,a) \triangleq d_h^\pi(s,a)$ be its state-action visitation distribution.

Using this data, we construct an estimates of transition probabilities $\{\widehat{p}_h\}_{h\in[H]}$: for each state-action-step triple (s,a,h) we define $n_h(s,a)$ as a number of visits in these N trajectories:

$$n_h(s,a) = \sum_{k=1}^{N} \mathbb{1}\{(s_h^k, a_h^k) = (s,a)\} \quad n_h(s'|s,a) = \sum_{k=1}^{N} \mathbb{1}\{(s_h^k, a_h^k, s_{h+1}^k) = (s,a,s')\},$$

where $z_k = (s_1^k, a_1^k, \dots, s_H^k, a_H^k, s_{H+1}^k)$, and then the model constructed in a usual way as a maximum likelihood estimate

$$\widehat{p}_h(s'|s,a) = \begin{cases} \frac{n_h(s'|s,a)}{n_h(s,a)} & n_h(s,a) > 0 \\ \frac{1}{S} & n_h(s,a) = 0 \end{cases}.$$

Lemma D.19. *Suppose that $\widehat{p}_h$ is the empirical transitions formed using N independent trajectories $\{z_k\}_{k\in[N]}$ sampled independently using policy π in the original MDP. Then we probability at least $1-\delta$ the following holds*

$$\forall h \in [H] : \mathbb{E}_{(s,a)\sim\mu_h}\Big[(\mathcal{H}(\widehat{p}_h(s,a)) - \mathcal{H}(p_h(s,a)))^2\Big] \leqslant \frac{12S^2A\log^2(SN)\cdot\log(4SAH/\delta)}{N},$$

where $\mu_h(s,a) = d_h^\pi(s,a)$.

Proof. Define $n_h(s,a)$ be a number of samples from a fixed state-action pair s,a. Then by Theorem D.21 we have with probability at least $1-\delta/2$ for all triples $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$

$$(\mathcal{H}(\widehat{p}_h(s,a)) - \mathcal{H}(p_h(s,a)))^2 \leqslant \frac{2\log^2(n_h(s,a))\cdot\log(4SAH/\delta)}{n_h(s,a)} + \frac{S\log(S)\log(n_h(s,a))}{n_h(s,a)}.$$

From other point of view, we have a trivial upper bound $\log^2(S)$. Thus, we obtain

$$(\mathcal{H}(\widehat{p}_h(s,a)) - \mathcal{H}(p_h(s,a)))^2 \leqslant \frac{2\log^2(Sn_h(s,a))\cdot\log(4SAH/\delta) + S\log(S)\log(n_h(s,a)}{n_h(s,a)} \wedge 1.$$

Next we define the following event

$$\mathcal{E}^{\mathrm{cnt}}(\delta) = \Big\{\forall(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H] : n_h(s,a) \geqslant \frac{N}{2}\mu_h(s,a) - \log(2SAH/\delta)\Big\}.$$

By Theorem D.14 it holds with probability at least $1-\delta/2$, then we can apply Lemma D.5 and obtain the following bound with probability at least $1-\delta$ by union bound

$$(\mathcal{H}(\widehat{p}_h(s,a)) - \mathcal{H}(p_h(s,a)))^2 \leqslant 4\frac{2\log^2(SN)\cdot\log(4SAH/\delta) + S\log(S)\log(N)}{(N\mu_h(s,a))\vee 1}.$$

Thus, taking expectation we have

$$\begin{aligned}\mathbb{E}_{(s,a)\sim\mu_h}\Big[(\mathcal{H}(\widehat{p}_h(s,a)) - \mathcal{H}(p_h(s,a)))^2\Big] &\leqslant \sum_{s,a}\frac{4\mu_h(s,a)}{(N\mu_h(s,a))\vee 1}\times \\ &\qquad\times\Big(2\log^2(SN)\log(4SAH/\delta) + S\log(S)\log(N)\Big) \\ &\leqslant \frac{12S^2A\cdot\log^2(SN)\cdot\log(4SAH/\delta)}{N}.\end{aligned}$$

□

Lemma D.20. *[Lemma C.2 by Jin et al. 2020c] Suppose that $\widehat{p}_h$ is the empirical transitions formed using N independent trajectories $\{z_k\}_{k\in[N]}$ sampled independently using policy π in the original MDP. Then there is an absolute constant $C>0$ such that with probability at least $1-\delta$ the following holds for all $h\in[H]$*

$$\max_{G:\,\mathcal{S}\to[0,H\mathrm{R}_{\max}]}\max_{\nu:\,\mathcal{S}\to\mathcal{A}}\mathbb{E}_{(s,a)\sim\mu_h}\Big[\Big([\widehat{p}_{h'} - p_{h'}]G(s,a)\Big)^2\mathbb{1}\{\nu(s)=a\}\Big] \leqslant \frac{CH^2\mathrm{R}_{\max}^2S}{N}\log\Big(\frac{AH\mathrm{R}_{\max}N}{\delta}\Big),$$

where $\mu_h(s,a) = d_h^\pi(s,a)$.

D.6.7 Deviation Inequality for Shannon Entropy

We denote by $\mathcal{H}(p)$ the (Shannon) entropy of $p \in \Delta_{m-1}$,

$$\mathcal{H}(p) \triangleq \sum_{k=1}^{m} p_k \log\left(\frac{1}{p_k}\right).$$

We will follow the ideas of Paninski (2003).

Theorem D.21. *For all $p \in \Delta_{m-1}$ and for all $\delta \in [0,1]$*

$$\mathbb{P}\left[|\mathcal{H}(\widehat{p}_n) - \mathcal{H}(p)| \geqslant \sqrt{\frac{2\log^2(n)\cdot\log(2/\delta)}{n}} + \left(\frac{(m-1)\log(\mathrm{e}(1+n/(m-1)))+1}{n} \wedge \log(m)\right)\right] \leqslant \delta.$$

Moreover,

$$\mathbb{P}\left[\exists n : |\mathcal{H}(\widehat{p}_n) - \mathcal{H}(p)| \geqslant \sqrt{\frac{2\log^2(n)\cdot\log(2n(n+1)/\delta)}{n}} + \left(\frac{m\log(\mathrm{e}(1+n))}{n} \wedge \log(m)\right)\right] \leqslant \delta.$$

Proof. We start from application of McDiarmid's inequality to entropy by Antos and Kontoyiannis (2001). For all $p \in \Delta_{m-1}$ with probability at least $1-\delta$ we have

$$\mathbb{P}\left[|\mathcal{H}(\widehat{p}_n) - \mathbb{E}[\mathcal{H}(\widehat{p}_n)]| \geqslant \sqrt{\frac{2\log^2(n)\cdot\log(2/\delta)}{n}}\right] \leqslant \delta.$$

To relate $\mathbb{E}[\mathcal{H}(\widehat{p}_n)]$ and $\mathcal{H}(p)$ we use the following observation

$$\mathcal{H}(\widehat{p}_n) - \mathcal{H}(p) = -\operatorname{KL}(\widehat{p}_n, p) + \sum_{k:p_k>0} (\widehat{p}_{n,k} - p_k)\log(1/p_k),$$

therefore by taking expectation we have

$$\mathbb{E}[\mathcal{H}(\widehat{p}_n)] - \mathcal{H}(p) = -\mathbb{E}[\operatorname{KL}(\widehat{p}_n, p)].$$

In the following our analysis differs from Paninski (2003) since we obtain a direct estimate on the KL-divergence using Theorem D.13,

$$\mathbb{E}[n\operatorname{KL}(\widehat{p}_n, p)] = \int_0^\infty \mathbb{P}[n\operatorname{KL}(\widehat{p}_n, p) > t]\mathrm{d}t \leqslant (m-1)\log(\mathrm{e}(1+n/(m-1))) + \int_0^\infty \mathrm{e}^{-t}\mathrm{d}t.$$

At the same time we have a trivial bound that concludes the first statement

$$\mathbb{E}[\mathcal{H}(\widehat{p}^n)] - \mathcal{H}(p) \leqslant \log(m).$$

To show the second statement of Theorem D.21, we apply the first part with $\delta'(n) = \delta/(n(n+1))$,

$$\mathbb{P}\left[|\mathcal{H}(\widehat{p}_n) - \mathcal{H}(p)| \geqslant \sqrt{\frac{2\log^2(n)\cdot\log(2/\delta'(n))}{n}} + \left(\frac{(m-1)\log(\mathrm{e}(1+n/(m-1)))+1}{n} \wedge \log(m)\right)\right] \leqslant \frac{\delta}{n(n+1)},$$

thus by union bound over $n \in \mathbb{N}$ we conclude the statement. □

Bibliography

Abbasi-Yadkori, Yasin, Dávid Pál, and Csaba Szepesvári (2011). "Improved Algorithms for Linear Stochastic Bandits". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor et al. Vol. 24. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Abbeel, Pieter and Andrew Y. Ng (2004). "Apprenticeship learning via inverse reinforcement learning". In: *Proceedings of the Twenty-First International Conference on Machine Learning*. ICML '04. Banff, Alberta, Canada: Association for Computing Machinery, p. 1. ISBN: 1581138385. DOI: 10.1145/1015330.1015430. URL: https://doi.org/10.1145/1015330.1015430.

Abernethy, Jacob D and Jun-Kun Wang (2017). "On Frank-Wolfe and Equilibrium Computation". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/7371364b3d72ac9a3ed8638e6f0be2c9-Paper.pdf.

Agarwal, Alekh et al. (2020a). "FLAMBE: Structural Complexity and Representation Learning of Low Rank MDPs". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 20095–20107. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/e894d787e2fd6c133af47140aa156f00-Paper.pdf.

Agarwal, Alekh et al. (July 2020b). "Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, pp. 64–66. URL: https://proceedings.mlr.press/v125/agarwal20a.html.

— (2021). "On the Theory of Policy Gradient Methods: Optimality, Approximation, and Distribution Shift". In: *Journal of Machine Learning Research* 22.98, pp. 1–76. URL: http://jmlr.org/papers/v22/19-736.html.

Agrawal, Shipra and Randy Jia (2017). "Optimistic posterior sampling for reinforcement learning: worst-case regret bounds". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/3621f1454cacf995530ea53652ddf8fb-Paper.pdf.

Al Marjani, Aymen, Aurélien Garivier, and Alexandre Proutiere (2021). "Navigating to the Best Policy in Markov Decision Processes". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 25852–25864. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/d9896106ca98d3d05b8cbdf4fd8b13a1-Paper.pdf.

Altman, Eitan (2021). *Constrained Markov decision processes*. Routledge. DOI: https://doi.org/10.1201/9781315140223.

Antos, András and Ioannis Kontoyiannis (2001). "Convergence properties of functional estimates for discrete distributions". In: *Random Structures & Algorithms* 19.3-4, pp. 163–193. DOI: https://doi.org/10.1002/rsa.10019. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rsa.10019. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.10019.

Atkeson, Christopher G and Stefan Schaal (1997). "Robot learning from demonstration". In: *ICML*. Vol. 97, pp. 12–20.

Auer, P., N. Cesa-Bianchi, and C. Gentile (2002). "Adaptive and self-confident on-line learning algorithms". In: *Journal of Computer and System Sciences* 64.1, pp. 48–75.

Auer, Peter, Thomas Jaksch, and Ronald Ortner (2008). "Near-optimal Regret Bounds for Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller et al. Vol. 21. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2008/file/e4a6222cdb5b34375400904f03d8e6a5-Paper.pdf.

Aytar, Yusuf et al. (2018). "Playing hard exploration games by watching YouTube". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/35309226eb45ec366ca86a4329a2b7c3-Paper.pdf.

Azar, Mohammad Gheshlaghi, Rémi Munos, and Hilbert J. Kappen (2013). "Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model". In: *Machine Learning* 91.3, pp. 325–349. URL: https://hal.archives-ouvertes.fr/hal-00831875.

Azar, Mohammad Gheshlaghi, Ian Osband, and Rémi Munos (Aug. 2017). "Minimax Regret Bounds for Reinforcement Learning". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 263–272. URL: https://proceedings.mlr.press/v70/azar17a.html.

Azuma, Kazuoki (1967). "Weighted sums of certain dependent random variables". In: *Tohoku Mathematical Journal, Second Series* 19.3, pp. 357–367.

Bartlett, Peter L and Shahar Mendelson (2006). "Empirical minimization". In: *Probability theory and related fields* 135.3, pp. 311–334.

Bauschke, Heinz H., Jérôme Bolte, and Marc Teboulle (2017). "A Descent Lemma Beyond Lipschitz Gradient Continuity: First-Order Methods Revisited and Applications". In: *Mathematics of Operations Research* 42.2, pp. 330–348. DOI: 10.1287/moor.2016.0817. eprint: https://doi.org/10.1287/moor.2016.0817. URL: https://doi.org/10.1287/moor.2016.0817.

Bellemare, Marc G. et al. (2016). "Unifying count-based exploration and intrinsic motivation". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., pp. 1479–1487. ISBN: 9781510838819.

Bellman, Richard (1957). *Dynamic Programming*. Dover Publications. ISBN: 9780486428093.

Belomestny, Denis et al. (2023). "Sharp Deviations Bounds for Dirichlet Weighted Sums with Application to analysis of Bayesian algorithms". In: *arXiv preprint arXiv:2304.03056*.

Bengio, Emmanuel et al. (2021). "Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 27381–27394. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/e614f646836aaed9f89ce58e837e2310-Paper.pdf.

Bernstein, S. (1926). *Leçons sur les propriétés extrémales et la meilleure approximation des fonctions analytiques d'une variable réelle*. Collection de monographies sur la théorie des fonctions. Gauthier-Villars. URL: https://books.google.fr/books?id=vRbCzwEACAAJ.

Boucheron, Stéphane, Gabor Lugosi, and Pascal Massart (2013). *Concentration inequalities : a non asymptotic theory of independence*. Oxford University Press, p. 481. URL: https://hal.inria.fr/hal-00942704.

Boyd, Stephen and Lieven Vandenberghe (2004). *Convex Optimization*. Cambridge University Press. DOI: 10.1017/CBO9780511804441.

Bradley, Ralph Allan and Milton E Terry (1952). "Rank analysis of incomplete block designs: I. The method of paired comparisons". In: *Biometrika* 39.3/4, pp. 324–345.

Bubeck, Sébastien (Nov. 2015). "Convex Optimization: Algorithms and Complexity". In: *Found. Trends Mach. Learn.* 8.3–4, pp. 231–357. ISSN: 1935-8237. DOI: 10.1561/2200000050. URL: https://doi.org/10.1561/2200000050.

Burda, Yuri et al. (2019). "Exploration by random network distillation". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=H1lJJnR5Ym.

Burnetas, Apostolos N. and Michael N. Katehakis (1997). "Optimal Adaptive Policies for Markov Decision Processes". In: *Mathematics of Operations Research* 22.1, pp. 222–255. DOI: 10.1287/moor.22.1.222. eprint: https://doi.org/10.1287/moor.22.1.222. URL: https://doi.org/10.1287/moor.22.1.222.

Busa-Fekete, Róbert et al. (Dec. 2014). "Preference-Based Reinforcement Learning: Evolutionary Direct Policy Search Using a Preference-Based Racing Algorithm". In: *Mach. Learn.* 97.3, pp. 327–351. ISSN: 0885-6125. DOI: 10.1007/s10994-014-5458-8. URL: https://doi.org/10.1007/s10994-014-5458-8.

Cai, Qi et al. (July 2020). "Provably Efficient Exploration in Policy Optimization". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 1283–1294. URL: https://proceedings.mlr.press/v119/cai20d.html.

Cassel, Asaf and Aviv Rosenberg (2024). "Warm-up Free Policy Optimization: Improved Regret in Linear Markov Decision Processes". In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson et al. Vol. 37. Curran Associates, Inc., pp. 3275–3303. DOI: 10.52202/079017-0108. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/05d42b7bd130ecbc57fdf02cbdcd370e-Paper-Conference.pdf.

Cen, Shicong et al. (2022). "Fast global convergence of natural policy gradient methods with entropy regularization". In: *Operations Research* 70.4, pp. 2563–2578.

Cesa-Bianchi, N. and G. Lugosi (2003). "Potential-Based Algorithms in On-Line Prediction and Game Theory". In: *Machine Learning* 51, pp. 239–261.

Cesa-Bianchi, N., Y. Mansour, and G. Stoltz (2007). "Improved second-order bounds for prediction with expert advice". In: *Machine Learning* 66, pp. 321–352.

Cesa-Bianchi, Nicolò and Gábor Lugosi (2006). *Prediction, learning, and games*. Cambridge University Press, pp. I–XII, 1–394. ISBN: 978-0-511-54692-1.

Cesa-Bianchi, Nicolò et al. (2017). "Boltzmann Exploration Done Right". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., pp. 6287–6296. ISBN: 9781510860964.

Chaves, Pedro Dalla Vecchia, Bruno L. Pereira, and Rodrygo L. T. Santos (2022). "Efficient Online Learning to Rank for Sequential Music Recommendation". In: *Proceedings of the ACM Web Conference 2022*. WWW '22. Virtual Event, Lyon, France: Association for Computing Machinery, pp. 2442–2450. ISBN: 9781450390965. DOI: 10.1145/3485447.3512116. URL: https://doi.org/10.1145/3485447.3512116.

Chentanez, Nuttapong, Andrew Barto, and Satinder Singh (2004). "Intrinsically Motivated Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. Ed. by L. Saul, Y. Weiss, and L. Bottou. Vol. 17. MIT Press. URL: https://proceedings.neurips.cc/paper/2004/file/4be5a36cbaca8ab9d2066debfe4e65c1-Paper.pdf.

Cheung, Wang Chi (2019). *Exploration-Exploitation Trade-off in Reinforcement Learning on Online Markov Decision Processes with Global Concave Rewards*. arXiv: 1905.06466 [cs.LG]. URL: https://arxiv.org/abs/1905.06466.

Chow, Y. and H. Teicher (1988). *Probability Theory*. Springer.

Christiano, Paul F et al. (2017). "Deep Reinforcement Learning from Human Preferences". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Cover, Thomas M. and Joy A. Thomas (2006). *Elements of information theory*. John Wiley & Sons. URL: https://www.amazon.com/Elements-Information-Theory-Telecommunications-Processing/dp/0471241954.

Dann, Christoph, Tor Lattimore, and Emma Brunskill (2017). "Unifying PAC and Regret: Uniform PAC bounds for episodic reinforcement learning". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/17d8da815fa21c57af9829fb0a869602-Paper.pdf.

Dann, Christoph et al. (June 2019). "Policy Certificates: Towards Accountable Reinforcement Learning". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 1507–1516. URL: https://proceedings.mlr.press/v97/dann19a.html.

Degrave, Jonas et al. (Feb. 2022). "Magnetic control of tokamak plasmas through deep reinforcement learning". In: *Nature* 602.7897, pp. 414–419. ISSN: 1476-4687. DOI: 10.1038/s41586-021-04301-9. URL: https://doi.org/10.1038/s41586-021-04301-9.

Domingues, Omar Darwiche et al. (Mar. 2021a). "Episodic Reinforcement Learning in Finite MDPs: Minimax Lower Bounds Revisited". In: *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*. Ed. by Vitaly Feldman, Katrina Ligett, and Sivan Sabato. Vol. 132. Proceedings of Machine Learning Research. PMLR, pp. 578–598. URL: https://proceedings.mlr.press/v132/domingues21a.html.

Domingues, Omar Darwiche et al. (July 2021b). "Kernel-Based Reinforcement Learning: A Finite-Time Analysis". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 2783–2792. URL: https://proceedings.mlr.press/v139/domingues21a.html.

Donsker, Monroe D and SR Srinivasa Varadhan (1983). "Asymptotic evaluation of certain Markov process expectations for large time. IV". In: *Communications on pure and applied mathematics* 36.2, pp. 183–212.

Doob, J.L. (1953). *Stochastic Processes*. Wiley Publications in Statistics. John Wiley & Sons.

Douc, Randal et al. (2018). *Markov chains*. Operation research and financial engineering. Springer, p. 757. DOI: 10.1007/978-3-319-97704-1. URL: https://hal.science/hal-02022651.

Dudley, Richard M (2014). *Uniform central limit theorems*. Vol. 142. Cambridge university press.

Ecoffet, Adrien et al. (Feb. 2021). "First return, then explore". In: *Nature* 590.7847, pp. 580–586. ISSN: 1476-4687. DOI: 10.1038/s41586-020-03157-9. URL: https://doi.org/10.1038/s41586-020-03157-9.

Ekroot, Laura and Thomas M. Cover (1993). "The entropy of Markov trajectories". In: *IEEE Transactions on Information Theory* 39.4, pp. 1418–1421. DOI: 10.1109/18.243461.

Erven, Tim et al. (2011). "Adaptive Hedge". In: *Advances in Neural Information Processing Systems* 24.

Even-Dar, Eyal, Sham. M. Kakade, and Yishay Mansour (2009). "Online Markov Decision Processes". In: *Mathematics of Operations Research* 34.3, pp. 726–736. ISSN: 0364765X, 15265471. URL: http://www.jstor.org/stable/40538442 (visited on 11/23/2025).

Eysenbach, Benjamin and Sergey Levine (2019). *If MaxEnt RL is the Answer, What is the Question?* arXiv: 1910.01913 [cs.LG]. URL: https://arxiv.org/abs/1910.01913.

Fiechter, Claude-Nicolas (1994). "Efficient reinforcement learning". In: *Proceedings of the Seventh Annual Conference on Computational Learning Theory*. COLT '94. New Brunswick, New Jersey, USA: Association for Computing Machinery, pp. 88–97. ISBN: 0897916557. DOI: 10.1145/180139.181019. URL: https://doi.org/10.1145/180139.181019.

Fox, Roy, Ari Pakman, and Naftali Tishby (2016). "Taming the Noise in Reinforcement Learning via Soft Updates". In: *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, UAI 2016, June 25-29, 2016, New York City, NY, USA*. Ed. by Alexander Ihler and Dominik Janzing. AUAI Press. URL: http://auai.org/uai2016/proceedings/papers/219.pdf.

Frank, Marguerite and Philip Wolfe (1956). "An algorithm for quadratic programming". In: *Naval Research Logistics Quarterly* 3.1-2, pp. 95–110. DOI: https://doi.org/10.1002/nav.3800030109. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nav.3800030109. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800030109.

Freund, Yoav and Robert E Schapire (1997). "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of computer and system sciences* 55.1, pp. 119–139.

Fruit, Ronan et al. (July 2018). "Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1578–1586. URL: https://proceedings.mlr.press/v80/fruit18a.html.

Gaillard, Pierre, Gilles Stoltz, and Tim van Erven (June 2014). "A second-order bound with excess losses". In: *Proceedings of The 27th Conference on Learning Theory*. Ed. by Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári. Vol. 35. Proceedings of Machine Learning Research. Barcelona, Spain: PMLR, pp. 176–196. URL: https://proceedings.mlr.press/v35/gaillard14.html.

Garivier, A. et al. (2022). "KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints". In: *Journal of Machine Learning Research* 23.179, pp. 1–66.

Geer, Sara A van de (2000). *Empirical Processes in M-estimation*. Vol. 6. Cambridge university press.

Geist, Matthieu, Bruno Scherrer, and Olivier Pietquin (June 2019). "A Theory of Regularized Markov Decision Processes". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 2160–2169. URL: https://proceedings.mlr.press/v97/geist19a.html.

Geist, Matthieu et al. (2022). "Concave Utility Reinforcement Learning: The Mean-field Game Viewpoint". In: *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. AAMAS '22. Virtual Event, New Zealand: International Foundation for Autonomous Agents and Multiagent Systems, pp. 489–497. ISBN: 9781450392136.

Goecks, Vinicius G. et al. (2020). "Integrating Behavior Cloning and Reinforcement Learning for Improved Performance in Dense and Sparse Reward Environments". In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '20. Auckland, New Zealand: International Foundation for Autonomous Agents and Multiagent Systems, pp. 465–473. ISBN: 9781450375184.

Grill, Jean-Bastien et al. (2019). "Planning in entropy-regularized Markov decision processes and games". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/50982fb2f2cfa186d335310461dfa2be-Paper.pdf.

Gritsaev, Timofei et al. (2025a). "Adaptive Destruction Processes for Diffusion Samplers". In: *arXiv preprint arXiv:2506.01541*.

Gritsaev, Timofei et al. (2025b). "Optimizing Backward Policies in GFlowNets via Trajectory Likelihood Maximization". In: *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Xj66fkrlTk.

Grünwald, Peter D. and A. Philip Dawid (2002). "Game theory, maximum generalized entropy, minimum discrepancy, robust Bayes and Pythagoras". In: *Proceedings of the 2002 IEEE Information Theory Workshop, ITW 2002, 20-25 October 2002, Bangalore, India*. IEEE, pp. 94–97. DOI: 10.1109/ITW.2002.1115425. URL: https://doi.org/10.1109/ITW.2002.1115425.

Haarnoja, Tuomas et al. (Aug. 2017). "Reinforcement Learning with Deep Energy-Based Policies". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup

and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1352–1361. URL: https://proceedings.mlr.press/v70/haarnoja17a.html.

Haarnoja, Tuomas et al. (July 2018). "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1861–1870. URL: https://proceedings.mlr.press/v80/haarnoja18b.html.

Handel, Ramon van (2016). "Probability in High Dimensions". In: URL: https://web.math.princeton.edu/~rvan/APC550.pdf.

Hazan, Elad, Tomer Koren, and Kfir Y. Levy (June 2014). "Logistic Regression: Tight Bounds for Stochastic and Online Optimization". In: *Proceedings of The 27th Conference on Learning Theory*. Ed. by Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári. Vol. 35. Proceedings of Machine Learning Research. Barcelona, Spain: PMLR, pp. 197–209. URL: https://proceedings.mlr.press/v35/hazan14a.html.

Hazan, Elad et al. (June 2019). "Provably Efficient Maximum Entropy Exploration". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 2681–2691. URL: https://proceedings.mlr.press/v97/hazan19a.html.

He, Jiafan, Dongruo Zhou, and Quanquan Gu (Mar. 2022). "Near-optimal Policy Optimization Algorithms for Learning Adversarial Linear Mixture MDPs". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, pp. 4259–4280. URL: https://proceedings.mlr.press/v151/he22a.html.

Hester, Todd et al. (2018). "Deep Q-Learning from Demonstrations". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI'18/IAAI'18/EAAI'18. New Orleans, Louisiana, USA: AAAI Press. ISBN: 978-1-57735-800-8.

Ho, Jonathan and Stefano Ermon (2016). "Generative Adversarial Imitation Learning". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/cc7e2b878868cbae992d1fb743995d8f-Paper.pdf.

Hoeffding, Wassily (1963). "Probability Inequalities for Sums of Bounded Random Variables". In: *Journal of the American Statistical Association* 58.301, pp. 13–30. DOI: 10.1080/01621459.1963.10500830. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1963.10500830. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830.

Hoeven, Dirk van der, Nikita Zhivotovskiy, and Nicolò Cesa-Bianchi (2023). *High-Probability Risk Bounds via Sequential Predictors*. arXiv: 2308.07588 [cs.LG].

Hosu, I.-A. and T. Rebedea (2016). "Playing Atari Games with Deep Reinforcement Learning and Human Checkpoint Replay". In: *ECAI Workshop on Evaluating General Purpose AI*.

Howard, Ronald A (1960). "Dynamic programming and markov processes." In.

Huang, Audrey et al. (2025). "Correcting the Mythos of KL-Regularization: Direct Alignment without Overoptimization via Chi-Squared Preference Optimization". In: *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=hXm0Wu2U9K.

Huang, Jiayi et al. (2023). "Tackling Heavy-Tailed Rewards in Reinforcement Learning with Function Approximation: Minimax Optimal and Instance-Dependent Regret Bounds". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc.,

pp. 56576–56588. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/b11393733b1ea5890100302ab8a0f74c-Paper-Conference.pdf.

Jain, Ashesh et al. (2013). "Learning Trajectory Preferences for Manipulators via Iterative Improvement". In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/c058f544c737782deacefa532d9add4c-Paper.pdf.

Jaksch, Thomas, Ronald Ortner, and Peter Auer (2010). "Near-optimal Regret Bounds for Reinforcement Learning". In: *Journal of Machine Learning Research* 11.51, pp. 1563–1600. URL: http://jmlr.org/papers/v11/jaksch10a.html.

Jin, Chi et al. (2018). "Is Q-Learning Provably Efficient?" In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/d3b1fb02964aa64e257f9f26a31f72cf-Paper.pdf.

Jin, Chi et al. (July 2020a). "Learning Adversarial Markov Decision Processes with Bandit Feedback and Unknown Transition". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 4860–4869. URL: https://proceedings.mlr.press/v119/jin20c.html.

Jin, Chi et al. (July 2020b). "Provably efficient reinforcement learning with linear function approximation". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, pp. 2137–2143. URL: https://proceedings.mlr.press/v125/jin20a.html.

Jin, Chi et al. (July 2020c). "Reward-Free Exploration for Reinforcement Learning". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 4870–4879. URL: https://proceedings.mlr.press/v119/jin20d.html.

Jin, Tiancheng, Longbo Huang, and Haipeng Luo (2021). "The best of both worlds: stochastic and adversarial episodic MDPs with unknown transition". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 20491–20502. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/abb9d15b3293a96a3ea116867b2b16d5-Paper.pdf.

Jonckheere, Matthieu, Chiara Mignacco, and Gilles Stoltz (2025). "Policy Optimization via Adv2: Adversarial Learning on Advantage Functions". In: *Transactions on Machine Learning Research*. ISSN: 2835-8856. URL: https://openreview.net/forum?id=Oyueig10Ed.

Jonsson, Anders et al. (2020). "Planning in Markov Decision Processes with Gap-Dependent Sample Complexity". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 1253–1263. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/0d85eb24e2add96ff1a7021f83c1abc9-Paper.pdf.

Kakade, S. and J. Langford (2002). "Approximately optimal approximate reinforcement learning". In: *Proceedings of the 19th International Conference on Machine Learning (ICML'02)*, pp. 267–274.

Kakade, Sham, Shai Shalev-Shwartz, Ambuj Tewari, et al. (2009). *On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization*. URL: http://ttic.uchicago.edu/shai/papers/KakadeShalevTewari09.pdf.

Kakade, Sham M (2001). "A Natural Policy Gradient". In: *Advances in Neural Information Processing Systems*. Ed. by T. Dietterich, S. Becker, and Z. Ghahramani. Vol. 14. MIT Press. URL: https://proceedings.neurips.cc/paper_files/paper/2001/file/4b86abe48d358ecf194c56c69108433e-Paper.pdf.

Kakade, Sham Machandranath (2003). *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom).

Kang, Bingyi, Zequn Jie, and Jiashi Feng (July 2018). "Policy Optimization with Demonstrations". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 2469–2478. URL: https://proceedings.mlr.press/v80/kang18a.html.

Kaufmann, Emilie, Olivier Cappe, and Aurelien Garivier (Apr. 2012). "On Bayesian Upper Confidence Bounds for Bandit Problems". In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Neil D. Lawrence and Mark Girolami. Vol. 22. Proceedings of Machine Learning Research. La Palma, Canary Islands: PMLR, pp. 592–600. URL: https://proceedings.mlr.press/v22/kaufmann12.html.

Kaufmann, Emilie et al. (Mar. 2021). "Adaptive Reward-Free Exploration". In: *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*. Ed. by Vitaly Feldman, Katrina Ligett, and Sivan Sabato. Vol. 132. Proceedings of Machine Learning Research. PMLR, pp. 865–891. URL: https://proceedings.mlr.press/v132/kaufmann21a.html.

Kivinen, Jyrki and Manfred K Warmuth (1997). "Exponentiated gradient versus gradient descent for linear predictors". In: *Information and computation* 132.1, pp. 1–63.

Labbi, Safwan et al. (May 2025a). "Federated UCBVI: Communication-Efficient Federated Regret Minimization with Heterogeneous Agents". In: *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*. Ed. by Yingzhen Li et al. Vol. 258. Proceedings of Machine Learning Research. PMLR, pp. 1315–1323. URL: https://proceedings.mlr.press/v258/labbi25a.html.

Labbi, Safwan et al. (2025b). "On Global Convergence Rates for Federated Policy Gradient under Heterogeneous Environment". In: *arXiv preprint arXiv:2505.23459*.

Lai, Tze Leung and Herbert Robbins (1985). "Asymptotically efficient adaptive allocation rules". In: *Advances in applied mathematics* 6.1, pp. 4–22.

Lakshminarayanan, A. S., S. Ozair, and Y. Bengio (2016). "Reinforcement Learning with Few Expert Demonstrations". In: *NIPS Workshop on Deep Learning for Action and Interaction*.

Lancewicki, T., A. Rosenberg, and Y. Mansour (2022). "Learning adversarial Markov decision processes with delayed feedback". In: *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI'22)*. Vol. 7, pp. 7281–7289.

Lee, Harrison et al. (July 2024). "RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback". In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov et al. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 26874–26901. URL: https://proceedings.mlr.press/v235/lee24t.html.

Lee, Lisa et al. (2020). *Efficient Exploration via State Marginal Matching*. arXiv: 1906.05274 [cs.LG]. URL: https://arxiv.org/abs/1906.05274.

Li, Gen et al. (2024). "Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Generative Model". In: *Operations Research* 72.1, pp. 203–221. DOI: 10.1287/opre.2023.2451. eprint: https://doi.org/10.1287/opre.2023.2451. URL: https://doi.org/10.1287/opre.2023.2451.

Li, Lihong et al. (2010). "A contextual-bandit approach to personalized news article recommendation". In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. Raleigh, North Carolina, USA: Association for Computing Machinery, pp. 661–670. ISBN: 9781605587998. DOI: 10.1145/1772690.1772758. URL: https://doi.org/10.1145/1772690.1772758.

Lim, Shiau Hong and Peter Auer (June 2012). "Autonomous Exploration For Navigating In MDPs". In: *Proceedings of the 25th Annual Conference on Learning Theory*. Ed. by Shie Mannor, Nathan Srebro, and Robert C. Williamson. Vol. 23. Proceedings of Machine Learning Research. Edinburgh, Scotland: PMLR, pp. 40.1–40.24. URL: https://proceedings.mlr.press/v23/lim12.html.

Littlestone, Nick and Manfred K Warmuth (1994). "The weighted majority algorithm". In: *Information and computation* 108.2, pp. 212–261.

Liu, H., C.-Y. Wei, and J. Zimmert (2024). "Towards Optimal Regret in Adversarial Linear MDPs with Bandit Feedback". In: *Proceedings of the 12th International Conference on Learning Representations (ICLR'24)*. URL: https://openreview.net/forum?id=6yv8UHVJn4.

Luo, H., C.-Y. Wei, and C.-W. Lee (2021). "Policy optimization in adversarial MDPs: Improved exploration via dilated bonuses". In: *Advances in Neural Information Processing Systems* 34, pp. 22931–22942.

Malkin, Nikolay et al. (2022). "Trajectory balance: Improved credit assignment in GFlowNets". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., pp. 5955–5967. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/27b51baca8377a0cf109f6ecc15a0f70-Paper-Conference.pdf.

Mankowitz, Daniel J. et al. (Jan. 2023). "Faster sorting algorithms discovered using deep reinforcement learning". In: *Nature* 618.7964, pp. 257–263. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06004-9. URL: https://doi.org/10.1038/s41586-023-06004-9.

McAllester, David A. (1999). "PAC-Bayesian model averaging". In: *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*. COLT '99. Santa Cruz, California, USA: Association for Computing Machinery, pp. 164–170. ISBN: 1581131674. DOI: 10.1145/307400.307435. URL: https://doi.org/10.1145/307400.307435.

Mei, Jincheng et al. (July 2020). "On the Global Convergence Rates of Softmax Policy Gradient Methods". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 6820–6829. URL: https://proceedings.mlr.press/v119/mei20b.html.

Menard, Pierre et al. (July 2021). "UCB Momentum Q-learning: Correcting the bias without forgetting". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 7609–7618. URL: https://proceedings.mlr.press/v139/menard21b.html.

Ménard, Pierre et al. (July 2021). "Fast active learning for pure exploration in reinforcement learning". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 7599–7608. URL: https://proceedings.mlr.press/v139/menard21a.html.

Meyer, Jean-Arcady and Stewart W. Wilson (1991). "A Possibility for Implementing Curiosity and Boredom in Model-Building Neural Controllers". In: *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, pp. 222–227.

Mnih, Volodymyr et al. (2015). "Human-level control through deep reinforcement learning". In: *nature* 518.7540, pp. 529–533.

Morozov, Nikita et al. (2024). "Improving gflownets with monte carlo tree search". In: *arXiv preprint arXiv:2406.13655*.

Morozov, Nikita et al. (2025). "Revisiting Non-Acyclic GFlowNets in Discrete Environments". In: *Forty-second International Conference on Machine Learning*. URL: https://openreview.net/forum?id=czqxB524oo.

Mutti, Mirco, Mattia Mancassola, and Marcello Restelli (June 2022). "Unsupervised Reinforcement Learning in Multiple Environments". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.7, pp. 7850–7858. DOI: 10.1609/aaai.v36i7.20754. URL: https://ojs.aaai.org/index.php/AAAI/article/view/20754.

Mutti, Mirco, Lorenzo Pratissoli, and Marcello Restelli (2021). "Task-Agnostic Exploration via Policy Gradient of a Non-Parametric State Entropy Estimate". In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Arti-*

ficial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. AAAI Press, pp. 9028–9036. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17091.

Mutti, Mirco and Marcello Restelli (Apr. 2020). "An Intrinsically-Motivated Approach for Learning Highly Exploring and Fast Mixing Policies". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.04, pp. 5232–5239. DOI: 10.1609/aaai.v34i04.5968. URL: https://ojs.aaai.org/index.php/AAAI/article/view/5968.

Nachum, Ofir et al. (2017). "Bridging the Gap Between Value and Policy Based Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/facf9f743b083008a894eee7baa16469-Paper.pdf.

Nair, Ashvin et al. (2018). "Overcoming Exploration in Reinforcement Learning with Demonstrations". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. Brisbane, Australia: IEEE Press, pp. 6292–6299. DOI: 10.1109/ICRA.2018.8463162. URL: https://doi.org/10.1109/ICRA.2018.8463162.

Nair, Ashvin et al. (2021). *AWAC: Accelerating Online Reinforcement Learning with Offline Datasets*. arXiv: 2006.09359 [cs.LG]. URL: https://arxiv.org/abs/2006.09359.

Neu, Gergely, Anders Jonsson, and Vicenç Gómez (2017). *A unified view of entropy-regularized Markov decision processes*. arXiv: 1705.07798 [cs.LG]. URL: https://arxiv.org/abs/1705.07798.

Neu, Gergely et al. (2010). "Online Markov Decision Processes under Bandit Feedback". In: *Advances in Neural Information Processing Systems*. Ed. by J. Lafferty et al. Vol. 23. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2010/file/7bb060764a818184ebb1cc0d43d382aa-Paper.pdf.

Ng, Andrew Y. and Stuart J. Russell (2000). "Algorithms for Inverse Reinforcement Learning". In: *Proceedings of the Seventeenth International Conference on Machine Learning*. ICML '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 663–670. ISBN: 1558607072.

Novoseller, Ellen et al. (Aug. 2020). "Dueling Posterior Sampling for Preference-Based Reinforcement Learning". In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*. Ed. by Jonas Peters and David Sontag. Vol. 124. Proceedings of Machine Learning Research. PMLR, pp. 1029–1038. URL: https://proceedings.mlr.press/v124/novoseller20a.html.

Ocello, Antonio et al. (2025). "Finite-Sample Convergence Bounds for Trust Region Policy Optimization in Mean Field Games". In: *Forty-second International Conference on Machine Learning*. URL: https://openreview.net/forum?id=NE9ajrddxX.

OpenAI et al. (2019). *Dota 2 with Large Scale Deep Reinforcement Learning*. arXiv: 1912.06680 [cs.LG]. URL: https://arxiv.org/abs/1912.06680.

Orabona, F. (2019). *A modern introduction to online learning*. Preprint, arXiv:1912.13213.

Orabona, F. and D. Pál (2015). "Scale-Free Algorithms for Online Linear Optimization". In: *Proceedings of the 26th International Conference on Algorithmic Learning Theory (ALT'15)*. Springer, pp. 287–301.

Oudeyer, Pierre-Yves, Frdric Kaplan, and Verena V. Hafner (2007). "Intrinsic Motivation Systems for Autonomous Mental Development". In: *IEEE Transactions on Evolutionary Computation* 11.2, pp. 265–286. DOI: 10.1109/TEVC.2006.890271.

Ouyang, Long et al. (2022). "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., pp. 27730–27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Paninski, Liam (June 2003). "Estimation of Entropy and Mutual Information". In: *Neural Computation* 15.6, pp. 1191–1253. ISSN: 0899-7667. DOI: 10.1162/089976603321780272. eprint: https

://direct.mit.edu/neco/article-pdf/15/6/1191/815550/089976603321780272.pdf. URL: https://doi.org/10.1162/089976603321780272.

Pathak, Deepak et al. (Aug. 2017). "Curiosity-driven Exploration by Self-supervised Prediction". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 2778–2787. URL: https://proceedings.mlr.press/v70/pathak17a.html.

Pertsch, Karl et al. (2021). "Demonstration-Guided Reinforcement Learning with Learned Skills". In: *5th Conference on Robot Learning*.

Pettigrew, Richard (2019). *Choosing for changing selves*. Oxford University Press.

Pomerleau, Dean A. (1988). "ALVINN: An Autonomous Land Vehicle in a Neural Network". In: *Advances in Neural Information Processing Systems*. Ed. by David S. Touretzky. Vol. 1. Morgan-Kaufmann, pp. 305–313. URL: https://proceedings.neurips.cc/paper_files/paper/1988/file/812b4ba287f5ee0bc9d43bbf5bbe87fb-Paper.pdf.

Puterman, Martin L (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Rajaraman, Nived et al. (2020). "Toward the Fundamental Limits of Imitation Learning". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 2914–2924. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1e7875cf32d306989d80c14308f3a099-Paper.pdf.

Rajaraman, Nived et al. (2021). "On the Value of Interaction and Function Approximation in Imitation Learning". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 1325–1336. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/09dbc1177211571ef3e1ca961cc39363-Paper.pdf.

Rajeswaran, Aravind et al. (2018). "Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations". In: *Proceedings of Robotics: Science and Systems (RSS)*.

Rashidinejad, Paria et al. (2021). "Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 11702–11716. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/60ce36723c17bbac504f2ef4c8a46995-Paper.pdf.

Rockafellar, R.T. (1997). *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press. ISBN: 9780691015866. URL: https://books.google.fr/books?id=1TiOka9bx3sC.

Rooij, S. de et al. (2014). "Follow the Leader If You Can, Hedge If You Must". In: *Journal of Machine Learning Research* 15.37, pp. 1281–1316.

Rosenberg, Aviv and Yishay Mansour (June 2019a). "Online Convex Optimization in Adversarial Markov Decision Processes". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 5478–5486. URL: https://proceedings.mlr.press/v97/rosenberg19a.html.

— (2019b). "Online Stochastic Shortest Path with Bandit Feedback and Unknown Transition Function". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/a0872cc5b5ca4cc25076f3d868e1bdf8-Paper.pdf.

Ross, Stephane and Drew Bagnell (May 2010). "Efficient Reductions for Imitation Learning". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, pp. 661–668. URL: https://proceedings.mlr.press/v9/ross10a.html.

Ross, Stephane, Geoffrey Gordon, and Drew Bagnell (Apr. 2011). "A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Geoffrey Gordon, David Dunson, and Miroslav Dudík. Vol. 15. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, pp. 627–635. URL: https://proceedings.mlr.press/v15/ross11a.html.

Russo, Daniel (2019). "Worst-Case Regret Bounds for Exploration via Randomized Value Functions". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/451ae86722d26a608c2e174b2b2773f1-Paper.pdf.

Saha, Aadirupa, Aldo Pacchiano, and Jonathan Lee (Apr. 2023). "Dueling RL: Reinforcement Learning with Trajectory Preferences". In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Ed. by Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent. Vol. 206. Proceedings of Machine Learning Research. PMLR, pp. 6263–6289. URL: https://proceedings.mlr.press/v206/saha23a.html.

Samsonov, Sergey et al. (May 2024). "Improved High-Probability Bounds for the Temporal Difference Learning Algorithm via Exponential Stability". In: *Proceedings of Thirty Seventh Conference on Learning Theory*. Ed. by Shipra Agrawal and Aaron Roth. Vol. 247. Proceedings of Machine Learning Research. PMLR, pp. 4511–4547. URL: https://proceedings.mlr.press/v247/samsonov24a.html.

Sason, Igal and Sergio Verdú (2016). "f -Divergence Inequalities". In: *IEEE Transactions on Information Theory* 62.11, pp. 5973–6006. DOI: 10.1109/TIT.2016.2603151.

Savas, Yagiz et al. (2020). "Entropy Maximization for Markov Decision Processes Under Temporal Logic Constraints". In: *IEEE Transactions on Automatic Control* 65.4, pp. 1552–1567. DOI: 10.1109/TAC.2019.2922583.

Scheid, Antoine et al. (July 2024). "Incentivized Learning in Principal-Agent Bandit Games". In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov et al. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 43608–43631. URL: https://proceedings.mlr.press/v235/scheid24a.html.

Schulman, John, Xi Chen, and Pieter Abbeel (2018). *Equivalence Between Policy Gradients and Soft Q-Learning*. arXiv: 1704.06440 [cs.LG]. URL: https://arxiv.org/abs/1704.06440.

Schulman, John et al. (July 2015). "Trust Region Policy Optimization". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 1889–1897. URL: https://proceedings.mlr.press/v37/schulman15.html.

Schulman, John et al. (2017). *Proximal Policy Optimization Algorithms*. arXiv: 1707.06347 [cs.LG]. URL: https://arxiv.org/abs/1707.06347.

Seo, Younggyo et al. (July 2021). "State Entropy Maximization with Random Encoders for Efficient Exploration". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 9443–9454. URL: https://proceedings.mlr.press/v139/seo21a.html.

Shani, Lior et al. (July 2020). "Optimistic Policy Optimization with Bandit Feedback". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 8604–8613. URL: https://proceedings.mlr.press/v119/shani20a.html.

Sherman, U. et al. (2023). *Rate-optimal policy optimization for linear Markov decision processes*. Preprint, arXiv:2308.14642.

Shi, Laixi et al. (July 2022). "Pessimistic Q-Learning for Offline Reinforcement Learning: Towards Optimal Sample Complexity". In: *Proceedings of the 39th International Conference on Machine*

Learning. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 19967–20025. URL: https://proceedings.mlr.press/v162/shi22c.html.

Silver, David et al. (2016). "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529.7587, pp. 484–489.

Silver, David et al. (2018). "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". In: *Science* 362 (6419).

Silver, David et al. (2021). "Reward is enough". In: *Artificial intelligence* 299, p. 103535.

Sinclair, Sean R, Siddhartha Banerjee, and Christina Lee Yu (2023). "Adaptive discretization in online reinforcement learning". In: *Operations Research* 71.5, pp. 1636–1652.

Sion, Maurice (1958). "On general minimax theorems." In: *Pacific Journal of Mathematics* 8.1, pp. 171–176. DOI: pjm/1103040253. URL: https://doi.org/.

Stiennon, Nisan et al. (2020). "Learning to summarize with human feedback". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 3008–3021. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

Sutton, Richard S, Andrew G Barto, et al. (1998). *Reinforcement learning: An introduction*. Vol. 1. 1. MIT press Cambridge.

Sutton, Richard S. (1990). "Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming". In: *Machine Learning Proceedings 1990*. Ed. by Bruce Porter and Raymond Mooney. San Francisco (CA): Morgan Kaufmann, pp. 216–224. ISBN: 978-1-55860-141-3. DOI: https://doi.org/10.1016/B978-1-55860-141-3.50030-4. URL: https://www.sciencedirect.com/science/article/pii/B9781558601413500304.

Talebi, Mohammad Sadegh and Odalric-Ambrym Maillard (Apr. 2018). "Variance-Aware Regret Bounds for Undiscounted Reinforcement Learning in MDPs". In: *Proceedings of Algorithmic Learning Theory*. Ed. by Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan. Vol. 83. Proceedings of Machine Learning Research. PMLR, pp. 770–805. URL: https://proceedings.mlr.press/v83/talebi18a.html.

Tang, Chen et al. (Apr. 2025). "Deep Reinforcement Learning for Robotics: A Survey of Real-World Successes". In: vol. 39. 27, pp. 28694–28698. DOI: 10.1609/aaai.v39i27.35095. URL: https://ojs.aaai.org/index.php/AAAI/article/view/35095.

Tarbouriech, Jean et al. (Aug. 2020a). "Active Model Estimation in Markov Decision Processes". In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*. Ed. by Jonas Peters and David Sontag. Vol. 124. Proceedings of Machine Learning Research. PMLR, pp. 1019–1028. URL: https://proceedings.mlr.press/v124/tarbouriech20a.html.

Tarbouriech, Jean et al. (2020b). "Improved Sample Complexity for Incremental Autonomous Exploration in MDPs". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 11273–11284. URL: https://proceedings.neurips.cc/paper/2020/file/81e793dc8317a3dbc3534ed3f242c418-Paper.pdf.

— (2021). "A Provably Efficient Sample Collection Strategy for Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 7611–7624. URL: https://proceedings.neurips.cc/paper/2021/file/3e98410c45ea98addec555019bbae8eb-Paper.pdf.

Taupin, Jérôme, Yassir Jedra, and Alexandre Proutiere (2023). "Best Policy Identification in Discounted Linear MDPs". In: *Sixteenth European Workshop on Reinforcement Learning*. URL: https://openreview.net/forum?id=SOCOgATRQiY.

Tiapkin, Daniil, Evgenii Chzhen, and Gilles Stoltz (May 2025a). "Narrowing the Gap between Adversarial and Stochastic MDPs via Policy Optimization". In: *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*. Ed. by Yingzhen Li et al. Vol. 258.

Proceedings of Machine Learning Research. PMLR, pp. 3331–3339. URL: https://proceedings.mlr.press/v258/tiapkin25a.html.

Tiapkin, Daniil et al. (July 2022a). "From Dirichlet to Rubin: Optimistic Exploration in RL without Bonuses". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 21380–21431. URL: https://proceedings.mlr.press/v162/tiapkin22a.html.

Tiapkin, Daniil et al. (2022b). "Optimistic Posterior Sampling for Reinforcement Learning with Few Samples and Tight Guarantees". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., pp. 10737–10751. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/45e15bae91a6f213d45e203b8a29be48-Paper-Conference.pdf.

Tiapkin, Daniil et al. (July 2023a). "Fast Rates for Maximum Entropy Exploration". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 34161–34221. URL: https://proceedings.mlr.press/v202/tiapkin23a.html.

Tiapkin, Daniil et al. (2023b). "Model-free Posterior Sampling via Learning Rate Randomization". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., pp. 73719–73774. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/e985dfca10e1167c0836a70880ef0858-Paper-Conference.pdf.

Tiapkin, Daniil et al. (2024a). "Demonstration-Regularized RL". In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=lF2aip4Scn.

Tiapkin, Daniil et al. (May 2024b). "Generative Flow Networks as Entropy-Regularized RL". In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. Ed. by Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li. Vol. 238. Proceedings of Machine Learning Research. PMLR, pp. 4213–4221. URL: https://proceedings.mlr.press/v238/tiapkin24a.html.

Tiapkin, Daniil et al. (2025b). "Accelerating Nash Learning from Human Feedback via Mirror Prox". In: *arXiv preprint arXiv:2505.19731*.

Tirinzoni, Andrea, Aymen Al-Marjani, and Emilie Kaufmann (Feb. 2023). "Optimistic PAC Reinforcement Learning: the Instance-Dependent View". In: *Proceedings of The 34th International Conference on Algorithmic Learning Theory*. Ed. by Shipra Agrawal and Francesco Orabona. Vol. 201. Proceedings of Machine Learning Research. PMLR, pp. 1460–1480. URL: https://proceedings.mlr.press/v201/tirinzoni23a.html.

Tsybakov, A.B. (2008). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York. ISBN: 9780387790527. URL: https://doi.org/10.1007/b13794.

Vecerik, Mel et al. (2018). *Leveraging Demonstrations for Deep Reinforcement Learning on Robotics Problems with Sparse Rewards*. arXiv: 1707.08817 [cs.AI]. URL: https://arxiv.org/abs/1707.08817.

Vershynin, Roman (2018). *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press.

Vieillard, Nino et al. (2020). "Leverage the Average: An Analysis of KL Regularization in Reinforcement Learning". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9781713829546.

Vinyals, Oriol et al. (2019). "Grandmaster level in StarCraft II using multi-agent reinforcement learning". In: *Nature* 575.7782, pp. 350–354.

Wang, Yuanhao, Qinghua Liu, and Chi Jin (2023). "Is RLHF More Difficult than Standard RL? A Theoretical Perspective". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh

et al. Vol. 36. Curran Associates, Inc., pp. 76006–76032. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/efb9629755e598c4f261c44aeb6fde5e-Paper-Conference.pdf.

Watkins, Christopher JCH and Peter Dayan (1992). "Q-learning". In: *Machine learning* 8, pp. 279–292.

Wirth, Christian et al. (2017). "A Survey of Preference-Based Reinforcement Learning Methods". In: *Journal of Machine Learning Research* 18.136, pp. 1–46. URL: http://jmlr.org/papers/v18/16-634.html.

Xie, Tengyang et al. (2021). "Policy Finetuning: Bridging Sample-Efficient Offline and Online Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 27395–27407. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/e61eaa38aed621dd776d0e67cfeee366-Paper.pdf.

Xu, Yichong et al. (2020). "Preference-based Reinforcement Learning with Finite-Time Guarantees". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 18784–18794. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/d9d3837ee7981e8c064774da6cdd98bf-Paper.pdf.

Yang, Qisong and Matthijs TJ Spaan (2023). "CEM: Constrained Entropy Maximization for Task-Agnostic Safe Exploration". In: *The Thirty-Seventh AAAI Conference on Artificial Intelligence*.

Yang, Yuhong and A.R. Barron (1998). "An asymptotic property of model selection criteria". In: *IEEE Transactions on Information Theory* 44.1, pp. 95–116. DOI: 10.1109/18.650993.

Yardim, Batuhan, Semih Cayci, and Niao He (July 2025). "A Variational Inequality Approach to Independent Learning in Static Mean-Field Games". In: *ACM / IMS J. Data Sci.* 2.2. DOI: 10.1145/3731553. URL: https://doi.org/10.1145/3731553.

Yin, Ming, Yu Bai, and Yu-Xiang Wang (2021). "Near-Optimal Offline Reinforcement Learning via Double Variance Reduction". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 7677–7688. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/3f24bb08a5741e4197af64e1f93a5029-Paper.pdf.

Yu, Jia Yuan, Shie Mannor, and Nahum Shimkin (2009). "Markov Decision Processes with Arbitrary Reward Processes". In: *Mathematics of Operations Research* 34.3, pp. 737–757. DOI: 10.1287/moor.1090.0397. eprint: https://doi.org/10.1287/moor.1090.0397. URL: https://doi.org/10.1287/moor.1090.0397.

Zahavy, Tom et al. (2021). "Reward is enough for convex MDPs". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 25746–25759. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/d7e4cdde82a894b8f633e6d61a01ef15-Paper.pdf.

Zanette, Andrea and Emma Brunskill (June 2019). "Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 7304–7312. URL: https://proceedings.mlr.press/v97/zanette19a.html.

Zhan, Wenhao et al. (2023). *Provable Offline Preference-Based Reinforcement Learning*. arXiv: 2305.14816 [cs.LG]. URL: https://arxiv.org/abs/2305.14816.

Zhan, Wenhao et al. (2024). "Provable Reward-Agnostic Preference-Based Reinforcement Learning". In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=yTBXeXdbMf.

Zhang, Chuheng et al. (May 2021a). "Exploration by Maximizing Renyi Entropy for Reward-Free RL Framework". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.12, pp. 10859–10867. DOI: 10.1609/aaai.v35i12.17297. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17297.

Zhang, Lijun, Shiyin Lu, and Zhi-Hua Zhou (2018). "Adaptive Online Learning in Dynamic Environments". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/10a5ab2db37feedfdeaab192ead4ac0e-Paper.pdf.

Zhang, Tong (2002). "Covering number bounds of certain regularized linear function classes". In: *Journal of Machine Learning Research* 2.Mar, pp. 527–550.

— (2006). "From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation". In: *The Annals of Statistics* 34.5, pp. 2180–2210. DOI: 10.1214/009053606000000704. URL: https://doi.org/10.1214/009053606000000704.

Zhang, Zihan, Xiangyang Ji, and Simon Du (Aug. 2021b). "Is Reinforcement Learning More Difficult Than Bandits? A Near-optimal Algorithm Escaping the Curse of Horizon". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, pp. 4528–4531. URL: https://proceedings.mlr.press/v134/zhang21b.html.

Zhang, Zihan, Yuan Zhou, and Xiangyang Ji (2020). "Almost Optimal Model-Free Reinforcement Learningvia Reference-Advantage Decomposition". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 15198–15207. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/ad71c82b22f4f65b9398f76d8be4c615-Paper.pdf.

Zhang, Zihan et al. (June 2024). "Settling the sample complexity of online reinforcement learning". In: *Proceedings of Thirty Seventh Conference on Learning Theory*. Ed. by Shipra Agrawal and Aaron Roth. Vol. 247. Proceedings of Machine Learning Research. PMLR, pp. 5213–5219. URL: https://proceedings.mlr.press/v247/zhang24a.html.

Zhao, Y. et al. (2011). "Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer". In: *Biometrics* 67.4, pp. 1422–1433.

Zheng, Yan et al. (2021). "Automatic Web Testing Using Curiosity-Driven Reinforcement Learning". In: *Proceedings of the 43rd International Conference on Software Engineering*. ICSE '21. Madrid, Spain: IEEE Press, pp. 423–435. ISBN: 9781450390859. DOI: 10.1109/ICSE43902.2021.00048. URL: https://doi.org/10.1109/ICSE43902.2021.00048.

Zhu, Banghua, Michael Jordan, and Jiantao Jiao (July 2023). "Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 43037–43067. URL: https://proceedings.mlr.press/v202/zhu23f.html.

Zhu, Yuke et al. (June 2018). "Reinforcement and Imitation Learning for Diverse Visuomotor Skills". In: *Proceedings of Robotics: Science and Systems*. Pittsburgh, Pennsylvania. DOI: 10.15607/RSS.2018.XIV.009.

Zhuang, Vincent and Yanan Sui (Apr. 2021). "No-Regret Reinforcement Learning with Heavy-Tailed Rewards". In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, pp. 3385–3393. URL: https://proceedings.mlr.press/v130/zhuang21a.html.

Ziegler, Daniel M. et al. (2020). *Fine-Tuning Language Models from Human Preferences*. arXiv: 1909.08593 [cs.CL]. URL: https://arxiv.org/abs/1909.08593.

Zimin, Alexander and Gergely Neu (2013). "Online learning in episodic Markovian decision processes by relative entropy policy search". In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre: Apprentissage par renforcement efficace en taille d'échantillon : exploration, imitation et apprentissage séquentiel

Mots clés: apprentissage par renforcement, exploration, apprentissage par imitation, MDPs adversaires, apprentissage par renforcement à partir du retour humain

Résumé: L'apprentissage par renforcement (RL) fournit un cadre formel pour la prise de décision séquentielle en environnement inconnu. Un agent apprend une politique afin de maximiser un signal de récompense cumulée en interagissant avec un processus de décision markovien (MDP). Cette thèse étudie trois extensions du cadre standard de l'apprentissage par renforcement, en proposant de nouveaux algorithmes ainsi que leurs analyses théoriques associées. Premièrement, nous étudions le problème de l'exploration en l'absence de signal de récompense. Nous considérons deux formulations distinctes de l'exploration par maximisation d'entropie et proposons des algorithmes computationnellement efficaces. Pour ces algorithmes, nous dérivons des bornes de complexité en tailles d'échantillons, qui améliorent les résultats précédemment connus. Notre analyse établit en outre que l'apprentissage de la politique de maximisation de l'entropie de visite est statistiquement moins exigeant que l'exploration sans récompense dans l'une de ses variantes usuelles, ainsi que l'apprentissage d'une politique optimale dans un PDM standard avec récompenses. Deuxièmement, nous abordons le problème de l'accélération de l'apprentissage par renforcement à l'aide de démonstrations expertes, un cadre motivé par des applications modernes telles que l'apprentissage par renforcement à partir de retours (*feedback*) humains (RLHF). Nous analysons le cadre de l'apprentissage par renforcement régularisé par des démonstrations, où la fonction objectif de l'agent est augmentée d'une pénalité de Kullback-Leibler pour toute déviation par rapport à une politique de référence apprise à partir de démonstrations. Nous prouvons que cette approche entraîne une réduction linéaire de la complexité en tailles d'échantillon par rapport au nombre de démonstrations. Ce résultat établit comment les données expertes peuvent améliorer l'efficacité de la tâche d'apprentissage par renforcement. Troisièmement, nous étudions le problème de l'apprentissage dans des MDP où les récompenses sont choisies par un adversaire oublieux (*oblivious*). Nous proposons un nouvel algorithme qui utilise un oracle boîte noire d'optimisation linéaire séquentielle, le rendant facile à mettre en œuvre. Nous fournissons une analyse de regret montrant que sa borne de performance correspond à la borne inférieure minimax du cas stochastique (non adverse) en termes de dépendance aux nombres d'états, d'actions et d'épisodes. Globalement, les contributions de cette thèse consistent en de nouveaux algorithmes pour divers problèmes en apprentissage par renforcement, accompagnés de preuves de leur performance en termes de complexité en tailles d'échantillons et de bornes de regret.

Title: Sample-Efficient Reinforcement Learning: Exploration, Imitation, and Online Learning

Keywords: reinforcement learning, exploration, imitation learning, adversarial MDPs, reinforcement learning from human feedback

Abstract: Reinforcement learning (RL) provides a formal framework for sequential decision-making in an unknown environment. An agent learns a policy to maximize a cumulative reward signal by interacting with a Markov Decision Process (MDP). This thesis investigates three extensions to the standard RL setting, providing new algorithms and associated theoretical analyses. First, we study the problem of exploration in the absence of a reward signal. We consider two distinct formulations of maximum entropy exploration and propose computationally efficient algorithms. For these algorithms, we derive sample complexity bounds that improve upon previously known results. Our analysis further establishes that learning the maximum visitation entropy policy is statistically less demanding than both a common variant of reward-free exploration and the task of learning an optimal policy in a standard rewarded MDP. Second, we address the problem of accelerating reinforcement learning using expert demonstrations, a setting motivated by modern applications such as Reinforcement Learning from Human Feedback (RLHF). We analyze the framework of demonstration-regularized RL, where the agent's learning objective is augmented with a Kullback-Leibler penalty for deviating from a reference policy learned from the demonstrations. We prove that this approach leads to a linear reduction in sample complexity with respect to the number of demonstrations. This result establishes how expert data can improve the efficiency of the subsequent RL task. Third, we address the problem of learning in MDPs with rewards selected by an (oblivious) adversary. We propose a new algorithm that uses a black-box online linear optimization oracle, making it easy to implement. We provide a regret analysis showing that its performance bound matches the minimax lower bound for the stochastic (non-adversarial) case in its dependence on the number of states, actions, and the number of episodes. Collectively, the contributions of this thesis consist of novel algorithms for various problems in reinforcement learning, accompanied by proofs of their performance in terms of sample complexity and regret bounds.

Institut Polytechnique de Paris
91120 Palaiseau, France