UNIVERSITY OF CALIFORNIA

Los Angeles

Efficient Reinforcement Learning through Uncertainties

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Computer Science

by

Dongruo Zhou

2023

ABSTRACT OF THE DISSERTATION

Efficient Reinforcement Learning through Uncertainties

by

Dongruo Zhou
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2023
Professor Quanquan Gu, Chair

This dissertation is centered around the concept of uncertainty-aware reinforcement learning (RL), which seeks to enhance the efficiency of RL by incorporating uncertainty. RL is a vital mathematical framework in the field of artificial intelligence (AI) for creating autonomous agents that can learn optimal behaviors through interaction with their environments. However, RL is often criticized for being sample inefficient and computationally demanding. To tackle these challenges, the primary goals of this dissertation are twofold: to offer theoretical understanding of uncertainty-aware RL and to develop practical algorithms that utilize uncertainty to enhance the efficiency of RL.

Our first objective is to develop an RL approach that is efficient in terms of sample usage for Markov Decision Processes (MDPs) with large state and action spaces. We present an uncertainty-aware RL algorithm that incorporates function approximation. We provide theoretical proof that this algorithm achieves near minimax optimal statistical complexity when learning the optimal policy. In our second objective, we address two specific scenarios: the batch learning setting and the rare policy switch setting. For both settings, we propose uncertainty-aware RL algorithms with limited adaptivity. These algorithms

significantly reduce the number of policy switches compared to previous baseline algorithms while maintaining a similar level of statistical complexity. Lastly, we focus on estimating uncertainties in neural network-based estimation models. We introduce a gradient-based method that effectively computes these uncertainties. Our approach is computationally efficient, and the resulting uncertainty estimates are both valid and reliable.

The methods and techniques presented in this dissertation contribute to the advancement of our understanding regarding the fundamental limits of RL. These research findings pave the way for further exploration and development in the field of decision-making algorithm design.

The dissertation of Dongruo Zhou is approved.

Adnan Darwiche

Baharan Mirzasoleiman

Csaba Szepesvári

Lieven Vandenberghe

Quanquan Gu, Committee Chair

University of California, Los Angeles

2023

To my beloved ones.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to my advisor, Quanquan Gu, for several reasons. Firstly, I am immensely grateful to him for providing me with the opportunity to join his esteemed research group, despite my lack of prior experience in machine learning research and an exceptional academic record when I was applying for graduate schools. His generosity in giving me this chance has been instrumental in shaping my journey as a researcher. Moreover, I deeply appreciate his unwavering belief in my potential to grow into a proficient machine learning researcher right from the beginning of my PhD studies. His faith in me has been a constant source of encouragement and motivation. Secondly, I would like to acknowledge and thank him for his candid suggestions regarding research, career paths, and life in general, from which I have gained a clear understanding of the qualities and attributes that define an outstanding and successful professor. Lastly, I am indebted to him for the diligent training he has provided me in enhancing my presentation and communication skills. The emphasis he placed on these essential aspects has made me realize their profound significance not only in academia but also in my future job pursuits. Thank you, Quanquan!

I would like to express my sincere appreciation to the members of my doctoral committee: Adnan Darwiche, Baharan Mirzasoleiman, Csaba Szepesvári, and Lieven Vandenberghe, for their invaluable feedback and suggestions regarding my research topic and presentation style. I extend special thanks to Csaba Szepesvári for his generous sharing of knowledge and understanding in the field of reinforcement learning during our discussions, which leads to one chapter of my dissertation. Additionally, I would like to express my gratitude to Jing Huang and Peng Qi for introducing me to the fascinating field of Natural Language Processing and providing me with an invaluable opportunity to intern at JD.com during the summer of 2021. I would also like to extend my thanks to Aparna Pandey and Fangzhou Cheng for hosting me as an intern at AWS AI during the summer of 2022.

Throughout the past years, I have had the privilege of collaborating with an exceptional

group of individuals in Quanquan's research team. I am immensely grateful for the opportunity to work alongside the following members: Yuan Cao, Jinghui Chen, Yuanzhou Chen, Zixiang Chen, Yihe Deng, Qiwei Di, Spencer Frei, Jiafan He, Yiwen Kou, Xuheng Li, Lu Lin, Lu Tian, Lingxiao Wang, Yue Wu, Pan Xu, Yaodong Yu, Junkai Zhang, Weitong Zhang, Xiao Zhang, Heyang Zhao, and Difan Zou. In particular, I would like to extend my thanks to Yuan Cao, Jinghui Chen, Zixiang Chen, Qiwei Di, Jiafan He, Yue Wu, Pan Xu, Weitong Zhang, Heyang Zhao, and Difan Zou for their exceptional contributions to our collaborated projects. A special thank-you to Weitong for his many helps during these years, including taking several long walks around UCLA with me, which helped me cope with the challenges of the COVID pandemic and the associated lockdown period.

Meanwhile, I would like to extend my gratitude to my collaborators outside of UCLA: Jiahao Chen, Jeffrey L. Chen, Jingdong Gao, Yiling Jia, Ruoxi Jiang, Michael I. Jordan, Amin Karbasi, Khashayar Khosravi, Chris Junchi Li, Lihong Li, Yifei Min, Vahab Mirrokni, Robert Nowak, Junhong Shen, Csaba Szepesvári, Yiqi Tang, Hongning Wang, Tianhao Wang, Rebecca Willett, Ying Nian Wu, Ziyan Yang, Jinfeng Yi, Luyao Yuan, Amy Zhang, Tong Zhang, Song-Chun Zhu, and Yinglun Zhu. Their collaboration and support have been instrumental in achieving high standards in our research endeavors.

Moreover, I would like to express my appreciation to a group of friends who were not mentioned earlier. I want to thank Jingling for her integrity and inspiring outlook on life, which have helped me overcome the most challenging periods during my Ph.D. studies. To Xuelu, Minhao, Zijun, Ziniu, Xuanqing, Yewen, Ming, and Yiqi, I am grateful for their valuable hints and advice during my job hunting process. To Pengyu, I want to thank him for his opinions and suggestions, which have been influential in various aspects since our undergraduate days. I would also like to thank Junheng for his recommendations on interesting things to explore in Westwood, like exceptional beers at Broxton.

Finally, I would like to extend my deepest gratitude to my parents, Ruojuan and Xiangjian, for instilling in me the invaluable qualities of critical thinking and limitless imagination. I also

extend a special appreciation to Lily, whose presence in my life has brought my Ph.D. journey to a magnificent end. Your presence reignites my passion to pursue all the possibilities that lie ahead and makes every challenge I have encountered in the past meaningful. I love you all.

VITA

2013-2017 Bachelor of Science in Pure and Applied Mathematics, Tsinghua University

2017-2018 Teaching Assistant, Department of System and Information Engineering, University of Virginia

2018–2023 Research Assistant, Computer Science Department, University of California, Los Angeles

2020–2021 Teaching Assistant, Computer Science Department, University of California, Los Angeles

PUBLICATIONS

*We select publications that are the most relevant to the topic of this dissertation. * indicates equal contribution.*

Dongruo Zhou, Lihong Li and Quanquan Gu. Neural Contextual Bandits with UCB-based Exploration. International Conference on Machine Learning (ICML), 2020.

Dongrou Zhou, Quanquan Gu and Csaba Szepesvári . Nearly Minimax Optimal Reinforcement Learning for Linear Mixture Markov Decision Processes. Conference of Learning Theory (COLT), 2021.

Dongruo Zhou, Jiafan He and Quanquan Gu. Provably Efficient Reinforcement Learning for Discounted MDPs with Feature Mapping. International Conference on Machine Learning

(ICML), 2021.

Tianhao Wang*, **Dongruo Zhou*** and Quanquan Gu. Provably Efficient Reinforcement Learning with Linear Function Approximation under Adaptivity Constraints. Advances in Neural Information Processing Systems (NeurIPS), 2021.

CHAPTER 1

Introduction

The field of artificial intelligence (AI) aims to develop autonomous agents that can learn optimal behaviors by interacting with their environments. Reinforcement learning (RL) is a mathematical framework that plays a crucial role in achieving state-of-the-art performance in various applications to develop intelligent agents, such as AlphaGo (Silver et al., 2016) for the game of Go and ChatGPT (OpenAI, 2023) for conversational systems. Despite its empirical success, RL is often considered inefficient for two reasons. Firstly, it requires a large number of samples to train modern RL models, which affects their statistical performance. Secondly, the learned behavior of the agent needs to be frequently updated to align with its most recent experiences, leading to computationally expensive model updates. To address these challenges, this dissertation focuses on enhancing the efficiency of RL through the incorporation of uncertainty, also known as uncertainty-aware RL. Uncertainty refers to the factors that cause the agent's inference about the unknown environment to deviate from the true environment. The main objectives of this dissertation are to provide theoretical insights into uncertainty-aware RL and develop practical algorithms that leverage uncertainty to improve the efficiency of RL.

In the initial stage, our objective is to enhance the efficiency of reinforcement learning (RL) by reducing the number of samples required. When the number of states and actions is finite, it is referred to as "tabular RL." Over the past decade, significant progress has been made in understanding the limitations of sample efficiency in RL. Breakthroughs have led to the development of algorithms that approach the minimax optimal sample complexity

for planning scenarios, assuming the availability of a generative model (Azar et al., 2013; Sidford et al., 2018; Agarwal et al., 2020). Further advancements have extended these nearly minimax optimal algorithms to the more challenging online learning setting, encompassing various objectives. These settings include episodic Markov Decision Processes (MDPs) (Azar et al., 2017; Zanette and Brunskill, 2019; Zhang et al., 2020), through discounted MDPs (Lattimore and Hutter, 2012; Zhang et al., 2021b; He et al., 2021b) to infinite horizon MDPs with the average reward criterion (Zhang and Ji, 2019; Tossou et al., 2019). In **Chapter** 2, we address a more general case where tabular RL is not suitable due to the large state and action spaces involved. To handle such large MDPs, a classical approach is to assume the availability of function approximation techniques that can compactly represent policies or value functions (Sutton and Barto, 1998). We propose an uncertainty-aware RL algorithm with function approximation, specifically designed for linear mixture MDPs (Jia et al., 2020; Ayoub et al., 2020). Our algorithm is proven to achieve nearly minimax optimal statistical complexity when learning the optimal policy.

Furthermore, our objective includes enhancing the computational efficiency of RL. In practical applications, it is often impractical to frequently change the policy due to factors such as large amounts of data, limited computing resources, and the associated costs of switching. Consequently, it becomes desirable to batch the data stream and update the policy at the end of each period or batch. To illustrate, in the context of clinical trials, each phase or batch of the trial involves administering a medical treatment to a group of patients simultaneously. The outcomes of the treatment are not observed until the phase concludes, and they are subsequently used to design experiments for the next phase. The selection of an appropriate number and size of batches becomes crucial for achieving nearly optimal efficiency in the clinical trial setting. This gives rise to the limited adaptivity setting, which has been extensively studied in many online learning problems including prediction-from-experts (PFE) (Kalai and Vempala, 2005; Cesa-Bianchi et al., 2013), multi-armed bandits (MAB) (Arora et al., 2012; Cesa-Bianchi et al., 2013) and online convex optimization (Jaghargh et al., 2019;

Chen et al., 2020), to mention a few. In **Chapter** 3, we introduce algorithms with limited adaptivity for two distinct settings: the batch learning setting and the rare policy switch setting. Our algorithms demonstrate a significant reduction in the number of policy switches while maintaining a comparable level of statistical complexity to previous baseline algorithms.

Lastly, we put forth practical techniques for efficiently estimating uncertainties in a general model class. To illustrate our approach, we focus on the stochastic contextual bandit problem, which has garnered extensive attention in the field of machine learning (Langford and Zhang, 2008; Bubeck and Cesa-Bianchi, 2012; Lattimore and Szepesvári, 2020) and is often considered a simplified model of more complex RL problems. Contextual bandit algorithms have found applications in various real-world scenarios, including personalized recommendation systems, advertising, and web search. The linear contextual bandit model (Auer, 2002; Abe et al., 2003; Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010) has received significant attention in the literature. This model assumes that the expected reward at each round is linearly related to the feature vector. While linear contextual bandits have shown success in both theory and practice (Li et al., 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011), they often fail to capture the true reward structure in practical applications. This motivates the exploration of nonlinear or nonparametric contextual bandit models (Filippi et al., 2010; Srinivas et al., 2010; Bubeck et al., 2011; Valko et al., 2013), which relax the restrictive assumptions imposed on the reward function. However, these nonlinear models still necessitate certain assumptions. For instance, Filippi et al. (2010) make a generalized linear model assumption on the reward, Bubeck et al. (2011) require it to have a Lipschitz continuous property in a proper metric space, and Valko et al. (2013) assume the reward function belongs to some Reproducing Kernel Hilbert Space (RKHS). In **Chapter** 4, we consider the general reward model which can be approximated by a neural network. We propose a gradient-based method for estimating uncertainties in neural network-based estimation models. Our approach exhibits computational efficiency, and the resulting uncertainty estimates are both valid and effective, as demonstrated through experiments on both simulated and real-world datasets.

1.1 Organization of the Dissertation

In **Chapter** 2, we propose an uncertainty-aware RL algorithm for linear mixture MDPs (Jia et al., 2020; Ayoub et al., 2020), a special class of MDPs where the transition dynamic can be represented as a linear combination of several basis transition dynamics. Our proposed algorithm improves existing baseline algorithms (Jia et al., 2020; Ayoub et al., 2020) by utilizing the uncertainty-based weighted regression to learn the transition dynamic. Furthermore, we prove that such an improvement is already minimax optimal by constructing a hard linear mixture MDP example. In **Chapter** 3, we study linear MDPs (Yang and Wang, 2019; Jin et al., 2020) and propose algorithms with limited adaptivity for two settings named by the batch learning setting and the rare policy switch setting. We show that our proposed algorithms enjoy smaller number of policy switches while maintaining the same order of statistical complexity, compared with previous baseline algorithms. We also prove that our proposed algorithms are nearly optimal with respect to the number of policy switches by constructing special linear MDP instances. In **Chapter** 4, we propose a gradient-based uncertainty estimate for neural network-based reward model, applying on the contextual bandit problem. We show that our method is computational efficient, the constructed uncertainty estimate is valid and effective, tested by both simulated and real-world datasets. We summarize this dissertation and highlight several future research directions in **Chapter** 5.

1.2 Notations and Basic Definitions

Notations We use lower case letters to denote scalars, and use lower and upper case bold face letters to denote vectors and matrices respectively. We denote by $[n]$ the set $\{1, \ldots, n\}$. For a vector $\mathbf{x} \in \mathbb{R}^d$ and matrix $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$, a positive semi-definite matrix, we denote by $\|\mathbf{x}\|_2$ the vector's Euclidean norm and define $\|\mathbf{x}\|_{\mathbf{\Sigma}} = \sqrt{\mathbf{x}^\top \mathbf{\Sigma} \mathbf{x}}$. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, let $\mathbf{x} \odot \mathbf{y}$ be the Hadamard (componentwise) product of $\mathbf{x}$ and $\mathbf{y}$. For two positive sequences $\{a_n\}$ and $\{b_n\}$

with $n = 1, 2, \ldots$, we write $a_n = O(b_n)$ if there exists an absolute constant $C > 0$ such that $a_n \le C b_n$ holds for all $n \ge 1$ and write $a_n = \Omega(b_n)$ if there exists an absolute constant $C > 0$ such that $a_n \ge C b_n$ holds for all $n \ge 1$. We use $\widetilde{O}(\cdot)$ to further hide the polylogarithmic factors. We use $\mathbb{1}\{\cdot\}$ to denote the indicator function. For $a, b \in \mathbb{R}$ satisfying $a \le b$, we use $[x]_{[a,b]}$ to denote the function $x \cdot \mathbb{1}\{a \le x \le b\} + a \cdot \mathbb{1}\{x < a\} + b \cdot \mathbb{1}\{x > b\}$, which truncates its argument to the $[a, b]$ interval. We say a random variable X is ν-sub-Gaussian if $\mathbb{E} \exp(\lambda(X - \mathbb{E}X)) \le \exp(\lambda^2 \nu^2 / 2)$ for any $\lambda > 0$.

Inhomogeneous Episodic MDP We denote an inhomogeneous, episodic MDP by a tuple $M = (\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h=1}^H, \{\mathbb{P}_h\}_{h=1}^H)$, where $\mathcal{S}$ is the state space with $|\mathcal{S}| = S$, $\mathcal{A}$ is the action space with $|\mathcal{A}| = A$, H is the length of the episode, $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, either deterministic or stochastic, and $\mathbb{P}_h$ is the transition probability function at stage h so that for $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$, $\mathbb{P}_h(s'|s, a)$ is the transition probability (or called transition dynamic) of arriving at stage $h + 1$ at state s' provided that the state at stage h is s and action a is chosen at this stage. For the sake of simplicity, we restrict ourselves to countable state and finite action spaces. A policy $\pi = \{\pi_h\}_{h=1}^H$ is a collection of H functions, where each of them maps a state s to an action a. For $(s, a) \in \mathcal{S} \times \mathcal{A}$, we define the action-values $Q_h^\pi(s, a)$ and (state) values $V_h^\pi(s)$ as follows:

$$Q_h^\pi(s, a) = \mathbb{E}_{\pi,h,s,a}\left[\sum_{h'=h}^{H} r_h(s_{h'}, a_{h'})\right],\ V_h^\pi(s) = Q_h^\pi(s, \pi_h(s)),\ V_{H+1}^\pi(s) = 0.$$

In the definition of Q_h^π, $\mathbb{E}_{\pi,h,s,a}$ means an expectation over the probability measure over state-action pairs of length $H - h + 1$ that is induced by the interconnection of policy π and the MDP M when initializing the process to start at stage h with the pair (s, a). In particular, the probability of sequence $(s_h, a_h, s_{h+1}, a_{h+1}, \ldots, s_H, a_H)$ under this sequence is $\mathbf{1}(s_h = s)\mathbf{1}(a_h = a)\mathbb{P}_h(s_{h+1}|s_h, a_h)\mathbf{1}_{\pi_{h+1}(s_{h+1})=a_{h+1}} \ldots \mathbb{P}_{H-1}(s_H|s_{H-1}, a_{H-1})\mathbf{1}_{\pi_H(s_H)=a_H}$. The optimal value function $V_h^*(\cdot)$ and the optimal action-value function $Q_h^*(\cdot, \cdot)$ are defined by $V_h^*(s) = \sup_\pi V_h^\pi(s)$ and $Q_h^*(s, a) = \sup_\pi Q_h^\pi(s, a)$, respectively. For any function $V : \mathcal{S} \to \mathbb{R}$,

we introduce the shorthands

$$[\mathbb{P}_h V](s,a) = \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} V(s'),\ [\mathbb{V}_h V](s,a) = [\mathbb{P}_h V^2](s,a) - ([\mathbb{P}_h V](s,a))^2,$$

where V^2 stands for the function whose value at s is $V^2(s)$. Using this notation, the Bellman equations for policy π and the Bellman optimality equation can be written as

$$Q_h^\pi(s,a) = r_h(s,a) + [\mathbb{P}_h V_{h+1}^\pi](s,a),\ Q_h^*(s,a) = r_h(s,a) + [\mathbb{P}_h V_{h+1}^*](s,a).$$

Note that both hold *simultaneously* for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $h \in [H]$. In general, the goal for an agent is to utilize any RL algorithm to learn the optimal value function V_h^*, the optimal action-value function Q_h^*, or the corresponding optimal policy $\pi_h^* : \mathcal{S} \to \mathcal{A}$ such that $\pi_h^*(s) = \mathrm{argmax}_{a \in \mathcal{A}} Q_h^*(s,a)$.

Uncertainty in RL Our approach to enhancing the efficiency of reinforcement learning (RL) involves leveraging the concept of uncertainty in machine learning, as discussed in Hora (1996); Hüllermeier and Waegeman (2021). Uncertainty in RL can be categorized into two main types: *epistemic* (or systematic) uncertainty, which arises from a lack of knowledge about the optimal model, and *aleatoric* (or statistical) uncertainty, which pertains to inherent randomness and variability in experimental outcomes. In RL, the goal is to learn the optimal value function or policy using the Bellman optimality equations. However, since the agent does not have access to the true reward function and transition dynamics, they can only utilize estimated versions denoted as $\widehat{r}_h$ and $\widehat{\mathbb{P}}_h$, respectively, which inherently possess non-zero estimation errors ($r_h - \widehat{r}_h$ and $\mathbb{P}_h - \widehat{\mathbb{P}}_h$). These estimation errors represent epistemic uncertainty in RL, while the random nature of rewards (r_h) and transition dynamics ($\mathbb{P}_h$) represents aleatoric uncertainty. The specific definitions of epistemic and aleatoric uncertainty may vary depending on the problem setting.

CHAPTER 2

Sample-Efficient Reinforcement Learning through Uncertainties

2.1 Introduction

We aim to propose sample-efficient RL through the use of uncertainties in this chapter. Recently, there is a growing body of work in understanding the interplay between RL and function approximation. When a generative model is available, Yang and Wang (2019) proposed a computationally efficient, nearly minimax optimal RL algorithm that works with such linear function approximation for a special case when the learner has access to a polynomially sized set of "anchor state-action pairs". Lattimore et al. (2020) proposed an optimal-design based RL algorithm without the anchor state-action pairs assumption. However, for online RL where no generative model is accessible, as of today a gap between the upper bounds (Yang and Wang, 2020; Jin et al., 2020; Wang et al., 2020c; Modi et al., 2020; Zanette et al., 2020a,b; Jia et al., 2020; Ayoub et al., 2020) and the lower bounds (Du et al., 2019b; Zhou et al., 2021b) still exist, with or without the anchor state-action assumption.

In this chapter, we propose a new RL algorithm with the near-minimax-optimality [1] for the special class of linear mixture MDPs, where the transition probability kernel is a linear mixture of a number of basis kernels (Modi et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021b). In detail, our contributions are listed as follows.

[1] In this chapter, we say an algorithm is nearly minimax optimal if this algorithm attains a regret/sample complexity that matches the minimax lower bound up to logarithmic factors.

- We propose a Bernstein-type self-normalized concentration inequality for vector-valued martingales, which improves the dominating term of the analog inequality of Abbasi-Yadkori et al. (2011) from $R\sqrt{d}$ to $\sigma\sqrt{d} + R$, where R and σ^2 are the magnitude and the variance of the noise respectively, and d is the dimension of the vectors involved. Following ideas developed for the tabular case (e.g., Azar et al. 2013) we replace the conservative Hoeffding-type confidence bounds used in UCRL-VTR of Ayoub et al. (2020) with a Bernstein-type confidence bound that is based on a new, Bernstein-type variant of the standard self-normalized concentration inequality of Abbasi-Yadkori et al. (2011). Our concentration inequality is a non-trivial extension of the Bernstein inequality from the scalar case to the vector case.

- With the Bernstein-type tail inequality, we consider a linear bandit problem as a "warm-up" example, whose noise at round t is R-bounded and of σ_t^2-variance. Note that bandits can be seen as a special instance of episodic RL where the length of the episode equals one. We propose a new algorithm called Weighted OFUL, which adapts a new linear regression scheme called *weighted ridge regression*, where the weights depend on the variance σ_t^2. We prove that Weighted OFUL enjoys an $\widetilde{O}(R\sqrt{dT} + d\sqrt{\sum_{t=1}^{T}\sigma_t^2})$ regret, which strictly improves the regret $\widetilde{O}(Rd\sqrt{T})$ obtained for the OFUL algorithm by Abbasi-Yadkori et al. (2011).

- We further apply the new tail inequality to the design and analysis of online RL algorithms for the aforementioned linear mixture MDPs (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021b). In the episodic setting, we propose a new algorithm, UCRL-VTR$^+$, which can be seen as an extension of UCRL-VTR studied by Jia et al. (2020); Ayoub et al. (2020). We show that UCRL-VTR$^+$ attains an $\widetilde{O}(dH\sqrt{T} + \sqrt{dH^3}\sqrt{T} + d^2H^3 + d^3H^2)$ regret, where T is the number of interactions with the MDP and H is the episode length. We also prove a nearly matching lower bound $\Omega(dH\sqrt{T})$ on the regret. When $d \geq H$ and $T \geq d^4H^2 + d^3H^3$, our UCRL-VTR$^+$ algorithm achieves an $\widetilde{O}(dH\sqrt{T})$ regret, which

matches our proved lower bound. Thus, our results imply that our algorithm is minimax optimal up to logarithmic factors in the high-dimensional large-sample regime.

- The weights adapted by UCRL-VTR$^+$ depend on the *epistemic uncertainty* and *aleatoric uncertainty*. We propose valid estimates of these two types of uncertainties and show that their modifications can serve as components of the weights which help to reduce the sample complexity of the original UCRL-VTR algorithm.

To the best of our knowledge, ignoring logarithmic factors, our proposed UCRL-VTR$^+$ is the first minimax optimal online RL algorithm with linear function approximation using the common case of a constant-dimension feature mapping. UCRL-VTR$^+$ is also computationally efficient with an access to a sampling or an integration oracle.

2.2 Related Work

The purpose of this section is to review prior works that are most relevant to our contributions.

Linear Bandits Linear bandits can be seen as the simplest version of RL with linear function approximation, where the episode length (i.e., planning horizon) $H = 1$. There is a huge body of literature on linear bandit problems (Auer, 2002; Chu et al., 2011; Li et al., 2010, 2019; Dani et al., 2008; Abbasi-Yadkori et al., 2011). Most of the linear bandit algorithms can be divided into two categories: algorithms for k-armed linear bandits, and algorithms for infinite-armed linear bandits. For the k-armed case, Auer (2002) proposed a SupLinRel algorithm, which makes use of the eigenvalue decomposition and enjoys an $O(\log^{3/2}(kT)\sqrt{dT})$ regret[2] . Li et al. (2010); Chu et al. (2011) proposed a SupLinUCB algorithm using the regularized least-squares estimator, which enjoys the same regret guarantees. Li et al. (2019) proposed a VCL-SupLinUCB algorithm with a refined confidence set design which enjoys an improved $O(\sqrt{\log(T)\log(k)dT})$ regret, which matches the lower bound up to a logarithmic

[2]We omit the poly$(\log\log(kT))$ factors for the simplicity of comparison.

factor. For the infinite-armed case, Dani et al. (2008) proposed an algorithm with a confidence ball, which enjoys $O(d\sqrt{T\log^3 T})$ regret. Abbasi-Yadkori et al. (2011) improved the regret to $O(d\sqrt{T\log^2 T})$ with a new self-normalized concentration inequality for vector-valued martingales. Li et al. (2021) further improved the regret to $O(d\sqrt{T\log T})$, which matches the lower bound up to a logarithmic factor. However, previous works only focus on the case where the reward noise is sub-Gaussian. In this chapter, we show that if the reward noise is restricted to a smaller class of distributions with bounded magnitude and variance, a better regret bound can be obtained. The main motivation to consider this problem is that linear bandits with bounded reward and variance can be seen as a special RL with linear function approximation when the episode length $H = 1$. Thus, this result immediately sheds light on the challenges involved in achieving minimax optimal regret for general RL with linear function approximation.

Reinforcement Learning with Linear Function Approximation Recent years have witnessed a flurry of activity on RL with linear function approximation (e.g., Jiang et al., 2017; Yang and Wang, 2019, 2020; Jin et al., 2020; Wang et al., 2020c; Modi et al., 2020; Dann et al., 2018; Du et al., 2019b; Sun et al., 2019; Zanette et al., 2020a,b; Cai et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Weisz et al., 2021; Zhou et al., 2021b; He et al., 2021a). These results can be generally grouped into four categories based on their assumptions on the underlying MDP. The first category of work uses the low Bellman-rank assumption (Jiang et al., 2017) which assumes that the Bellman error "matrix" where "rows" are index by a test function and columns are indexed by a distribution generating function from the set of test functions assumes a low-rank factorization. Representative work includes Jiang et al. (2017); Dann et al. (2018); Sun et al. (2019). The second category of work considers the *linear MDP* assumption (Yang and Wang, 2019; Jin et al., 2020) which assumes taht both the transition probability function and reward function are parameterized as a linear function of a given feature mapping over state-action pairs. Representative work includes Yang and Wang (2019); Jin et al. (2020); Wang et al. (2020c); Du et al. (2019b); Zanette et al. (2020a);

Wang et al. (2020b); He et al. (2021a). The third category of work focuses on the low inherent Bellman error assumption (Zanette et al., 2020b), which assumes the Bellman backup can be parameterized as a linear function up to some misspecification error. Zanette et al. (2020b) proposed an ELEANOR algorithm with a regret $\widetilde{O}(\sum_{h=1}^{H} d_h\sqrt{K})$, where d_h is the dimension of the feature mapping at the h-th stage within the episodes and K is the number of episodes. They also proved a lower bound $\Omega(\sum_{h=1}^{H} d_h\sqrt{K})$ under the sub-Gaussian norm assumption of the rewards and transitions but only for the special case when $d_1 = \sum_{h=2}^{H} d_h$. It can be seen that in this special case, their upper bound matches their lower bound up to logarithmic factors, and thus their algorithm is statistically near optimal. However, in the general case when $d_1 = \cdots = d_H = d$, there still exists a gap of H between their upper and lower bounds. Furthermore, as noted by the authors, the ELEANOR algorithm is not computationally efficient. The last category considers linear mixture MDPs (a.k.a., linear kernel MDPs) (Modi et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021b), which assumes that the transition probability function is parameterized as a linear function of a given feature mapping over state-action-next-state triples. Representative work includes Yang and Wang (2020); Modi et al. (2020); Jia et al. (2020); Ayoub et al. (2020); Cai et al. (2020); Zhou et al. (2021b); He et al. (2021a) (of these, Yang and Wang (2020) considers a special case, but their results extend to the linear mixture case seamlessly). Our work also considers linear mixture MDPs.

Bernstein Bonuses for Tabular MDPs There is a series of work proposing algorithms with nearly minimax optimal sample complexity or regret for the tabular MDP under different settings, including average-reward, discounted, and episodic MDPs (Azar et al., 2013, 2017; Zanette and Brunskill, 2019; Zhang and Ji, 2019; Simchowitz and Jamieson, 2019; Zhang et al., 2020; He et al., 2021b; Zhang et al., 2021a). The key idea at the heart of these works is the usage of the law of total variance to obtain tighter bounds on the expected sum of the variances for the estimated value function. These works have designed tighter confidence sets or upper confidence bounds by replacing the Hoeffding-type exploration bonuses with

Bernstein-type exploration bonuses, and obtained more accurate estimates of the optimal value function, a technique pioneered by Lattimore and Hutter (2012). Our work shows how this idea extends to algorithms with linear function approximation. To the best of our knowledge, our work is the first work using Bernstein bonus and law of total variance to achieve nearly minimax optimal regret for RL with linear function approximation.

2.3 Preliminaries

We consider RL with linear function approximation for episodic MDPs. The definition of episodic MDPs is in Section 1.2. Since the main difficulty of learning a MDP comes from learning the underlying transition dynamic $\{\mathbb{P}_h\}_h$, in this chapter we assume the reward function $\{r_h\}_h$ is deterministic and known to the agent. In the following, we will introduce the necessary background and definitions. For further background, the reader is advised to consult, e.g., Puterman (2014).

Online Reinforcement Learning A learning agent who does not know the kernels $\{\mathbb{P}_h\}_h$ but, for the sake of simplicity, knows the rewards $\{r_h\}_h$, aims to learn to take good actions by interacting with the environment. For each $k \geq 1$, at the beginning of the k-th episode, the environment picks the initial state s_1^k and the agent chooses a policy π^k to be followed in this episode. As the agent follows the policy through the episode, it observes the sequence of states $\{s_h^k\}_h$ with $s_{h+1}^k \sim \mathbb{P}_h(\cdot|s_h^k, \pi^k(s_h^k))$. The goal is to design a learning algorithm that constructs the sequence $\{\pi^k\}_k$ based on past information so that the K-episode regret,

$$\text{Regret}(M, K) = \sum_{k=1}^{K} \left[V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\right]$$

is kept small. In this chapter, we focus on proving high probability bounds on the regret $\text{Regret}(M, K)$, as well as lower bounds in expectation.

Linear Mixture MDPs We consider a special class of MDPs called *linear mixture MDPs* (a.k.a., linear kernel MDPs), where the transition probability kernel is a linear mixture of a

number of basis kernels. This class has been considered by a number of previous authors (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021b) and is defined as follows: Firstly, let $\boldsymbol{\phi}(s'|s,a) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ be a feature mapping satisfying that for any bounded function $V : \mathcal{S} \to [0,1]$ and any tuple $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\|\boldsymbol{\phi}_V(s,a)\|_2 \leq 1, \text{where } \boldsymbol{\phi}_V(s,a) = \sum_{s' \in \mathcal{S}} \boldsymbol{\phi}(s'|s,a)V(s')\,. \tag{2.3.1}$$

We define **episodic linear mixture MDPs** as follows:

Definition 2.3.1 (Jia et al. 2020; Ayoub et al. 2020)**.** $M = (\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h=1}^H, \{\mathbb{P}_h\}_{h=1}^H)$ is called an inhomogeneous, episodic B-bounded linear mixture MDP if there exist vectors $\boldsymbol{\theta}_h \in \mathbb{R}^d$ with $\|\boldsymbol{\theta}_h\|_2 \leq B$ and $\boldsymbol{\phi}(\cdot|\cdot,\cdot)$ satisfying (2.3.1), such that $\mathbb{P}_h(s'|s,a) = \langle \boldsymbol{\phi}(s'|s,a), \boldsymbol{\theta}_h \rangle$ for any state-action-next-state triplet $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and stage h.

Note that in the learning problem, the vectors introduced in the above definition are initially unknown to the learning agent. In the rest of this chapter, we assume that the learning agent is given access to $\boldsymbol{\phi}$ and the unknown episodic linear mixture MDP is parameterized by $\boldsymbol{\Theta}^* = \{\boldsymbol{\theta}_h^*\}_{h=1}^H$. We denote this MDP by $M_{\boldsymbol{\Theta}^*}$.

2.4 Challenges and New Technical Tools

To motivate our approach, we start this section with a recap of previous work addressing online learning in episodic linear mixture MDPs. This allows us to argue for how this work falls short of achieving minimax optimal regret and motivates us to develop new theoretical tools to achieve that.

2.4.1 Barriers to Minimax Optimality in RL with Linear Function Approximation

To understand the key technical challenges that underly achieving minimax optimality in RL with linear function approximation, we first look into the UCRL with "value-targeted

regression" (UCRL-VTR) method of Jia et al. (2020) (for a longer exposition, with refined results see Ayoub et al. (2020)) for episodic linear mixture MDPs. The key idea of UCRL-VTR is using a model-based supervised learning framework to learn the underlying unknown parameter vector $\boldsymbol{\theta}_h^*$ of linear mixture MDP, and use the learned parameter vector $\boldsymbol{\theta}_{k,h}$ to build an optimistic estimator $Q_{k,h}(\cdot,\cdot)$ for the optimal action-value function $Q^*(\cdot,\cdot)$. In detail, for any stage h of the k-th episode, the following equation holds: For value functions $V_k = \{V_{k,h}\}_h$ constructed based on data received before episode k and the state action pair (s_h^k, a_h^k) visited in stage h of episode k,

$$[\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) = \left\langle \sum_{s'} \boldsymbol{\phi}(s'|s_h^k, a_h^k) V_{k,h+1}(s'), \boldsymbol{\theta}_h^* \right\rangle = \left\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^* \right\rangle,$$

where the first equation holds due to the definition of linear mixture MDPs (cf. Definition 2.3.1), the second equation holds due to the definition of $\boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)$ in (2.3.1). As it turns out, taking actions that maximize the value shown above with appropriately constructed value functions V_k is sufficient for minimizing regret. Therefore, learning the underlying $\boldsymbol{\theta}_h^*$ can be regarded as solving a "linear bandit" problem (Part V, Lattimore and Szepesvári, 2020), where the context is $\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k) \in \mathbb{R}^d$, and the noise is $V_{k,h+1}(s_{h+1}^k) - [\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k)$. Previous work (Jia et al., 2020; Ayoub et al., 2020) proposed an estimator $\boldsymbol{\theta}_{k,h}$ as the minimizer to the following regularized linear regression problem:

$$\boldsymbol{\theta}_{k,h} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \lambda \|\boldsymbol{\theta}\|_2^2 + \sum_{j=1}^{k-1} \left[\left\langle \boldsymbol{\phi}_{V_{j,h+1}}(s_h^j, a_h^j), \boldsymbol{\theta} \right\rangle - V_{j,h+1}(s_{h+1}^j) \right]^2. \tag{2.4.1}$$

By using the standard self-normalized concentration inequality for vector-valued martingales of Abbasi-Yadkori et al. (2011), one can show then that, with high probability, $\boldsymbol{\theta}_h^*$ lies in the ellipsoid

$$\mathcal{C}_{k,h} = \left\{ \boldsymbol{\theta} : \left\| \boldsymbol{\Sigma}_{k,h}^{1/2} (\boldsymbol{\theta} - \boldsymbol{\theta}_{k,h}) \right\|_2 \leq \beta_k \right\}$$

which is centered at $\boldsymbol{\theta}_{k,h}$, with shape parameter $\boldsymbol{\Sigma}_{k,h} = \lambda \mathbf{I} + \sum_{j=1}^{k-1} \boldsymbol{\phi}_{V_{j,h+1}}(s_h^j, a_h^j) \boldsymbol{\phi}_{V_{j,h+1}}(s_h^j, a_h^j)^\top$ and where β_k is the radius chosen to be proportional to the magnitude of the value function

$V_{k,h+1}(\cdot)$, which eventually gives $\beta_k = \widetilde{O}(\sqrt{d}H)$. It follows that if we define

$$Q_{k,h}(\cdot,\cdot) = \left[r_h(\cdot,\cdot) + \max_{\boldsymbol{\theta} \in \mathcal{C}_{k,h}} \left\langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot) \right\rangle \right]_{[0,H]},$$

then, with high probability, $Q_{k,1}(\cdot,\cdot)$ is an overestimate of $Q_1^*(\cdot,\cdot)$, and the summation of "suboptimality gaps" can be bounded by $\sum_{k=1}^{K} \sum_{h=1}^{H} \beta_k \|\boldsymbol{\Sigma}_{k,h}^{-1/2} \boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\|_2$. This leads to the $\widetilde{O}(dH^{3/2}\sqrt{T})$ regret by further applying the elliptical potential lemma from linear bandits (Abbasi-Yadkori et al., 2011).

However, we note that the above reasoning has a number of shortcomings. First, it chooses the confidence radius β_k proportional to the *magnitude* of the value function $V_{k,h+1}(\cdot)$ rather than its *variance* $[\mathbb{V}_h V_{k,h+1}](\cdot,\cdot)$. This is known to be too conservative: Tabular RL is a special case of linear mixture MDPs and here it is known by the *law of total variance* (Lattimore and Hutter, 2012; Azar et al., 2013) that the variance of the value function is smaller than its magnitude by a factor $\sqrt{H}$. This inspires us to derive a Bernstein-type self-normalized concentration bound for vector-valued martingales which is sensitive to the variance of the martingale terms. Second, even if we were able to build such a tighter concentration bound, we still need to carefully design an algorithm because the variances of the value functions $\{\mathbb{V}_h V_{k,h+1}(s_h^k, a_h^k)\}_h$ at different stages of the episodes are non-uniform: We face a so-called *heteroscedastic* linear bandit problem. Naively choosing a uniform upper bound for all the variances $\{[\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)\}_h$ yields no improvement compared with previous results. To address this challenge, we will need to build variance estimates and use these in a weighted least-squares estimator to achieve a better aggregation of the heteroscedastic data.

2.4.2 A Bernstein Self-normalized Concentration Inequality for Vector-valued Martingales

One of the key results of this chapter is the following Bernstein self-normalized concentration inequality:

Theorem 2.4.1 (Bernstein inequality for vector-valued martingales)**.** Let $\{\mathcal{G}_t\}_{t=1}^{\infty}$ be a

filtration, $\{\mathbf{x}_t, \eta_t\}_{t\geq 1}$ be a stochastic process so that $\mathbf{x}_t \in \mathbb{R}^d$ is $\mathcal{G}_t$-measurable and $\eta_t \in \mathbb{R}$ is $\mathcal{G}_{t+1}$-measurable. Fix $R, L, \sigma, \lambda > 0$, $\boldsymbol{\mu}^* \in \mathbb{R}^d$. For $t \geq 1$ let $y_t = \langle \boldsymbol{\mu}^*, \mathbf{x}_t \rangle + \eta_t$ and suppose that $\eta_t, \mathbf{x}_t$ also satisfy

$$|\eta_t| \leq R,\ \mathbb{E}[\eta_t|\mathcal{G}_t] = 0,\ \mathbb{E}[\eta_t^2|\mathcal{G}_t] \leq \sigma^2,\ \|\mathbf{x}_t\|_2 \leq L.$$

Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$ we have

$$\forall t > 0,\ \left\| \sum_{i=1}^{t} \mathbf{x}_i \eta_i \right\|_{\mathbf{Z}_t^{-1}} \leq \beta_t,\ \|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} \leq \beta_t + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2, \tag{2.4.2}$$

where for $t \geq 1$, $\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1}\mathbf{b}_t$, $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^{t} \mathbf{x}_i\mathbf{x}_i^\top$, $\mathbf{b}_t = \sum_{i=1}^{t} y_i\mathbf{x}_i$ and

$$\beta_t = 8\sigma\sqrt{d\log(1 + tL^2/(d\lambda))\log(4t^2/\delta)} + 4R\log(4t^2/\delta)\,.$$

Proof. The proof adapts the proof technique of Dani et al. (2008); for details see Section 2.7.1. □

Theorem 2.4.1 can be viewed as a non-trivial extension of the Bernstein concentration inequality from scalar-valued martingales to self-normalized vector-valued martingales. It is a strengthened version of self-normalized tail inequality for vector-valued martingales when the magnitude and the variance of the noise are bounded. Abbasi-Yadkori et al. (2011) considered the setting where η_t is R-sub-Gaussian and showed that (2.4.2) holds when $\beta_t = R\sqrt{d\log((1 + tL^2/\lambda)/\delta)} = \widetilde{O}(R\sqrt{d})$, while our result improves this to $\beta_t = \widetilde{O}(\sigma\sqrt{d} + R)$.

It is worth to compare Theorem 2.4.1 with a few Hoeffding-Azuma-type results proved in prior work (Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011). In particular, Dani et al. (2008) considered the setting where η_t is R-bounded and showed that for large enough t, the following holds with probability at least $1 - \delta$:

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} \leq R\max\{\sqrt{128d\log(tL^2)\log(t^2/\delta)}, 8/3 \cdot \log(t^2/\delta)\}.$$

Rusmevichientong and Tsitsiklis (2010) considered a more general setting than Dani et al. (2008) where η_t is R-sub-Gaussian and showed that (2.4.2) holds when

$\beta_t = 2\kappa^2 R\sqrt{\log t}\sqrt{d\log t + \log(t^2/\delta)}$, where $\kappa = \sqrt{3 + 2\log(L^2/\lambda + d)}$. Abbasi-Yadkori et al. (2011) considered the same setting as Rusmevichientong and Tsitsiklis (2010) where η_t is R-sub-Gaussian and showed that (2.4.2) holds when $\beta_t = R\sqrt{d\log((1 + tL^2/\lambda)/\delta)}$, which improves the bound of Rusmevichientong and Tsitsiklis (2010) in terms of logarithmic factors. By selecting proper λ, all these results yield an $\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} = \widetilde{O}(R\sqrt{d})$ bound. As a comparison, with the choice $\lambda = \sigma^2 d/\|\boldsymbol{\mu}^*\|_2^2$, our result gives

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} = \widetilde{O}(\sigma\sqrt{d} + R). \tag{2.4.3}$$

Note that for any random variable, its standard deviation is always upper bounded by its magnitude or sub-Gaussian norm, therefore our result strictly improves the mentioned previous results. This improvement is due to the fact that here we consider a subclass of sub-Gaussian noise variables which allows us to derive a tighter upper bound. Indeed, Exercise 20.1 in the book of Lattimore and Szepesvári (2020) shows that the previous inequalities are tight in the worst-case for R-sub-Gaussian noise.

Even more closely related are results by Lattimore et al. (2015); Kirschner and Krause (2018) and Faury et al. (2020). In all these papers the strategy is to use a weighted ridge regression estimator, which we will also make use of in the next section. In particular, Lattimore et al. (2015) study the special case of Bernoulli payoffs. For this special case, with our notation, they show a result implying that with high probability $\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} = \widetilde{O}(\sigma\sqrt{d})$. The lack of the scale term R is due to that Bernoulli's are single-parameter: The variance and the mean control each other, which the proof exploits. As such, this result does not lead in a straightforward way to ours, where the scale and variance are independently controlled. A similar comment applies to the result of Kirschner and Krause (2018) who considered the case when the noise in the responses are sub-Gaussian.

2.4.3 Weighted Ridge Regression and Heteroscedastic Linear Bandits

In this subsection we consider the problem of linear bandits where the learner is given at the end of each round an upper bound on the (conditional) variance of the noise in the responses as input. This abstract problem is studied to work out the tools needed to handle the heteroscedasticity of the noise that arises in the linear mixture MDPs in a cleaner setting. In more details, let $\{\mathcal{D}_t\}_{t=1}^{\infty}$ be a fixed sequence of decision sets. The agent selects an action $\mathbf{a}_t \in \mathcal{D}_t$ and then observes the reward $r_t = \langle \boldsymbol{\mu}^*, \mathbf{a}_t \rangle + \epsilon_t$, where $\boldsymbol{\mu}^* \in \mathbb{R}^d$ is a vector unknown to the agent and ϵ_t is a random noise satisfying the following properties almost surely:

$$\forall t,\ |\epsilon_t| \le R,\ \mathbb{E}[\epsilon_t | \mathbf{a}_{1:t}, \epsilon_{1:t-1}] = 0,\ \mathbb{E}[\epsilon_t^2 | \mathbf{a}_{1:t}, \epsilon_{1:t-1}] \le \sigma_t^2,\ \|\mathbf{a}_t\|_2 \le A. \tag{2.4.4}$$

As noted above, the learner gets to observe σ_t together with r_t after each choice it makes. We assume that σ_t is $(\mathbf{a}_{1:t}, \epsilon_{1:t-1})$-measurable. The goal of the agent is to minimize its *pseudo-regret*, defined as follows:

$$\mathrm{Regret}(T) = \sum_{t=1}^{T} \langle \mathbf{a}_t^*, \boldsymbol{\mu}^* \rangle - \sum_{t=1}^{T} \langle \mathbf{a}_t, \boldsymbol{\mu}^* \rangle,\ \text{where } \mathbf{a}_t^* = \operatorname*{argmax}_{\mathbf{a} \in \mathcal{D}_t} \langle \mathbf{a}, \boldsymbol{\mu}^* \rangle.$$

Our problem setup is similar to the setting studied by Kirschner and Krause (2018), where it is not the variance, but the sub-Gaussianity parameter that the learner observes at the end of the rounds. The learner's goal is then to make use of this information to achieve a smaller regret as a function of the sum of squared variances (a "second-order bound"). This is also related to the Gaussian side-observation setting and partial monitoring with feedback graphs considered in Wu et al. (2015).

To make use of the variance information, we propose *Weighted OFUL*, which is an extension of the "Optimism in the Face of Uncertainty for Linear bandits" algorithm (OFUL) of Abbasi-Yadkori et al. (2011). The algorithm's pseudocode is shown in Algorithm 1.

In round t, Weighted OFUL selects the estimate $\widehat{\boldsymbol{\mu}}_t$ of the unknown $\boldsymbol{\mu}^*$ as the minimizer to the following *weighted ridge regression* problem:

$$\widehat{\boldsymbol{\mu}}_t \leftarrow \operatorname*{argmin}_{\boldsymbol{\mu} \in \mathbb{R}^d} \lambda \|\boldsymbol{\mu}\|_2^2 + \sum_{i=1}^{t} [\langle \boldsymbol{\mu}, \mathbf{a}_i \rangle - r_i]^2 / \bar{\sigma}_i^2, \tag{2.4.5}$$

Algorithm 1 Weighted OFUL

Require: Regularization parameter $\lambda > 0$, and B, an upper bound on the ℓ_2-norm of $\boldsymbol{\mu}^*$

1: $\mathbf{A}_0 \leftarrow \lambda \mathbf{I}$, $\mathbf{c}_0 \leftarrow \mathbf{0}$, $\widehat{\boldsymbol{\mu}}_0 \leftarrow \mathbf{A}_0^{-1}\mathbf{c}_0$, $\widehat{\beta}_0 = 0$, $\mathcal{C}_0 \leftarrow \{\boldsymbol{\mu} : \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_0\|_{\mathbf{A}_0} \leq \widehat{\beta}_0 + \sqrt{\lambda}B\}$

2: **for** $t = 1, \ldots, T$ **do**

3: Observe $\mathcal{D}_t$

4: Let $(\mathbf{a}_t, \widetilde{\boldsymbol{\mu}}_t) \leftarrow \operatorname{argmax}_{\mathbf{a} \in \mathcal{D}_t, \boldsymbol{\mu} \in \mathcal{C}_{t-1}} \langle \mathbf{a}, \boldsymbol{\mu} \rangle$

5: Select $\mathbf{a}_t$ and observe (r_t, σ_t), set $\bar{\sigma}_t$ based on σ_t, set radius $\widehat{\beta}_t$ as defined in (2.4.6)

6: $\mathbf{A}_t \leftarrow \mathbf{A}_{t-1} + \mathbf{a}_t \mathbf{a}_t^\top / \bar{\sigma}_t^2$, $\mathbf{c}_t \leftarrow \mathbf{c}_{t-1} + r_t \mathbf{a}_t / \bar{\sigma}_t^2$, $\widehat{\boldsymbol{\mu}}_t \leftarrow \mathbf{A}_t^{-1}\mathbf{c}_t$, $\mathcal{C}_t \leftarrow \{\boldsymbol{\mu} : \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_t\|_{\mathbf{A}_t} \leq \widehat{\beta}_t + \sqrt{\lambda}B\}$

7: **end for**

where $\bar{\sigma}_i$ is a selected upper bound of σ_i. The closed-form solution to (2.4.5) is in Line 6 of Algorithm 1. The term "weighted" refers to the normalization constant $\bar{\sigma}_i$ used in (2.4.5). The estimator in (2.4.5) is closely related to the best linear unbiased estimator (BLUE) (Henderson, 1975). In particular, in the language of linear regression, with $\lambda = 0$ and when $\bar{\sigma}_t^2$ is the variance of r_t, with a fixed design, $\widehat{\boldsymbol{\mu}}_t$ is known to be the lowest variance estimator of $\boldsymbol{\mu}^*$ in the class of linear unbiased estimators. Note that both Lattimore et al. (2015) and Kirschner and Krause (2018) used a similar weighted ridge-regression estimator for their respective problem settings.

By adapting the new Bernstein-type self-normalized concentration inequality in Theorem 2.4.1, we obtain the following bound on the regret of Weighted OFUL:

Theorem 2.4.2. Suppose that for all $t \geq 1$ and all $\mathbf{a} \in \mathcal{D}_t$, $\langle \mathbf{a}, \boldsymbol{\mu}^* \rangle \in [-1, 1]$, $\|\boldsymbol{\mu}^*\|_2 \leq B$. Set $\bar{\sigma}_t = \max\{R/\sqrt{d}, \sigma_t\}$, $\lambda = 1/B^2$ and

$$\widehat{\beta}_0 = 0,\ \widehat{\beta}_t = 8\sqrt{d \log(1 + tA^2/([\bar{\sigma}_{\min}^t]^2 d\lambda)) \log(4t^2/\delta)} + 4R/\bar{\sigma}_{\min}^t \cdot \log(4t^2/\delta),\ t \geq 1\,. \quad (2.4.6)$$

where $\bar{\sigma}_{\min}^t = \min_{1 \leq i \leq t} \bar{\sigma}_i$. Then, with probability at least $1 - \delta$, the regret of Weighted OFUL

for the first T rounds is bounded as follows:

$$\text{Regret}(T) = \widetilde{O}\bigg(R\sqrt{dT} + d\sqrt{\sum_{t=1}^{T}\sigma_t^2}\bigg). \tag{2.4.7}$$

Proof. See Section 2.7.2. □

Remark 2.4.3. Comparing (2.4.7) of Theorem 2.4.2 with the regret bound $\text{Regret}(T) = \widetilde{O}(Rd\sqrt{T})$ achieved by OFUL in Abbasi-Yadkori et al. (2011), it can be seen that the regret of Weighted OFUL is strictly better than that of OFUL since $\sigma_t \leq R$.

2.5 Optimal Exploration for Episodic Linear Mixture MDPs

In this section, equipped with the new technical tools discussed in Section 2.4, we propose a new algorithm UCRL-VTR$^+$ for episodic linear mixture MDPs (see Definition 2.3.1). We also prove its near minimax optimality by providing matching upper and lower bounds.

2.5.1 The Proposed Algorithm

At a high level, UCRL-VTR$^+$ is an improved version of the UCRL-VTR algorithm by Jia et al. (2020) and refined and generalized by Ayoub et al. (2020). UCRL-VTR$^+$, shares the basic structure of UCRL-VTR, which constructs the optimistic estimate of the optimal action-value function at k-th episode and h-th stage as follows, following the optimism in the face of uncertainty principle:

$$Q_{k,h}(\cdot,\cdot) = \Big[r_h(\cdot,\cdot) + \max_{\boldsymbol{\theta}\in\widehat{\mathcal{C}}_{k,h}} \big\langle \boldsymbol{\theta}, \boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\big\rangle\Big]_{[0,H]}. \tag{2.5.1}$$

where the confidence set $\widehat{\mathcal{C}}_{k,h}$ constructed is an ellipsoid in the parameter space, centered at the parameter vector $\widehat{\boldsymbol{\theta}}_{k,h}$ and shape given by the "covariance" matrix $\widehat{\boldsymbol{\Sigma}}_{k,h}$ and having a radius of $\widehat{\beta}_k$:

$$\widehat{\mathcal{C}}_{k,h} = \bigg\{\boldsymbol{\theta} : \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{k,h})\Big\|_2 \leq \widehat{\beta}_k\bigg\}, \tag{2.5.2}$$

Algorithm 2 UCRL-VTR$^+$ for Episodic Linear Mixture MDPs

Require: Regularization parameter λ, an upper bound B of the ℓ_2-norm of $\boldsymbol{\theta}_h^*$

For $h \in [H]$, set $\widehat{\boldsymbol{\Sigma}}_{1,h}, \widetilde{\boldsymbol{\Sigma}}_{1,h} \leftarrow \lambda \mathbf{I}$, $\widehat{\mathbf{b}}_{1,h}, \widetilde{\mathbf{b}}_{1,h} \leftarrow \mathbf{0}$, $\widehat{\boldsymbol{\theta}}_{1,h}, \widetilde{\boldsymbol{\theta}}_{1,h} \leftarrow \mathbf{0}$, $V_{1,H+1}(\cdot) \leftarrow 0$

for $k = 1, \ldots, K$ **do**

  for $h = H, \ldots, 1$ **do**

    $Q_{k,h}(\cdot,\cdot) \leftarrow \Big[r_h(\cdot,\cdot) + \big\langle \widehat{\boldsymbol{\theta}}_{k,h}, \boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\big\rangle + \widehat{\beta}_k \Big\| \widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2} \boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot) \Big\|_2 \Big]_{[0,H]}$, where $\widehat{\beta}_k$ is defined in (2.5.8)

    $\pi_h^k(\cdot) \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} Q_{k,h}(\cdot, a)$

    $V_{k,h}(\cdot) \leftarrow \max_{a \in \mathcal{A}} Q_{k,h}(\cdot, a)$

  end for

  Receive s_1^k

  for $h = 1, \ldots, H$ **do**

    Take action $a_h^k \leftarrow \pi_h^k(s_h^k)$, receive $s_{h+1}^k \sim \mathbb{P}_h(\cdot | s_h^k, a_h^k)$

    Set $[\bar{\mathbb{V}}_{k,h} V_{k,h+1}](s_h^k, a_h^k)$ as in (2.5.7) and $E_{k,h}$ as in (2.5.9)

    $\bar{\sigma}_{k,h} \leftarrow \sqrt{\max\big\{H^2/d, [\bar{\mathbb{V}}_{k,h} V_{k,h+1}](s_h^k, a_h^k) + E_{k,h}\big\}}$ {Variance upper bound}

    $\widehat{\boldsymbol{\Sigma}}_{k+1,h} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k,h} + \bar{\sigma}_{k,h}^{-2} \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k) \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)^\top$ {"Covariance", 1st moment}

    $\widehat{\mathbf{b}}_{k+1,h} \leftarrow \widehat{\mathbf{b}}_{k,h} + \bar{\sigma}_{k,h}^{-2} \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k) V_{k,h+1}(s_{h+1}^k)$ {Response, 1st moment}

    $\widetilde{\boldsymbol{\Sigma}}_{k+1,h} \leftarrow \widetilde{\boldsymbol{\Sigma}}_{k,h} + \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k) \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)$ {"Covariance", 2nd moment}

    $\widetilde{\mathbf{b}}_{k+1,h} \leftarrow \widetilde{\mathbf{b}}_{k,h} + \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k) V_{k,h+1}^2(s_{h+1}^k)$ {Response, 2nd moment}

    $\widehat{\boldsymbol{\theta}}_{k+1,h} \leftarrow \widehat{\boldsymbol{\Sigma}}_{k+1,h}^{-1} \widehat{\mathbf{b}}_{k+1,h}$, $\widetilde{\boldsymbol{\theta}}_{k+1,h} \leftarrow \widetilde{\boldsymbol{\Sigma}}_{k+1,h}^{-1} \widetilde{\mathbf{b}}_{k+1,h}$ {1st and 2nd moment parameters}

  end for

end for

Given $\{Q_{k,h}\}_h$, in each episode k, at h-th stage, UCRL-VTR$^+$ executes actions that are greedy with respect to $Q_{k,h}$ (Line 5).

Epistemic Uncertainty Estimate Intuitively speaking, we want to measure the epistemic uncertainty of the transition dynamic $\mathbb{P}_h = \langle \boldsymbol{\theta}_h^*, \boldsymbol{\phi} \rangle$. At k-th episode, the estimate of $\boldsymbol{\theta}_h^*$ is $\widehat{\boldsymbol{\theta}}_{k,h}$, therefore the epistemic uncertainty is the error between the estimated transition dynamic $\widehat{\mathbb{P}}_h = \langle \widehat{\boldsymbol{\theta}}_{k,h}, \boldsymbol{\phi} \rangle$, and the true dynamic $\mathbb{P}_h = \langle \boldsymbol{\theta}_h^*, \boldsymbol{\phi} \rangle$. For any value function V, the difference between the estimated transition dynamic and the true transition dynamic applied over V can be bounded as

$$\begin{aligned}
\widehat{\mathbb{P}}_h V(s,a) - \mathbb{P}V(s,a) &= \langle \boldsymbol{\theta}_h^* - \widehat{\boldsymbol{\theta}}_{k,h}, \boldsymbol{\phi}_V \rangle \\
&\leq \|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2}(\boldsymbol{\theta}_h^* - \widehat{\boldsymbol{\theta}}_{k,h})\|_2 \cdot \|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_V(s,a)\|_2 \\
&\leq \widehat{\beta}_k \|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_V(s,a)\|_2, \qquad (2.5.3)
\end{aligned}$$

where the last inequality holds due to the definition of $\widehat{\mathcal{C}}_{k,h}$ and Lemma 2.5.1 (will be showed later). (2.5.3) gives the *epistemic uncertainty estimate* for $\widehat{\mathbb{P}}_h$ with respect to the value function V. Given the choice of $\widehat{\mathcal{C}}_{k,h}$, it is not hard to see that the update in Line 4 is equivalent to (2.5.1).

Weighted Ridge Regression and Optimistic Estimates of Value Functions The key novelty of UCRL-VTR$^+$ is the use of the covariance matrix $\widehat{\boldsymbol{\Sigma}}_{k,h}$ (Line 13) and the parameter vector $\widehat{\boldsymbol{\theta}}_{k,h}$ (Line 17) based on weighted ridge regression (cf. Section 2.4) to learn the underlying $\boldsymbol{\theta}_h^*$. To understand the mechanism behind UCRL-VTR$^+$, recall the discussion in Section 2.4.1: $V_{k,h+1}(s_{h+1}^k)$ and $\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)$ can be seen as the stochastic reward and context of a linear bandits problem. Then, letting $\sigma_{k,h}^2 = [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)$ be the variance of the value function, the analysis in Section 2.4 suggests that one should use a weighted ridge regression estimator, such as

$$\widehat{\boldsymbol{\theta}}_{k,h} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \lambda \|\boldsymbol{\theta}\|_2^2 + \sum_{j=1}^{k-1} \left[\left\langle \boldsymbol{\phi}_{V_{j,h+1}}(s_h^j, a_h^j), \boldsymbol{\theta} \right\rangle - V_{j,h+1}(s_{h+1}^j) \right]^2 / \bar{\sigma}_{j,h}^2, \qquad (2.5.4)$$

where $\bar{\sigma}_{j,h}$ is an appropriate upper bound on $\sigma_{j,h}$. We propose to set

$$\bar{\sigma}_{k,h} = \sqrt{\max\left\{H^2/d, [\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) + E_{k,h}\right\}},$$

where $[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k)$ is a scalar-valued empirical estimate for the variance of the value function $V_{k,h+1}$ under the transition probability $\mathbb{P}_h(\cdot|s_k, a_k)$, and $E_{k,h}$ is an offset term that is used to guarantee that $[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) + E_{k,h}$ upper bounds $\sigma_{k,h}^2$ with high probability. The detailed specifications of these are deferred later. Moreover, by construction, we have $\bar{\sigma}_{k,h} \geq H/\sqrt{d}$. Our construction of $\bar{\sigma}_{k,h}$ shares a similar spirit as the variance estimator used in *empirical Bernstein inequalities* (Audibert et al., 2009; Maurer and Pontil, 2009), which proved to be pivotal to achieve nearly minimax optimal sample complexity/regret in tabular MDPs (Azar et al., 2013, 2017; Zanette and Brunskill, 2019; He et al., 2021b).

Several nontrivial questions remain to be resolved. First, we need to specify how to calculate the empirical variance $[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k)$. Second, in order to ensure $Q_{k,h}(\cdot,\cdot)$ is an overestimate of $Q_h^*(\cdot,\cdot)$, we need to choose an appropriate $\widehat{\beta}_k$ such that $\widehat{\mathcal{C}}_{k,h}$ contains $\boldsymbol{\theta}_h^*$ with high probability. Third, we need to select $E_{k,h}$ to guarantee that $[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) + E_{k,h}$ upper bounds $\sigma_{k,h}^2$ with high probability.

Aleatoric Uncertainty Estimate We denote the variance of the value function $V_{k,h+1}$ as the *aleatoric uncertainty* of the MDP $M_{\boldsymbol{\Theta}^*}$, since it measures the stochasticity level of $\mathbb{P}_h$ and r_h. We only need to consider $\mathbb{P}_h$ since r_h is deterministic and known to the agent. By definition, we have

$$\begin{aligned}[\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k) &= [\mathbb{P}_h V_{k,h+1}^2](s_h^k, a_h^k) - \big([\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k)\big)^2 \\ &= \big\langle \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), \boldsymbol{\theta}_h^* \big\rangle - \big[\big\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^* \big\rangle\big]^2, \end{aligned} \tag{2.5.5}$$

where the second equality holds due to the definition of linear mixture MDPs. By (2.5.5) we conclude that the expectation of $V_{k,h+1}^2(s_{h+1}^k)$ over the next state, s_{h+1}^k, is a linear function of $\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)$. Therefore, we use $\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widetilde{\boldsymbol{\theta}}_{k,h} \rangle$ to estimate this term, where $\widetilde{\boldsymbol{\theta}}_{k,h}$ is

the solution to the following ridge regression problem:

$$\widetilde{\boldsymbol{\theta}}_{k,h} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d} \lambda\|\boldsymbol{\theta}\|_2^2 + \sum_{j=1}^{k-1} \left[\left\langle \boldsymbol{\phi}_{V_{j,h+1}^2}(s_h^j, a_h^j), \boldsymbol{\theta}\right\rangle - V_{j,h+1}^2(s_{h+1}^j)\right]^2. \tag{2.5.6}$$

The closed-form solution to (2.5.6) is in Line 17. In addition, we use $\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h}\rangle$ to estimate the second term in (2.5.5). Meanwhile, since $[\mathbb{P}_h V_{k,h+1}^2](s_h^k, a_h^k) \in [0, H^2]$ and $[\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) \in [0, H]$ hold, we add clipping to control the range of our variance estimator, which gives the final expression of $[\bar{\mathbb{V}}_{k,h} V_{k,h+1}](s_h^k, a_h^k)$, the *aleatoric uncertainty estimate*:

$$[\bar{\mathbb{V}}_{k,h} V_{k,h+1}](s_h^k, a_h^k) = \left[\left\langle \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), \widetilde{\boldsymbol{\theta}}_{k,h}\right\rangle\right]_{[0,H^2]} - \left[\left\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \widehat{\boldsymbol{\theta}}_{k,h}\right\rangle\right]_{[0,H]}^2. \tag{2.5.7}$$

UCRL-VTR$^+$ with single estimation sequence Currently UCRL-VTR$^+$ uses two estimate sequences $\check{\boldsymbol{\theta}}_{k,h}$ and $\widetilde{\boldsymbol{\theta}}_{k,h}$ to estimate the first-order moment $\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^*\rangle$ and second-order moment $\langle \boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), \boldsymbol{\theta}_h^*\rangle$ separately. We would like to point out that it is possible to use only one sequence to estimate both. Such an estimator can be constructed as a weighted ridge regression estimator based on both $\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)$'s and $\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)$, and the corresponding responses $V_{k,h+1}(s_{h+1}^k)$ and $V_{k,h+1}^2(s_{h+1}^k)$. However, since second-order moments generally have larger variance than the first-order moments, we need to use different weights for the square loss evaluated at $\left\{\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), V_{k,h+1}(s_{h+1}^k)\right\}_{k,h}$ and $\left\{\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k), V_{k,h+1}^2(s_{h+1}^k)\right\}_{k,h}$. Also, by merging the data, even with using perfect weighting, we would expect to win at best a (small) constant factor on the regret since the effect of not merging the data can be seen as not worse than throwing away "half of the data". As a result, for the sake of simplicity, we chose to use two estimate sequences instead of one in our algorithm.

Computational Efficiency of UCRL-VTR$^+$ Similar to UCRL-VTR (Ayoub et al., 2020), the computational complexity of UCRL-VTR$^+$ depends on the specific family of feature mapping $\boldsymbol{\phi}(\cdot|\cdot,\cdot)$. As an example, let us consider a special class of linear mixture MDPs studied by Yang and Wang (2020); Zhou et al. (2021b). In this setting, $\boldsymbol{\phi}(s'|s,a) = \boldsymbol{\psi}(s') \odot \boldsymbol{\mu}(s,a)$,

$\boldsymbol{\psi}(\cdot) : \mathcal{S} \to \mathbb{R}^d$ and $\boldsymbol{\mu}(\cdot,\cdot) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ are two features maps and $\odot$ denotes componentwise product. Recall that, by assumption, the action space $\mathcal{A}$ is finite.

We now argue that UCRL-VTR$^+$ is computationally efficient for this class of MDPs as long as we have access to an integration oracle $\mathcal{O}$ underlying the basis kernels. In particular, the assumption is that $\sum_{s'} \psi(s')V(s')$ can be evaluated at the cost of evaluating V at $p(d)$ states with some polynomial p. Now, for $1 \le h \le H$, $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ let

$$Q_{h,\boldsymbol{\theta},\boldsymbol{\Sigma}}(\cdot,\cdot) = \left[r_h(\cdot,\cdot) + \langle \boldsymbol{\theta}, \boldsymbol{\mu}(\cdot,\cdot)\rangle + \|\boldsymbol{\Sigma}\boldsymbol{\mu}(\cdot,\cdot)\|_2\right]_{[0,H]}.$$

It is easy to verify that for any k, h, $Q_{k,h} = Q_{h,\boldsymbol{\theta}_{k,h},\boldsymbol{\Sigma}_{k,h}}$ where $\boldsymbol{\theta}_{k,h} = \widehat{\boldsymbol{\theta}}_{k,h} \odot [\sum_{s'} \boldsymbol{\psi}(s')V_{k,h+1}(s')]$ and the (i,j)-th entry of $\boldsymbol{\Sigma}_{k,h}$ is $\widehat{\beta}_k(\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2})_{i,j}[\sum_{s'} \boldsymbol{\psi}_j(s')V_{k,h+1}(s')]$. Now notice that $\boldsymbol{\theta}_{k,H} = \mathbf{0}$, $\boldsymbol{\Sigma}_{k,H} = \mathbf{0}$. Thus, for $1 \le h \le H-1$, assuming that $\boldsymbol{\theta}_{k,h+1}$ and $\boldsymbol{\Sigma}_{k,h+1}$ have been calculated, evaluating $V_{k,h+1}$ at any state $s \in \mathcal{S}$ costs $O(d^2|\mathcal{A}|)$ arithmetic operations. Now, calculating $\boldsymbol{\theta}_{k,h}$ and $\boldsymbol{\Sigma}_{k,h}$ costs $O(d^2)$ arithmetic operations given access $\widehat{\boldsymbol{\theta}}_{k,h}$ and $\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}$, in addition to $p(d)$ evaluations of $V_{k,h+1}$. Since each evaluation of $V_{k,h+1}$ takes $O(d^2|\mathcal{A}|)$ operations, as established, calculating $\boldsymbol{\theta}_{k,h}$ and $\boldsymbol{\Sigma}_{k,h}$ cost a total of $O(p(d)d^2|\mathcal{A}|)$ operations. From this, it is clear that calculating the H actions to be taken in episode k takes a total of $O(p(d)d^2|\mathcal{A}|H)$ operations (Line 10). It also follows that calculating either $\boldsymbol{\phi}_{V_{k,h+1}}$ or $\boldsymbol{\phi}_{V_{k,h+1}^2}$ at any state-action pair costs $O(p(d)d^2|\mathcal{A}|)$ operations.

To calculate the quantities appearing in Lines 11–17, first $\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)$ and $\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)$ ($h \in [H]$) are evaluated at the cost of $O(p(d)d^2|\mathcal{A}|H)$. It is then clear that the rest of the calculation costs at most $O(d^3H)$: the most expensive step is to obtain $\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}$ (the cost could be reduced to $O(d^2H)$ by using the matrix inversion lemma and organizing the calculation of $Q_{k,h}$ slightly differently). It follows that the total computational complexity of UCRL-VTR$^+$ is $O(\mathrm{poly}(d)|\mathcal{A}|HK) = O(\mathrm{poly}(d)|\mathcal{A}|T)$.

Confidence Set To address the choice of $\widehat{\beta}_k$ and $E_{k,h}$, we need the following key technical lemma:

Lemma 2.5.1. Let $\widehat{\mathcal{C}}_{k,h}$ be defined in (2.5.2) and set $\widehat{\beta}_k$ as

$$\widehat{\beta}_k = 8\sqrt{d\log(1+k/\lambda)\log(4k^2H/\delta)} + 4\sqrt{d}\log(4k^2H/\delta) + \sqrt{\lambda}B\,. \tag{2.5.8}$$

Then, with probability at least $1-3\delta$, we have that simultaneously for all $k \in [K]$ and $h \in [H]$,

$$\boldsymbol{\theta}_h^* \in \widehat{\mathcal{C}}_{k,h},\ |[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)| \le E_{k,h},$$

where $E_{k,h}$ is defined as follows:

$$E_{k,h} = \min\left\{H^2, 2H\check{\beta}_k \left\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\right\|_2\right\} + \min\left\{H^2, \widetilde{\beta}_k \left\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\right\|_2\right\}, \tag{2.5.9}$$

with

$$\begin{aligned}
\check{\beta}_k &= 8d\sqrt{\log(1+k/\lambda)\log(4k^2H/\delta)} + 4\sqrt{d}\log(4k^2H/\delta) + \sqrt{\lambda}B, \\
\widetilde{\beta}_k &= 8\sqrt{dH^4\log(1+kH^4/(d\lambda))\log(4k^2H/\delta)} + 4H^2\log(4k^2H/\delta) + \sqrt{\lambda}B.
\end{aligned}$$

Proof. See Section 2.8.1. □

Lemma 2.5.1 shows that with high probability, for all stages h and episodes k, $\boldsymbol{\theta}_h^*$ lies in the confidence set centered at its estimate $\widehat{\boldsymbol{\theta}}_{k,h}$, and the error between the estimated variance and the true variance is bounded by the offset term $E_{k,h}$. Equipped with Lemma 2.5.1, we can verify the following facts: First, since $\boldsymbol{\theta}_h^* \in \widehat{\mathcal{C}}_{k,h}$, it can be easily verified that $\langle\widehat{\boldsymbol{\theta}}_{k,h}, \boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\rangle + \widehat{\beta}_k\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\|_2 \ge \langle\boldsymbol{\theta}_h^*, \boldsymbol{\phi}_{V_{k,h+1}}(\cdot,\cdot)\rangle = [\mathbb{P}_h V_{k,h+1}](\cdot,\cdot)$, which shows that our constructed $Q_{k,h}(\cdot,\cdot)$ in Line 4 is indeed an overestimate of $Q_h^*(\cdot,\cdot)$. Second, recalling the definition of $\bar{\sigma}_{k,h}$ defined in Line 12, since $\left|[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)\right| \le E_{k,h}$, we have $\bar{\sigma}_{k,h}^2 \ge [\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) + E_{k,h} \ge [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)$, which shows that $\bar{\sigma}_{k,h}$ is indeed an overestimate of the true variance $[\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)$.

2.5.2 Regret Upper Bound

Now we present the regret upper bound of UCRL-VTR$^+$.

Theorem 2.5.2. Set $\lambda = 1/B^2$. Then, with probability at least $1 - 5\delta$, the regret of UCRL-VTR$^+$ on MDP $M_{\boldsymbol{\Theta}^*}$ is upper bounded as follows:

$$\mathrm{Regret}\big(M_{\boldsymbol{\Theta}^*}, K\big) = \widetilde{O}\Big(\sqrt{d^2H^2 + dH^3}\sqrt{T} + d^2H^3 + d^3H^2\Big),\ T = KH. \tag{2.5.10}$$

Proof Sketch. The detailed proof is given in Section 2.8.2. By Lemma 2.5.1, it suffices to prove the result on the event $\mathcal{E}$ when the conclusions of this lemma hold. Hence, in what follows assume that this event holds. By using the standard regret decomposition and using the definition of the confidence sets $\{\widehat{\mathcal{C}}_{k,h}\}_{k,h}$, we can show that the total regret is bounded by the summation of the bonus terms, $\sum_{k=1}^K \sum_{h=1}^H \widehat{\beta}_k \big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\big\|_2$, which, by the Cauchy-Schwarz inequality, can be further bounded by $\widehat{\beta}_K\sqrt{dH\sum_{k=1}^K\sum_{h=1}^H \bar{\sigma}_{k,h}^2}$. Finally, by the definition of $\bar{\sigma}_{k,h}^2$ we have $\bar{\sigma}_{k,h}^2 \le H^2/d + E_{k,h} + [\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) \le H^2/d + 2E_{k,h} + [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)$. Therefore the summation of $\bar{\sigma}_{k,h}^2$ can be bounded as

$$\begin{aligned}\sum_{k=1}^K\sum_{h=1}^H \bar{\sigma}_{k,h}^2 &\le H^3K/d + 2\sum_{k=1}^K\sum_{h=1}^H E_{k,h} + \sum_{k=1}^K\sum_{h=1}^H [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)\\ &= \widetilde{O}(HT + H^2T/d + dH^3\sqrt{T}),\end{aligned} \tag{2.5.11}$$

where the equality holds since by the *law of total variance* (Lattimore and Hutter, 2012; Azar et al., 2013), $\sum_{k=1}^K\sum_{h=1}^H [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k) = \widetilde{O}(HT)$, and $\sum_{k=1}^K\sum_{h=1}^H E_{k,h} = \widetilde{O}(dH^3\sqrt{T} + d^{1.5}H^{2.5}\sqrt{T})$ by the elliptical potential lemma. □

Remark 2.5.3. When $d \ge H$ and $T \ge d^4H^2 + d^3H^3$, the regret in (2.5.10) can be simplified to $\widetilde{O}(dH\sqrt{T})$. Compared with the regret $\widetilde{O}(dH^{3/2}\sqrt{T})$ of UCRL-VTR in Jia et al. (2020); Ayoub et al. (2020)[3], the regret of UCRL-VTR$^+$ is improved by a factor of $\sqrt{H}$.

[3]Jia et al. (2020); Ayoub et al. (2020) report a regret of order $\widetilde{O}(dH\sqrt{T})$. However, these works considered the time-homogeneous case where $\mathbb{P}_1 = \cdots = \mathbb{P}_H$. In particular, in the time-homogeneous setting parameters are shared between the stages of an episode, and this reduces the regret. When UCRL-VTR is modified for the inhomogenous case, the regret picks up an additional $\sqrt{H}$ factor. Similar observation has also been made by Jin et al. (2018).

2.5.3 Lower Bound

In this subsection, we present a lower bound for episodic linear mixture MDPs, which shows the optimality of UCRL-VTR$^+$.

Theorem 2.5.4. Let $B > 1$ and suppose $K \geq \max\{(d-1)^2 H/2, (d-1)/(32H(B-1))\}$, $d \geq 4$, $H \geq 3$. Then for any algorithm there exists an episodic, B-bounded linear mixture MDP parameterized by $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_H)$ such that the expected regret is lower bounded as follows:

$$\mathbb{E}_{\boldsymbol{\Theta}} \text{Regret}\big(M_{\boldsymbol{\Theta}}, K\big) \geq \Omega\big(dH\sqrt{T}\big),$$

where $T = KH$ and $\mathbb{E}_{\boldsymbol{\Theta}}$ denotes the expectation over the probability distribution generated by the interconnection of the algorithm and the MDP.

Proof Sketch. We construct a hard-to-learn MDP instance M. The detailed construction and proof are given in Section 2.9.1 and 2.9.2. We show that learning the optimal policy of such an MDP is no *harder* than minimizing the regret on H linear bandit problems, where the payoff for the first $H/2$ bandits is $\Omega(H)Z$. Here Z is a Bernoulli random variable with mean equal to $\Theta(1/H)$. Utilizing existing lower bound results for linear bandits (Lattimore and Szepesvári, 2020) yields our result. □

Remark 2.5.5. Theorem 2.5.4 shows that for any algorithm running on episodic linear mixture MDPs, its regret is lower bounded by $\Omega(dH\sqrt{T})$. The lower bound together with the upper bound of UCRL-VTR$^+$ in Theorem 2.5.2 shows that UCRL-VTR$^+$ is minimax optimal up to logarithmic factors.

2.6 Conclusion

In this chapter, we proposed a new Bernstein-type concentration inequality for self-normalized vector-valued martingales, which was shown to tighten existing confidence sets for linear

bandits when the reward noise has low variance σ_t^2 and is almost surely uniformly bounded by a constant $R > 0$. This also allowed us to derive a bandit algorithm for the stochastic linear bandit problem with changing actions sets. The proposed algorithm uses weighted least-squares estimates and achieves a second-order regret bound of order $\widetilde{O}(R\sqrt{dT} + d\sqrt{\sum_{t=1}^{T} \sigma_t^2})$, which is a significant improvement on the dimension dependence in the low-noise regime. Based on the new tail inequality, we propose a new, computationally efficient algorithm, UCRL-VTR$^+$ for episodic MDPs with an $\widetilde{O}(dH\sqrt{T} + \sqrt{dH^3}\sqrt{T} + d^2H^3 + d^3H^2)$ regret.

2.7 Proofs of Theorems in Section 2.4

2.7.1 Proof of Theorem 2.4.1

We follow the proof in Dani et al. (2008) with a refined analysis. Let us start with recalling two well known results that we will need:

Lemma 2.7.1 (Freedman 1975)**.** Let $M, v > 0$ be fixed constants. Let $\{x_i\}_{i=1}^n$ be a stochastic process,$\{\mathcal{G}_i\}_i$ be a filtration so that so that for all $i \in [n]$ x_i is $\mathcal{G}_i$-measurable, while almost surely $\mathbb{E}[x_i|\mathcal{G}_{i-1}] = 0$, $|x_i| \leq M$ and

$$\sum_{i=1}^{n} \mathbb{E}(x_i^2|\mathcal{G}_i) \leq v\,.$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\sum_{i=1}^{n} x_i \leq \sqrt{2v\log(1/\delta)} + 2/3 \cdot M\log(1/\delta).$$

Lemma 2.7.2 (Lemma 11, Abbasi-Yadkori et al. 2011)**.** For any $\lambda > 0$ and sequence $\{\mathbf{x}_t\}_{t=1}^T \subset \mathbb{R}^d$ for $t \in \{0, 1, \ldots, T\}$, define $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^{t} \mathbf{x}_i\mathbf{x}_i^\top$. Then, provided that $\|\mathbf{x}_t\|_2 \leq L$ holds for all $t \in [T]$, we have

$$\sum_{t=1}^{T} \min\{1, \|\mathbf{x}_t\|_{\mathbf{Z}_{t-1}^{-1}}^2\} \leq 2d\log\frac{d\lambda + TL^2}{d\lambda}\,.$$

Recall that for $t \geq 0$, $\mathbf{Z}_t = \lambda \mathbf{I} + \sum_{i=1}^{t} \mathbf{x}_i \mathbf{x}_i^\top$. Since $\mathbf{Z}_t = \mathbf{Z}_{t-1} + \mathbf{x}_t \mathbf{x}_t^\top$, by the matrix inversion lemma

$$\mathbf{Z}_t^{-1} = \mathbf{Z}_{t-1}^{-1} - \frac{\mathbf{Z}_{t-1}^{-1} \mathbf{x}_t \mathbf{x}_t^\top \mathbf{Z}_{t-1}^{-1}}{1 + w_t^2}. \tag{2.7.1}$$

We need the following definitions:

$$\mathbf{d}_0 = 0,\ Z_0 = 0,\ \mathbf{d}_t = \sum_{i=1}^{t} \mathbf{x}_i \eta_i,\ Z_t = \|\mathbf{d}_t\|_{\mathbf{Z}_t^{-1}},\ w_t = \|\mathbf{x}_t\|_{\mathbf{Z}_{t-1}^{-1}},\ \mathcal{E}_t = \mathbb{1}\{0 \leq s \leq t, Z_s \leq \beta_s\}, \tag{2.7.2}$$

where $t \geq 1$ and we define $\beta_0 = 0$. Recalling that x_t is $\mathcal{G}_t$-measurable and η_t is $\mathcal{G}_{t+1}$-measurable, we find that d_t, Z_t and $\mathcal{E}_t$ are $\mathcal{G}_{t+1}$-measurable while w_t is $\mathcal{G}_t$ measurable. We now prove the following result:

Lemma 2.7.3. Let $\mathbf{d}_i, w_i, \mathcal{E}_i$ be as defined in (2.7.2). Then, with probability at least $1 - \delta/2$, simultaneously for all $t \geq 1$ it holds that

$$\sum_{i=1}^{t} \frac{2\eta_i \mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1} \mathbf{d}_{i-1}}{1 + w_i^2} \mathcal{E}_{i-1} \leq 3\beta_t^2/4.$$

Proof. We have

$$\left| \frac{2\mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1} \mathbf{d}_{i-1}}{1 + w_i^2} \mathcal{E}_{i-1} \right| \leq \frac{2\|\mathbf{x}_i\|_{\mathbf{Z}_{i-1}^{-1}} [\|\mathbf{d}_{i-1}\|_{\mathbf{Z}_{i-1}^{-1}} \mathcal{E}_{i-1}]}{1 + w_i^2} \leq \frac{2w_i \beta_{i-1}}{1 + w_i^2} \leq \min\{1, 2w_i\} \beta_{i-1}, \tag{2.7.3}$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds due to the definition of $\mathcal{E}_{i-1}$, the last inequality holds by algebra. For simplicity, let ℓ_i denote

$$\ell_i = \frac{2\eta_i \mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1} \mathbf{d}_{i-1}}{1 + w_i^2} \mathcal{E}_{i-1}. \tag{2.7.4}$$

We are preparing to apply Freedman's inequality from Lemma 2.7.1 to $\{\ell_i\}_i$ and $\{\mathcal{G}_i\}_i$. First note that $\mathbb{E}[\ell_i | \mathcal{G}_i] = 0$. Meanwhile, by (2.7.3), the inequalities

$$|\ell_i| \leq R\beta_{i-1} \min\{1, 2w_i\} \leq R\beta_{i-1} \leq R\beta_t \tag{2.7.5}$$

almost surely hold (the last inequality follows since $\{\beta_i\}_i$ is increasing). We also have

$$\begin{aligned}\sum_{i=1}^{t}\mathbb{E}[\ell_i^2|\mathcal{G}_i] &\le \sigma^2\sum_{i=1}^{t}\left(\frac{2\mathbf{x}_i^\top\mathbf{Z}_{i-1}^{-1}\mathbf{d}_{i-1}}{1+w_i^2}\mathcal{E}_{i-1}\right)^2\\ &\le \sigma^2\sum_{i=1}^{t}[\min\{1,2w_i\}\beta_{i-1}]^2\\ &\le 4\sigma^2\beta_t^2\sum_{i=1}^{t}\min\{1,w_i^2\}\\ &\le 8\sigma^2\beta_t^2 d\log(1+tL^2/(d\lambda)),\end{aligned} \tag{2.7.6}$$

where the first inequality holds since $\mathbb{E}[\eta_i^2|\mathcal{G}_i]\le\sigma^2$, the second inequality holds due to (2.7.3), the third inequality holds again since $\{\beta_i\}_i$ is increasing, the last inequality holds due to Lemma 2.7.2. Therefore, by (2.7.5) and (2.7.6), using Lemma 2.7.1, we know that for any t, with probability at least $1-\delta/(4t^2)$, we have

$$\begin{aligned}\sum_{i=1}^{t}\ell_i &\le \sqrt{16\sigma^2\beta_t^2 d\log(1+tL^2/(d\lambda))\log(4t^2/\delta)}+2/3\cdot R\beta_t\log(4t^2/\delta)\\ &\le \frac{\beta_t^2}{4}+16\sigma^2 d\log(1+tL^2/(d\lambda))\log(4t^2/\delta)+\frac{\beta_t^2}{4}+4R^2\log^2(4t^2/\delta)\\ &\le \beta_t^2/2+\frac{1}{4}\big(8\sigma\sqrt{d\log(1+tL^2/(d\lambda))\log(4t^2/\delta)}+4R\log(4t^2/\delta)\big)^2\\ &= 3\beta_t^2/4,\end{aligned} \tag{2.7.7}$$

where the first inequality holds due to Lemma 2.7.1, the second inequality holds due to $2\sqrt{|ab|}\le|a|+|b|$, the last equality holds due to the definition of β_t. Taking union bound for (2.7.7) from $t=1$ to ∞ and using the fact that $\sum_{t=1}^{\infty}t^{-2}<2$ finishes the proof. □

We also need the following lemma.

Lemma 2.7.4. Let w_i be as defined in (2.7.2). Then, with probability at least $1-\delta/2$, simultaneously for all $t\ge 1$ it holds that

$$\sum_{i=1}^{t}\frac{\eta_i^2 w_i^2}{1+w_i^2}\le\beta_t^2/4.$$

Proof. We are preparing to apply Freedman's inequality (Lemma 2.7.1) to $\{\ell_i\}_i$ and $\{\mathcal{G}_i\}_i$ where now

$$\ell_i = \frac{\eta_i^2 w_i^2}{1+w_i^2} - \mathbb{E}\left[\frac{\eta_i^2 w_i^2}{1+w_i^2}\middle|\mathcal{G}_i\right]. \tag{2.7.8}$$

Clearly, for any i, we have $\mathbb{E}[\ell_i|\mathcal{G}_i] = 0$ almost surely (a.s.). We further have that a.s.

$$\begin{aligned}
\sum_{i=1}^{t} \mathbb{E}[\ell_i^2|\mathcal{G}_i] &\le \sum_{i=1}^{t} \mathbb{E}\left[\frac{\eta_i^4 w_i^4}{(1+w_i^2)^2}\middle|\mathcal{G}_i\right] \\
&\le R^2 \sum_{i=1}^{t} \mathbb{E}\left[\frac{\eta_i^2 w_i^2}{1+w_i^2}\middle|\mathcal{G}_i\right] \\
&\le R^2\sigma^2 \sum_{i=1}^{t} \frac{w_i^2}{1+w_i^2} \\
&\le 2R^2\sigma^2 d\log(1+tL^2/(d\lambda)),
\end{aligned} \tag{2.7.9}$$

where the first inequality holds due to the fact $\mathbb{E}(X-\mathbb{E}X)^2 \le \mathbb{E}X^2$, the second inequality holds since $|\eta_t| \le R$ a.s., the third inequality holds since $\mathbb{E}[\eta_i^2|\mathcal{G}_i] \le \sigma^2$ a.s. and w_i is $\mathcal{G}_i$-measurable, the fourth inequality holds due to the fact $w_i^2/(1+w_i^2) \le \min\{1, w_i^2\}$ and Lemma 2.7.2. Furthermore, by the fact that $|\eta_i| \le R$ a.s., we have

$$|\ell_i| \le \left|\frac{\eta_i^2 w_i^2}{1+w_i^2}\right| + \left|\mathbb{E}\left[\frac{\eta_i^2 w_i^2}{1+w_i^2}\middle|\mathcal{G}_i\right]\right| \le 2R^2 \text{ a.s.} \tag{2.7.10}$$

Therefore, by (2.7.9) and (2.7.10), using Lemma 2.7.1, we know that for any t, with probability at least $1-\delta/(4t^2)$, we have that a.s.,

$$\begin{aligned}
\sum_{i=1}^{t} \frac{\eta_i^2 w_i^2}{1+w_i^2} &\le \sum_{i=1}^{t} \mathbb{E}\left[\frac{\eta_i^2 w_i^2}{1+w_i^2}\middle|\mathcal{G}_i\right] + \sqrt{4R^2\sigma^2 d\log(1+tL^2/(d\lambda))\log(4t^2/\delta)} + 4/3\cdot R^2\log(4t^2/\delta) \\
&\le \sigma^2 \sum_{i=1}^{t} \frac{w_i^2}{1+w_i^2} + 2R\sigma\sqrt{d\log(1+tL^2/(d\lambda))\log(4t^2/\delta)} + 2R^2\log(4t^2/\delta) \\
&\le 2\sigma^2 d\log(1+tL^2/(d\lambda)) + 2R\sigma\sqrt{d\log(1+tL^2/(d\lambda))\log(4t^2/\delta)} + 2R^2\log(4t^2/\delta) \\
&\le 1/4\cdot\left(8\sigma\sqrt{d}\sqrt{\log(1+tL^2/(d\lambda))\log(4t^2/\delta)} + 4R\log(4t^2/\delta)\right)^2 \\
&= \beta_t^2/4,
\end{aligned} \tag{2.7.11}$$

where the first inequality holds due to Lemma 2.7.1, the second inequality holds due to $\mathbb{E}[\eta_i^2|\mathcal{G}_i] \leq \sigma^2$, the third inequality holds due to the fact $w_i^2/(1+w_i^2) \leq \min\{1, w_i^2\}$ and Lemma 2.7.2, the last inequality holds due to the definition of β_t. Taking union bound for (2.7.11) from $t=1$ to ∞ and using the fact that $\sum_{t=1}^{\infty} t^{-2} < 2$ finishes the proof. □

With this, we are ready to prove Theorem 2.4.1.

Proof of Theorem 2.4.1. We first give a crude upper bound on Z_t. We have

$$\begin{aligned} Z_t^2 &= (\mathbf{d}_{t-1} + \mathbf{x}_t\eta_t)^\top \mathbf{Z}_t^{-1}(\mathbf{d}_{t-1} + \mathbf{x}_t\eta_t) \\ &= \mathbf{d}_{t-1}^\top \mathbf{Z}_t^{-1}\mathbf{d}_{t-1} + 2\eta_t \mathbf{x}_t^\top \mathbf{Z}_t^{-1}\mathbf{d}_{t-1} + \eta_t^2 \mathbf{x}_t^\top \mathbf{Z}_t^{-1}\mathbf{x}_t \\ &\leq Z_{t-1}^2 + \underbrace{2\eta_t \mathbf{x}_t^\top \mathbf{Z}_t^{-1}\mathbf{d}_{t-1}}_{I_1} + \underbrace{\eta_t^2 \mathbf{x}_t^\top \mathbf{Z}_t^{-1}\mathbf{x}_t}_{I_2}, \end{aligned}$$

where the inequality holds since $\mathbf{Z}_t \succeq \mathbf{Z}_{t-1}$. For term I_1, from the matrix inversion lemma (cf. (2.7.1)), we have

$$\begin{aligned} I_1 &= 2\eta_t \left(\mathbf{x}_t^\top \mathbf{Z}_{t-1}^{-1}\mathbf{d}_{t-1} - \frac{\mathbf{x}_t^\top \mathbf{Z}_{t-1}^{-1}\mathbf{x}_t \mathbf{x}_t^\top \mathbf{Z}_{t-1}^{-1}\mathbf{d}_{t-1}}{1+w_t^2} \right) \\ &= 2\eta_t \left(\mathbf{x}_t^\top \mathbf{Z}_{t-1}^{-1}\mathbf{d}_{t-1} - \frac{w_t^2 \mathbf{x}_t^\top \mathbf{Z}_{t-1}^{-1}\mathbf{d}_{t-1}}{1+w_t^2} \right) \\ &= \frac{2\eta_t \mathbf{x}_t^\top \mathbf{Z}_{t-1}^{-1}\mathbf{d}_{t-1}}{1+w_t^2}\,. \end{aligned}$$

For term I_2, again from the matrix inversion lemma (cf. (2.7.1)), we have

$$I_2 = \eta_t^2 \left(\mathbf{x}_t^\top \mathbf{Z}_{t-1}^{-1}\mathbf{x}_t^\top - \frac{\mathbf{x}_t^\top \mathbf{Z}_{t-1}^{-1}\mathbf{x}_t \mathbf{x}_t^\top \mathbf{Z}_{t-1}^{-1}\mathbf{x}_t}{1+w_t^2} \right) = \eta_t^2 \left(w_t^2 - \frac{w_t^4}{1+w_t^2} \right) = \frac{\eta_t^2 w_t^2}{1+w_t^2}\,.$$

Therefore, we have

$$Z_t^2 \leq \sum_{i=1}^{t} \frac{2\eta_i \mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1}\mathbf{d}_{i-1}}{1+w_i^2} + \sum_{i=1}^{t} \frac{\eta_i^2 w_i^2}{1+w_i^2}. \tag{2.7.12}$$

Consider now the event $\mathcal{E}$ where the conclusions of Lemma 2.7.3 and Lemma 2.7.4 hold. We claim that on this event for any $i \geq 0$, $Z_i \leq \beta_i$. We prove this by induction on i. Let the said event hold. The base case of $i=0$ holds since $\beta_0 = 0 = Z_0$, by definition. Now

fix some $t \geq 1$ and assume that for all $0 \leq i < t$, we have $Z_i \leq \beta_i$. This implies that $\mathcal{E}_1 = \mathcal{E}_2 = \cdots = \mathcal{E}_{t-1} = 1$. Then by (2.7.12), we have

$$Z_t^2 \leq \sum_{i=1}^{t} \frac{2\eta_i \mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1} \mathbf{d}_{i-1}}{1+w_i^2} + \sum_{i=1}^{t} \frac{\eta_i^2 w_i^2}{1+w_i^2} = \sum_{i=1}^{t} \frac{2\eta_i \mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1} \mathbf{d}_{i-1}}{1+w_i^2} \mathcal{E}_{i-1} + \sum_{i=1}^{t} \frac{\eta_i^2 w_i^2}{1+w_i^2}. \quad (2.7.13)$$

Since on the event $\mathcal{E}$ the conclusions of Lemma 2.7.3 and Lemma 2.7.4 hold, we have

$$\sum_{i=1}^{t} \frac{2\eta_i \mathbf{x}_i^\top \mathbf{Z}_{i-1}^{-1} \mathbf{d}_{i-1}}{1+w_i^2} \mathcal{E}_{i-1} \leq 3\beta_t^2/4, \ \sum_{i=1}^{t} \frac{\eta_i^2 w_i^2}{1+w_i^2} \leq \beta_t^2/4. \quad (2.7.14)$$

Therefore, substituting (2.7.14) into (2.7.13), we have $Z_t \leq \beta_t$, which ends the induction. Taking the union bound, the events in Lemma 2.7.3 and Lemma 2.7.4 hold with probability at least $1-\delta$, which implies that with probability at least $1-\delta$, for any t, $Z_t \leq \beta_t$.

Finally, we bound $\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t}$ as follows. First,

$$\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1}\mathbf{b}_t = \mathbf{Z}_t^{-1} \sum_{i=1}^{t} \mathbf{x}_i(\mathbf{x}_i^\top \boldsymbol{\mu}^* + \eta_i) = \boldsymbol{\mu}^* - \lambda \mathbf{Z}_t^{-1}\boldsymbol{\mu}^* + \mathbf{Z}_t^{-1}\mathbf{d}_t\,.$$

Then, on $\mathcal{E}$ we have

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} = \left\|\mathbf{d}_t - \lambda\boldsymbol{\mu}^*\right\|_{\mathbf{Z}_t^{-1}} \leq Z_t + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2 \leq \beta_t + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2, \quad (2.7.15)$$

where the first inequality holds due to triangle inequality and $\mathbf{Z}_t \succeq \lambda\mathbf{I}$, while the last one holds since we have shown that on $\mathcal{E}$, $Z_t \leq \beta_t$ for all $t \geq 0$, thus finishing the proof.

□

2.7.2 Proof of Theorem 2.4.2

Proof of Theorem 2.4.2. By the assumption on ϵ_t, we know that

$$|\epsilon_t/\bar{\sigma}_t| \leq R/\bar{\sigma}_{\min}^t, \ \mathbb{E}[\epsilon_t|\mathbf{a}_{1:t}, \epsilon_{1:t-1}] = 0, \ \mathbb{E}[(\epsilon_t/\bar{\sigma}_t)^2|\mathbf{a}_{1:t}, \epsilon_{1:t-1}] \leq 1, \ \|\mathbf{a}_t/\bar{\sigma}_t\|_2 \leq A/\bar{\sigma}_{\min}^t,$$

Then, taking $\mathcal{G}_t = \sigma(\mathbf{a}_{1:t}, \epsilon_{1:t-1})$, using that σ_t is $\mathcal{G}_t$-measurable, we can apply Theorem 2.4.1 to $(\boldsymbol{x}_t, \eta_t) = (\boldsymbol{a}_t/\sigma_t, \epsilon_t/\sigma_t)$ to get that with probability at least $1-\delta$,

$$\forall t \geq 1, \ \left\|\widehat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}^*\right\|_{\mathbf{A}_t} \leq \widehat{\beta}_t + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2 \leq \widehat{\beta}_t + \sqrt{\lambda}B, \quad (2.7.16)$$

where $\widehat{\beta}_t = 8\sqrt{d\log(1+tA^2/([\bar{\sigma}^t_{\min}]^2 d\lambda))\log(4t^2/\delta)} + 4R/\bar{\sigma}^t_{\min} \cdot \log(4t^2/\delta)$. Thus, in the remainder of the proof, we will assume that the event $\mathcal{E}$ when (2.7.16) is true holds and proceed to bound the regret on this event.

Note that on $\mathcal{E}$, $\boldsymbol{\mu}^* \in \mathcal{C}_t$. Recall that $\widetilde{\boldsymbol{\mu}}_t$ is the optimistic parameter choice of the algorithm (cf. Line 4 in Algorithm 1). Then, using the standard argument for linear bandits, the pseudo-regret for round t is bounded by

$$\langle \mathbf{a}^*_t, \boldsymbol{\mu}^* \rangle - \langle \mathbf{a}_t, \boldsymbol{\mu}^* \rangle \leq \langle \mathbf{a}_t, \widetilde{\boldsymbol{\mu}}_t \rangle - \langle \mathbf{a}_t, \boldsymbol{\mu}^* \rangle = \langle \mathbf{a}_t, \widetilde{\boldsymbol{\mu}}_t - \widehat{\boldsymbol{\mu}}_{t-1} \rangle + \langle \mathbf{a}_t, \widehat{\boldsymbol{\mu}}_{t-1} - \boldsymbol{\mu}^* \rangle, \tag{2.7.17}$$

where the inequality holds due to the choice $\widetilde{\boldsymbol{\mu}}_t$. To further bound (2.7.17), we have

$$\begin{aligned}
&\langle \mathbf{a}_t, \widetilde{\boldsymbol{\mu}}_t - \widehat{\boldsymbol{\mu}}_{t-1} \rangle + \langle \mathbf{a}_t, \widehat{\boldsymbol{\mu}}_{t-1} - \boldsymbol{\mu}^* \rangle \\
&\quad \leq \|\mathbf{a}_t\|_{\mathbf{A}^{-1}_{t-1}} (\|\widetilde{\boldsymbol{\mu}}_t - \widehat{\boldsymbol{\mu}}_{t-1}\|_{\mathbf{A}_{t-1}} + \|\boldsymbol{\mu}^* - \widehat{\boldsymbol{\mu}}_{t-1}\|_{\mathbf{A}_{t-1}}) \\
&\quad \leq 2(\widehat{\beta}_{t-1} + \sqrt{\lambda} B) \|\mathbf{a}_t\|_{\mathbf{A}^{-1}_{t-1}},
\end{aligned} \tag{2.7.18}$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second one holds since $\widetilde{\boldsymbol{\mu}}_t, \boldsymbol{\mu}^* \in \mathcal{C}_{t-1}$. Meanwhile, we have $0 \leq \langle \mathbf{a}^*_t, \boldsymbol{\mu}^* \rangle - \langle \mathbf{a}_t, \boldsymbol{\mu}^* \rangle \leq 2$. Thus, substituting (2.7.18) into (2.7.17) and summing up (2.7.17) for $t = 1, \ldots, T$, we have

$$\text{Regret}(T) = \sum_{t=1}^{T} \left[\langle \mathbf{a}^*_t, \boldsymbol{\mu}^* \rangle - \langle \mathbf{a}_t, \boldsymbol{\mu}^* \rangle \right] \leq 2 \sum_{t=1}^{T} \min \left\{ 1, \bar{\sigma}_t (\widehat{\beta}_{t-1} + \sqrt{\lambda} B) \|\mathbf{a}_t / \bar{\sigma}_t\|_{\mathbf{A}^{-1}_{t-1}} \right\}. \tag{2.7.19}$$

To further bound the right-hand side above, we decompose the set $[T]$ into a union of two disjoint subsets $[T] = \mathcal{I}_1 \cup \mathcal{I}_2$, where

$$\mathcal{I}_1 = \left\{ t \in [T] : \|\mathbf{a}_t / \bar{\sigma}_t\|_{\mathbf{A}^{-1}_{t-1}} \geq 1 \right\}, \ \mathcal{I}_2 = [T] \setminus \mathcal{I}_1. \tag{2.7.20}$$

Then the following upper bound of $|\mathcal{I}_1|$ holds:

$$|\mathcal{I}_1| \leq \sum_{t \in \mathcal{I}_1} \min \left\{ 1, \|\mathbf{a}_t / \bar{\sigma}_t\|^2_{\mathbf{A}^{-1}_{t-1}} \right\} \leq \sum_{t=1}^{T} \min \left\{ 1, \|\mathbf{a}_t / \bar{\sigma}_t\|^2_{\mathbf{A}^{-1}_{t-1}} \right\} \leq 2d \log(1 + TA^2/(d\lambda [\bar{\sigma}^T_{\min}]^2)), \tag{2.7.21}$$

where the first inequality holds since $\|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}} \geq 1$ for $t \in \mathcal{I}_1$, the third inequality holds due to Lemma 2.7.2 together with the fact $\|\mathbf{a}_t/\bar{\sigma}_t\|_2 \leq A/\bar{\sigma}_{\min}^T$. Therefore, by (2.7.19),

$$\begin{aligned}
&\text{Regret}(T)/2 = \\
&\sum_{t\in\mathcal{I}_1} \min\left\{1, \bar{\sigma}_t(\widehat{\beta}_{t-1} + \sqrt{\lambda}B)\|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}}\right\} + \sum_{t\in\mathcal{I}_2} \min\left\{1, \bar{\sigma}_t(\widehat{\beta}_{t-1} + \sqrt{\lambda}B)\|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}}\right\} \\
&\leq \left[\sum_{t\in\mathcal{I}_1} 1\right] + \sum_{t\in\mathcal{I}_2} (\widehat{\beta}_{t-1} + \sqrt{\lambda}B)\bar{\sigma}_t\|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}} \\
&= |\mathcal{I}_1| + \sum_{t\in\mathcal{I}_2} (\widehat{\beta}_{t-1} + \sqrt{\lambda}B)\bar{\sigma}_t \min\left\{1, \|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}}\right\} \\
&\leq 2d\log(1 + TA^2/(d\lambda[\bar{\sigma}_{\min}^T]^2)) + \sum_{t=1}^{T} (\widehat{\beta}_{t-1} + \sqrt{\lambda}B)\bar{\sigma}_t \min\left\{1, \|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}}\right\}, \qquad (2.7.22)
\end{aligned}$$

where the first inequality holds since for any x real, $\min\{1, x\} \leq 1$ and also $\min\{1, x\} \leq x$, the second inequality holds since $\|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}} \leq 1$ for $t \in \mathcal{I}_2$ and the last one holds due to (2.7.21). Finally, to further bound (2.7.22), notice that

$$\begin{aligned}
&\sum_{t=1}^{T} (\widehat{\beta}_{t-1} + \sqrt{\lambda}B)\bar{\sigma}_t \min\left\{1, \|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}}\right\} \\
&\quad\leq \sqrt{\sum_{t=1}^{T} (\widehat{\beta}_{t-1} + \sqrt{\lambda}B)^2\bar{\sigma}_t^2}\sqrt{\sum_{t=1}^{T} \min\left\{1, \|\mathbf{a}_t/\bar{\sigma}_t\|_{\mathbf{A}_{t-1}^{-1}}^2\right\}} \\
&\quad\leq \sqrt{\sum_{t=1}^{T} (\widehat{\beta}_{t-1} + \sqrt{\lambda}B)^2\bar{\sigma}_t^2}\sqrt{2d\log(1 + TA^2/(d\lambda[\bar{\sigma}_{\min}^T]^2))}, \qquad (2.7.23)
\end{aligned}$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second one holds due to Lemma 2.7.2 and the the fact that $\|\mathbf{a}_t/\sigma_t\|_2 \leq A/\bar{\sigma}_{\min}^T$. Substituting (2.7.23) into (2.7.22), we have

$$\begin{aligned}
\text{Regret}(T) \leq\ & 2\sqrt{2d\log(1 + TA^2/(d\lambda[\bar{\sigma}_{\min}^T]^2))}\sqrt{\sum_{t=1}^{T} (\widehat{\beta}_{t-1} + \sqrt{\lambda}B)^2\bar{\sigma}_t^2} \\
&+ 4d\log\left(1 + TA^2/(d\lambda[\bar{\sigma}_{\min}^T]^2)\right), \qquad (2.7.24)
\end{aligned}$$

Next, since $\bar{\sigma}_t = \max\{R/\sqrt{d}, \sigma_t\}$, then we have $\bar{\sigma}_{\min}^t \geq R/\sqrt{d}$. Therefore, with $\lambda = 1/B^2$, we

have

$$\log(1 + TA^2/(d\lambda[\bar{\sigma}^T_{\min}]^2)) \leq \log(1 + TB^2A^2/R^2) = \widetilde{O}(1), \tag{2.7.25}$$

and

$$\begin{aligned}
\widehat{\beta}_t + \sqrt{\lambda}B &= 8\sqrt{d\log(1 + tA^2/([\bar{\sigma}^t_{\min}]^2 d\lambda))\log(4t^2/\delta)} + 4R/\bar{\sigma}^t_{\min} \cdot \log(4t^2/\delta) + \sqrt{\lambda}B \\
&\leq 8\sqrt{d\log(1 + TB^2A^2/R^2)\log(4T^2/\delta)} + 4\sqrt{d}\log(4T^2/\delta) + 1 \\
&= \widetilde{O}(\sqrt{d}).
\end{aligned} \tag{2.7.26}$$

Substituting (2.7.25) and (2.7.26) into (2.7.24), we have our second result.

$$\text{Regret}(T) = \widetilde{O}\bigg(d\sqrt{\sum_{t=1}^{T}\bar{\sigma}_t^2}\bigg) = \widetilde{O}\bigg(d\sqrt{\sum_{t=1}^{T}(R^2/d + \sigma_t^2)}\bigg) = \widetilde{O}\bigg(R\sqrt{dT} + d\sqrt{\sum_{t=1}^{T}\sigma_t^2}\bigg),$$

where the second equality holds since $\bar{\sigma}_t^2 = \max\{R^2/d, \sigma_t^2\} \leq R^2/d + \sigma_t^2$, the third equality holds since $\sqrt{|x| + |y|} \leq \sqrt{|x|} + \sqrt{|y|}$. □

2.8 Proof of Upper Bound Results in Section 2.5

Let $\mathbb{P}$ be the distribution over $(\mathcal{S} \times \mathcal{A})^{\mathbb{N}}$ induced by the interconnection of UCRL-VTR$^+$ (treated as a nonstationary, history dependent policy) and the episodic MDP M. Further, let $\mathbb{E}$ be the corresponding expectation operator. Note that the only source of randomness are the stochastic transitions in the MDP, hence, all random variables can be defined over the sample space $\Omega = (\mathcal{S} \times \mathcal{A})^{\mathbb{N}}$. Thus, we work with the probability space given by the triplet $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ is the product σ-algebra generated by the discrete σ-algebras underlying $\mathcal{S}$ and $\mathcal{A}$, respectively.

For $1 \leq k \leq K$, $1 \leq h \leq H$, let $\mathcal{F}_{k,h}$ be the σ-algebra generated by the random variables representing the state-action pairs up to and including those that appear stage h of episode k. That is, $\mathcal{F}_{k,h}$ is generated by

$$s_1^1, a_1^1, \ldots, s_h^1, a_h^1, \ldots, s_H^1, a_H^1\,,$$

$$s_1^2, a_1^2, \ldots, s_h^2, a_h^2, \ldots, s_H^2, a_H^2\,,$$
$$\vdots$$
$$s_1^k, a_1^k, \ldots, s_h^k, a_h^k\,.$$

Note that, by construction,

$$\bar{\mathbb{V}}_{k,h}V_{k,h+1}(s_h^k, a_h^k), E_{k,h}, \bar{\sigma}_{k,h}, \widehat{\boldsymbol{\Sigma}}_{k+1,h}, \widetilde{\boldsymbol{\Sigma}}_{k+1,h},$$

are $\mathcal{F}_{k,h}$-measurable, $\widehat{\boldsymbol{b}}_{k+1,h}, \widetilde{\boldsymbol{b}}_{k+1,h}, \widehat{\boldsymbol{\theta}}_{k+1,h}, \widetilde{\boldsymbol{\theta}}_{k+1,h}$ are $\mathcal{F}_{k,h+1}$-measurable, and $Q_{k,h}, V_{k,h}, \pi_h^k, \phi_{V_{k,h+1}}$ are $\mathcal{F}_{k-1,H}$ measurable. Note also that $Q_{k,h}, V_{k,h}, \pi_h^k, \phi_{V_{k,h+1}}$ are *not* $\mathcal{F}_{k-1,h}$ measurable: The get their values only after episode $k-1$ is *over*, due to their "backwards" construction.

2.8.1 Proof of Lemma 2.5.1

The main idea of the proof is to use a (crude) two-step, "peeling" device. Let $\check{\mathcal{C}}_{k,h}, \widetilde{\mathcal{C}}_{k,h}$ denote the following confidence sets:

$$\check{\mathcal{C}}_{k,h} = \left\{\boldsymbol{\theta} : \left\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{k,h})\right\|_2 \le \check{\beta}_k\right\},\ \widetilde{\mathcal{C}}_{k,h} = \left\{\boldsymbol{\theta} : \left\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{1/2}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_{k,h})\right\|_2 \le \widetilde{\beta}_k\right\}.$$

Note that $\widehat{\mathcal{C}}_{k,h} \subset \check{\mathcal{C}}_{k,h}$: The "leading term" in the definition of $\check{\beta}_k$ is larger than that in $\widehat{\beta}_k$ by a factor of $\sqrt{d}$. The idea of our proof is to show that $\boldsymbol{\theta}_h^*$ is included in $\check{\mathcal{C}}_{k,h} \cap \widetilde{\mathcal{C}}_{k,h}$ with high probability (for this, a standard self-normalized tail inequality suffices) and then use that when this holds, the weights used in constructing $\widehat{\boldsymbol{\theta}}_{k,h}$ are sufficiently precise to "balance" the noise term, which allows to reduce $\check{\beta}_k$ by the extra $\sqrt{d}$ factor without significantly increasing the probability of the bad event when $\boldsymbol{\theta}_h^* \notin \widehat{\mathcal{C}}_{k,h}$.

We start with the following lemma.

Lemma 2.8.1. Let $V_{k,h+1}, \widehat{\boldsymbol{\theta}}_{k,h}, \widehat{\boldsymbol{\Sigma}}_{k,h}, \widetilde{\boldsymbol{\theta}}_{k,h}, \widetilde{\boldsymbol{\Sigma}}_{k,h}$ be defined in Algorithm 2, then we have

$$\begin{aligned}
&\left|\mathbb{V}_h V_{k,h+1}(s_h^k, a_h^k) - \bar{\mathbb{V}}_{k,h}V_{k,h+1}(s_h^k, a_h^k)\right| \\
&\quad\le \min\left\{H^2, \left\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\right\|_2 \left\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{1/2}\big(\widetilde{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^*\big)\right\|_2\right\} \\
&\qquad + \min\left\{H^2, 2H\left\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\right\|_2 \left\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2}\big(\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^*\big)\right\|_2\right\}.
\end{aligned}$$

Proof. We have

$$
\begin{aligned}
&\left|[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k)-[\mathbb{V}_hV_{k,h+1}](s_h^k,a_h^k)\right| \\
&=\Big|\left[\left\langle\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k),\widetilde{\boldsymbol{\theta}}_{k,h}\right\rangle\right]_{[0,H^2]}-\left\langle\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k),\boldsymbol{\theta}_h^*\right\rangle \\
&\qquad+\left(\left\langle\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k),\boldsymbol{\theta}_h^*\right\rangle\right)^2-\left[\left\langle\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k),\widehat{\boldsymbol{\theta}}_{k,h}\right\rangle\right]_{[0,H]}^2\Big| \\
&\le\underbrace{\left|\left[\left\langle\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k),\widetilde{\boldsymbol{\theta}}_{k,h}\right\rangle\right]_{[0,H^2]}-\left\langle\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k),\boldsymbol{\theta}_h^*\right\rangle\right|}_{I_1} \\
&\qquad+\underbrace{\left|\left(\left\langle\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k),\boldsymbol{\theta}_h^*\right\rangle\right)^2-\left[\left\langle\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k),\widehat{\boldsymbol{\theta}}_{k,h}\right\rangle\right]_{[0,H]}^2\right|}_{I_2},
\end{aligned}
$$

where the inequality holds due to the triangle inequality. We bound I_1 first. We have $I_1\le H^2$ since both terms in I_1 belong to the interval $[0,H^2]$. Furthermore,

$$
\begin{aligned}
I_1&\le\left|\left\langle\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k),\widetilde{\boldsymbol{\theta}}_{k,h}\right\rangle-\left\langle\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k),\boldsymbol{\theta}_h^*\right\rangle\right| \\
&=\left|\left\langle\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k),\widetilde{\boldsymbol{\theta}}_{k,h}-\boldsymbol{\theta}_h^*\right\rangle\right| \\
&\le\left\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k)\right\|_2\left\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{1/2}\big(\widetilde{\boldsymbol{\theta}}_{k,h}-\boldsymbol{\theta}_h^*\big)\right\|_2,
\end{aligned}
$$

where the first inequality holds since $\langle\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k),\boldsymbol{\theta}_h^*\rangle\in[0,H^2]$ and the second inequality holds due to the Cauchy-Schwarz inequality. Thus, we have

$$
I_1\le\min\left\{H^2,\left\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k)\right\|_2\left\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{1/2}\big(\widetilde{\boldsymbol{\theta}}_{k,h}-\boldsymbol{\theta}_h^*\big)\right\|_2\right\}. \tag{2.8.1}
$$

For the term I_2, since both terms in I_2 belong to the interval $[0,H^2]$, we have $I_2\le H^2$. Meanwhile,

$$
\begin{aligned}
I_2&=\left|\left\langle\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k),\boldsymbol{\theta}_h^*\right\rangle+\left[\left\langle\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k),\widehat{\boldsymbol{\theta}}_{k,h}\right\rangle\right]_{[0,H]}\right| \\
&\qquad\cdot\left|\left\langle\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k),\boldsymbol{\theta}_h^*\right\rangle-\left[\left\langle\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k),\widehat{\boldsymbol{\theta}}_{k,h}\right\rangle\right]_{[0,H]}\right| \\
&\le 2H\left|\left\langle\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k),\boldsymbol{\theta}_h^*\right\rangle-\left\langle\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k),\widehat{\boldsymbol{\theta}}_{k,h}\right\rangle\right| \\
&=2H\left|\left\langle\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k),\boldsymbol{\theta}_h^*-\widehat{\boldsymbol{\theta}}_{k,h}\right\rangle\right| \\
&\le 2H\left\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\right\|_2\left\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2}\big(\widehat{\boldsymbol{\theta}}_{k,h}-\boldsymbol{\theta}_h^*\big)\right\|_2,
\end{aligned}
\tag{2.8.2}
$$

where the first inequality holds since both terms in this line are less than H and the fact $\left\langle \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k), \boldsymbol{\theta}_h^* \right\rangle \in [0, H]$, the second inequality holds due to the Cauchy-Schwarz inequality. Thus, we have

$$I_2 \leq \min\left\{ H^2, 2H \left\| \widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2} \boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k) \right\|_2 \left\| \widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2} \big(\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}_h^* \big) \right\|_2 \right\}. \tag{2.8.3}$$

Combining (2.8.1) and (2.8.3) gives the desired result. □

Proof of Lemma 2.5.1. Fix $h \in [H]$. We first show that with probability at least $1 - \delta/H$, for all k, $\boldsymbol{\theta}_h^* \in \check{\mathcal{C}}_{k,h}$. To show this, we apply Theorem 2.4.1. Let $\mathbf{x}_i = \bar{\sigma}_{i,h}^{-1} \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)$ and $\eta_i = \bar{\sigma}_{i,h}^{-1} V_{i,h+1}(s_{h+1}^i) - \bar{\sigma}_{i,h}^{-1} \langle \boldsymbol{\phi}_{V_{i,h+1}}(s_{i,h}, a_{i,h}), \boldsymbol{\theta}_h^* \rangle$, $\mathcal{G}_i = \mathcal{F}_{i,h}$, $\boldsymbol{\mu}^* = \boldsymbol{\theta}_h^*$, $y_i = \langle \boldsymbol{\mu}^*, \mathbf{x}_i \rangle + \eta_i$, $\mathbf{Z}_i = \lambda \mathbf{I} + \sum_{i'=1}^{i} \mathbf{x}_{i'} \mathbf{x}_{i'}^\top$, $\mathbf{b}_i = \sum_{i'=1}^{i} \mathbf{x}_{i'} y_{i'}$ and $\boldsymbol{\mu}_i = \mathbf{Z}_i^{-1} \mathbf{b}_i$. Then it can be verified that $y_i = \bar{\sigma}_{i,h}^{-1} V_{i,h+1}(s_{h+1}^i)$ and $\boldsymbol{\mu}_i = \widehat{\boldsymbol{\theta}}_{i+1,h}$. Moreover, almost surely,

$$\|\mathbf{x}_i\|_2 \leq \bar{\sigma}_{i,h}^{-1} H \leq \sqrt{d}, \quad |\eta_i| \leq \bar{\sigma}_{i,h}^{-1} H \leq \sqrt{d}, \quad \mathbb{E}[\eta_i | \mathcal{G}_i] = 0, \quad \mathbb{E}[\eta_i^2 | \mathcal{G}_i] \leq d\,,$$

where we used that $V_{i,h+1}$ takes values in $[0, H]$ and that $\|\boldsymbol{\phi}_{V_{i,h+1}}(s, a)\|_2 \leq H$ by (2.3.1). Since we also have that $\boldsymbol{x}_i$ is $\mathcal{G}_i$ measurable and η_i is $\mathcal{G}_{i+1}$ measurable, by Theorem 2.4.1, we obtain that with probability at least $1 - \delta/H$, for all $k \leq K$,

$$\big\| \boldsymbol{\theta}_h^* - \widehat{\boldsymbol{\theta}}_{k,h} \big\|_{\widehat{\boldsymbol{\Sigma}}_{k,h}} \leq 8d \sqrt{\log(1 + k/\lambda) \log(4k^2 H/\delta)} + 4\sqrt{d} \log(4k^2 H/\delta) + \sqrt{\lambda} B = \check{\beta}_k, \tag{2.8.4}$$

implying that with probability $1 - \delta/H$, for any $k \leq K$, $\boldsymbol{\theta}_h^* \in \check{\mathcal{C}}_{k,h}$.

An argument, which is analogous to the one just used (except that now the range of the "noise" matches the range of "squared values" and is thus bounded by H^2, rather than being bounded by $\sqrt{d}$) gives that with probability at least $1 - \delta/H$, for any $k \leq K$ we have

$$\big\| \boldsymbol{\theta}_h^* - \widetilde{\boldsymbol{\theta}}_{k,h} \big\|_{\widetilde{\boldsymbol{\Sigma}}_{k,h}} \leq 8 \sqrt{dH^4 \log(1 + kH^4/(d\lambda)) \log(4k^2 H/\delta)} + 4H^2 \log(4k^2 H/\delta) + \sqrt{\lambda} B = \widetilde{\beta}_k, \tag{2.8.5}$$

which implies that with the said probability, $\boldsymbol{\theta}_h^* \in \widetilde{\mathcal{C}}_{k,h}$.

We now show that $\boldsymbol{\theta}_h^* \in \widehat{\mathcal{C}}_{k,h}$ with high probability. We again apply Theorem 2.4.1. Let $\mathbf{x}_i = \bar{\sigma}_{i,h}^{-1}\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)$ and

$$\eta_i = \bar{\sigma}_{i,h}^{-1}\mathbb{1}\{\boldsymbol{\theta}_h^* \in \check{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h}\}\big[V_{i,h+1}(s_{h+1}^i) - \langle \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i), \boldsymbol{\theta}_h^*\rangle\big],$$

$\mathcal{G}_i = \mathcal{F}_{i,h}$, $\boldsymbol{\mu}^* = \boldsymbol{\theta}_h^*$. Clearly $\mathbb{E}[\eta_i|\mathcal{G}_i] = 0$, $|\eta_i| \le \bar{\sigma}_{i,h}^{-1}H \le \sqrt{d}$ since $|V_{i,h+1}(\cdot)| \le H$ and $\bar{\sigma}_{i,h} \ge H/\sqrt{d}$, $\|\mathbf{x}_i\|_2 \le \bar{\sigma}_{i,h}^{-1}H \le \sqrt{d}$. Furthermore, owning to that $\mathbb{1}\{\boldsymbol{\theta}_h^* \in \check{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h}\}$ is $\mathcal{G}_i$-measurable, it holds that

$$\begin{aligned}
\mathbb{E}[\eta_i^2|\mathcal{G}_i] &= \bar{\sigma}_{i,h}^{-2}\,\mathbb{1}\{\boldsymbol{\theta}_h^* \in \check{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h}\}[\mathbb{V}_h V_{i,h+1}](s_h^i, a_h^i) \\
&\le \bar{\sigma}_{i,h}^{-2}\,\mathbb{1}\{\boldsymbol{\theta}_h^* \in \check{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h}\}\Bigg[[\bar{\mathbb{V}}_{i,h}V_{i,h+1}](s_h^i, a_h^i) \\
&\qquad + \min\Big\{H^2, \Big\|\widetilde{\boldsymbol{\Sigma}}_{i,h}^{-1/2}\boldsymbol{\phi}_{V_{i,h+1}^2}(s_h^i, a_h^i)\Big\|_2\Big\|\widetilde{\boldsymbol{\Sigma}}_{i,h}^{1/2}\big(\widetilde{\boldsymbol{\theta}}_{i,h} - \boldsymbol{\theta}_h^*\big)\Big\|_2\Big\} \\
&\qquad + \min\Big\{H^2, 2H\Big\|\widehat{\boldsymbol{\Sigma}}_{i,h}^{-1/2}\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)\Big\|_2\Big\|\widehat{\boldsymbol{\Sigma}}_{i,h}^{1/2}\big(\widehat{\boldsymbol{\theta}}_{i,h} - \boldsymbol{\theta}_h^*\big)\Big\|_2\Big\}\Bigg] \\
&\le \bar{\sigma}_{i,h}^{-2}\Bigg[[\bar{\mathbb{V}}_{i,h}V_{i,h+1}](s_h^i, a_h^i) + \min\Big\{H^2, \widetilde{\beta}_i\Big\|\widetilde{\boldsymbol{\Sigma}}_{i,h}^{-1/2}\boldsymbol{\phi}_{V_{i,h+1}^2}(s_h^i, a_h^i)\Big\|_2\Big\} \\
&\qquad + \min\Big\{H^2, 2H\check{\beta}_i\Big\|\widehat{\boldsymbol{\Sigma}}_{i,h}^{-1/2}\boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)\Big\|_2\Big\}\Bigg] \\
&= 1,
\end{aligned}$$

where the first inequality holds due to Lemma 2.8.1, the second inequality holds due to the indicator function, the last equality holds due to the definition of $\bar{\sigma}_{i,h}$. Now, let $y_i = \langle \boldsymbol{\mu}^*, \mathbf{x}_i\rangle + \eta_i$, $\mathbf{Z}_i = \lambda\mathbf{I} + \sum_{i'=1}^i \mathbf{x}_{i'}\mathbf{x}_{i'}^\top$, $\mathbf{b}_i = \sum_{i'=1}^i \mathbf{x}_{i'}y_{i'}$ and $\boldsymbol{\mu}_i = \mathbf{Z}_i^{-1}\mathbf{b}_i$. Then, by Theorem 2.4.1, with probability at least $1 - \delta/H$, $\forall k \le K$,

$$\|\boldsymbol{\mu}_k - \boldsymbol{\mu}^*\|_{\mathbf{Z}_i} \le 8\sqrt{d\log(1 + k/\lambda)\log(4k^2H/\delta)} + 4\sqrt{d}\log(4k^2H/\delta) + \sqrt{\lambda}B = \widehat{\beta}_k, \quad (2.8.6)$$

where the equality uses the definition of $\widehat{\beta}_k$. Let $\mathcal{E}'$ be the event when $\boldsymbol{\theta}_h^* \in \cap_{k\le K}\check{\mathcal{C}}_{k,h} \cap \widetilde{\mathcal{C}}_{k,h}$ and (2.8.6) hold. By the union bound, $\mathbb{P}(\mathcal{E}') \ge 1 - 3\delta/H$.

We now show that $\boldsymbol{\theta}_h^* \in \widehat{\mathcal{C}}_{k,h}$ holds on $\mathcal{E}'$. For this note that on $\mathcal{E}'$, for all $k \le K$, $\boldsymbol{\mu}_k = \widehat{\boldsymbol{\theta}}_{k+1,h}$ for any $k \le K$. Indeed, on this event, for any $i \le K$,

$$y_i = \bar{\sigma}_{i,h}^{-1}\big(\langle \boldsymbol{\theta}_h^*, \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i)\rangle + \mathbb{1}\{\boldsymbol{\theta}_h^* \in \check{\mathcal{C}}_{i,h} \cap \widetilde{\mathcal{C}}_{i,h}\}\big[V_{i,h+1}(s_{h+1}^i) - \langle \boldsymbol{\phi}_{V_{i,h+1}}(s_h^i, a_h^i), \boldsymbol{\theta}^*\rangle\big]\big)$$

$= \bar{\sigma}_{i,h}^{-1} V_{i,h+1}(s_{h+1}^i),$

which does imply the claim. Therefore, by the definition of $\widehat{\mathcal{C}}_{k,h}$ and since on $\mathcal{E}'$ (2.8.6) holds, we get that on $\mathcal{E}'$, the relation $\boldsymbol{\theta}_h^* \in \widehat{\mathcal{C}}_{k,h}$ also holds. Finally, taking union bound over h and substituting (2.8.4) and (2.8.5) into Lemma 2.8.1 shows that with probability at least $1 - 3\delta$,

$$\boldsymbol{\theta}_h^* \in \cap_{k,h} \widehat{\mathcal{C}}_{k,h} \cap \widetilde{\mathcal{C}}_{k,h} \tag{2.8.7}$$

To finish our proof, it is thus sufficient to show that on the event when (2.8.7) holds, it also holds that

$$\left|[\bar{\mathbb{V}}_{k,h} V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k)\right| \leq E_{k,h}.$$

However, this is immediate from Lemma 2.8.1 and the definition of $E_{k,h}$. □

2.8.2 Proof of Theorem 2.5.2

In this subsection we prove Theorem 2.5.2. The proof is broken down into a number of lemmas. However, first we need the Azuma-Hoeffding inequality:

Lemma 2.8.2 (Azuma-Hoeffding inequality, Azuma 1967)**.** Let $M > 0$ be a constant. Let $\{x_i\}_{i=1}^n$ be a martingale difference sequence with respect to a filtration $\{\mathcal{G}_i\}_i$ ($\mathbb{E}[x_i|\mathcal{G}_i] = 0$ a.s. and x_i is $\mathcal{G}_{i+1}$-measurable) such that for all $i \in [n]$, $|x_i| \leq M$ holds almost surely. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, we have

$$\sum_{i=1}^n x_i \leq M\sqrt{2n \log(1/\delta)}.$$

For the remainder of this subsection, let $\mathcal{E}$ denote the event when the conclusion of Lemma 2.5.1 holds. Then Lemma 2.5.1 suggests $\mathbb{P}(\mathcal{E}) \geq 1 - 3\delta$. We introduce another two events $\mathcal{E}_1$ and $\mathcal{E}_2$:

$$\mathcal{E}_1 = \left\{ \forall h' \in [H], \sum_{k=1}^K \sum_{h=h'}^H \left[[\mathbb{P}_h(V_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [V_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k) \right] \right.$$

$$\leq 4H\sqrt{2T\log(H/\delta)}\Big\},$$

$$\mathcal{E}_2 = \Big\{\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}_h V_{h+1}^{\pi^k}](s_h^k, a_h^k) \leq 3(HT + H^3\log(1/\delta))\Big\}.$$

Then we have $\mathbb{P}(\mathcal{E}_1) \geq 1-\delta$ and $\mathbb{P}(\mathcal{E}_2) \geq 1-\delta$. The first one holds since $[\mathbb{P}_h(V_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [V_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k)$ forms a martingale difference sequence and $|[\mathbb{P}_h(V_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [V_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k)| \leq 4H$. Applying the Azuma-Hoeffding inequality (Lemma 2.8.2), we find that with probability at least $1-\delta$, simultaneously for all $h' \in [H]$, we have

$$\sum_{k=1}^{K}\sum_{h=h'}^{H}\Big[[\mathbb{P}_h(V_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [V_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k)\Big] \leq 4H\sqrt{2T\log(H/\delta)}, \quad (2.8.8)$$

which implies $\mathbb{P}(\mathcal{E}_1) \geq 1-\delta$. That $\mathbb{P}(\mathcal{E}_2) \geq 1-\delta$ holds is due to the following lemma:

Lemma 2.8.3 (Total variance lemma, Lemma C.5, Jin et al. 2018)**.** With probability at least $1-\delta$, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}[\mathbb{V}_h V_{h+1}^{\pi^k}](s_h^k, a_h^k) \leq 3(HT + H^3\log(1/\delta)).$$

We now prove the following three lemmas based on $\mathcal{E}, \mathcal{E}_1, \mathcal{E}_2$.

Lemma 2.8.4. Let $Q_{k,h}, V_{k,h}$ be defined in Algorithm 2. Then, on the event $\mathcal{E}$, for any s, a, k, h we have that $Q_h^*(s,a) \leq Q_{k,h}(s,a)$, $V_h^*(s) \leq V_{k,h}(s)$.

Proof. Since $\mathcal{E}$ holds, we have for any $k \in [K]$ and $h \in [H]$, $\boldsymbol{\theta}_h^* \in \widehat{\mathcal{C}}_{k,h}$. We prove the statement by induction. The statement holds for $h = H+1$ since $Q_{k,H+1}(\cdot,\cdot) = 0 = Q_{H+1}^*(\cdot,\cdot)$. Assume the statement holds for $h+1$. That is, $Q_{k,h+1}(\cdot,\cdot) \geq Q_{h+1}^*(\cdot,\cdot)$, $V_{k,h+1}(\cdot) \geq V_{h+1}^*(\cdot)$. Given s, a, if $Q_{k,h}(s,a) \geq H$, then $Q_{k,h}(s,a) \geq H \geq Q_h^*(s,a)$. Otherwise, we have

$$\begin{aligned}
&Q_{k,h}(s,a) - Q_h^*(s,a) \\
&= \langle \boldsymbol{\phi}_{V_{k,h+1}}(s,a), \widehat{\boldsymbol{\theta}}_{k,h}\rangle + \widehat{\beta}_k\Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s,a)\Big\|_2 - \langle \boldsymbol{\phi}_{V_{k,h+1}}(s,a), \boldsymbol{\theta}_h^*\rangle
\end{aligned}$$

$$
\begin{aligned}
&\quad + \mathbb{P}_h V_{k,h+1}(s,a) - \mathbb{P}_h V^*_{h+1}(s,a) \\
&\geq \widehat{\beta}_k \left\| \widehat{\boldsymbol{\Sigma}}^{-1/2}_{k,h} \boldsymbol{\phi}_{V_{k,h+1}}(s,a) \right\|_2 - \left\| \widehat{\boldsymbol{\Sigma}}^{1/2}_{k,h} (\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}^*_h) \right\|_2 \left\| \widehat{\boldsymbol{\Sigma}}^{-1/2}_{k,h} \boldsymbol{\phi}_{V_{k,h+1}}(s,a) \right\|_2 \\
&\quad + \mathbb{P}_h V_{k,h+1}(s,a) - \mathbb{P}_h V^*_{h+1}(s,a) \\
&\geq \mathbb{P}_h V_{k,h+1}(s,a) - \mathbb{P}_h V^*_{h+1}(s,a) \\
&\geq 0,
\end{aligned}
$$

where the first inequality holds due to Cauchy-Schwarz, the second inequality holds by the assumption that $\boldsymbol{\theta}^*_h \in \widehat{\mathcal{C}}_{k,h}$, the third inequality holds by the induction assumption and because $\mathbb{P}_h$ is a monotone operator with respect to the partial ordering of functions. Therefore, for all s, a, we have $Q_{k,h}(s,a) \geq Q^*_h(s,a)$, which implies $V_{k,h}(s) \geq V^*_h(s)$, finishing the inductive step and thus the proof. □

Lemma 2.8.5. Let $V_{k,h}, \bar{\sigma}_{k,h}$ be defined in Algorithm 2. Then, on the event $\mathcal{E} \cap \mathcal{E}_1$, we have

$$
\sum_{k=1}^{K} \left[V_{k,1}(s^k_1) - V^{\pi^k}_1(s^k_1) \right] \leq 2\widehat{\beta}_K \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \bar{\sigma}^2_{k,h}} \sqrt{2Hd \log(1 + K/\lambda)} + 4H\sqrt{2T \log(H/\delta)},
$$

$$
\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{P}_h [V_{k,h+1} - V^{\pi^k}_{h+1}](s^k_h, a^k_h) \leq 2\widehat{\beta}_K \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \bar{\sigma}^2_{k,h}} \sqrt{2dH^3 \log(1 + K/\lambda)} + 4H^2\sqrt{2T \log(H/\delta)}.
$$

Proof. Assume that $\mathcal{E} \cap \mathcal{E}_1$ holds. We have

$$
\begin{aligned}
V_{k,h}(s^k_h) - V^{\pi^k}_h(s^k_h) &\leq \langle \widehat{\boldsymbol{\theta}}_{k,h}, \boldsymbol{\phi}_{V_{k,h+1}}(s^k_h, a^k_h) \rangle - [\mathbb{P}_h V^{\pi^k}_{h+1}](s^k_h, a^k_h) + \widehat{\beta}_k \left\| \widehat{\boldsymbol{\Sigma}}^{-1/2}_{k,h} \boldsymbol{\phi}_{V_{k,h+1}}(s^k_h, a^k_h) \right\|_2 \\
&\leq \left\| \widehat{\boldsymbol{\Sigma}}^{1/2}_{k,h} (\widehat{\boldsymbol{\theta}}_{k,h} - \boldsymbol{\theta}^*_h) \right\|_2 \left\| \widehat{\boldsymbol{\Sigma}}^{-1/2}_{k,h} \boldsymbol{\phi}_{V_{k,h+1}}(s^k_h, a^k_h) \right\|_2 \\
&\qquad + [\mathbb{P}_h V_{k,h+1}](s^k_h, a^k_h) - [\mathbb{P}_h V^{\pi^k}_{h+1}](s^k_h, a^k_h) + \widehat{\beta}_k \left\| \widehat{\boldsymbol{\Sigma}}^{-1/2}_{k,h} \boldsymbol{\phi}_{V_{k,h+1}}(s^k_h, a^k_h) \right\|_2 \\
&\leq [\mathbb{P}_h V_{k,h+1}](s^k_h, a^k_h) - [\mathbb{P}_h V^{\pi^k}_{h+1}](s^k_h, a^k_h) + 2\widehat{\beta}_k \left\| \widehat{\boldsymbol{\Sigma}}^{1/2}_{k,h} \boldsymbol{\phi}_{V_{k,h+1}}(s^k_h, a^k_h) \right\|_2,
\end{aligned}
\tag{2.8.9}
$$

where the first inequality holds due to the definition of $V_{k,h}$ and the Bellman equation for $V^{\pi^k}_h$, the second inequality holds due to Cauchy-Schwarz inequality and because we are in a

linear MDP, the third inequality holds by the fact that on $\mathcal{E}$, $\boldsymbol{\theta}_h^* \in \widehat{\mathcal{C}}_{k,h}$. Meanwhile, since $V_{k,h}(s_h^k) - V_h^{\pi^k}(s_h^k) \le H$, we also have

$$
\begin{aligned}
&V_{k,h}(s_h^k) - V_h^{\pi^k}(s_h^k) \\
&\le \min\Big\{H, 2\widehat{\beta}_k \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\Big\|_2 + [\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{P}_h V_{h+1}^{\pi^k}](s_h^k, a_h^k)\Big\} \\
&\le \min\Big\{H, 2\widehat{\beta}_k \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\Big\|_2\Big\} + [\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{P}_h V_{h+1}^{\pi^k}](s_h^k, a_h^k) \\
&\le 2\widehat{\beta}_k \bar{\sigma}_{k,h} \min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\Big\|_2\Big\} + [\mathbb{P}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{P}_h V_{h+1}^{\pi^k}](s_h^k, a_h^k),
\end{aligned} \tag{2.8.10}
$$

where the second inequality holds since the optimal value function dominates the value function of any policy, and thus on $\mathcal{E}$, by Lemma 2.8.4, $V_{k,h+1}(\cdot) \ge V_{h+1}^{\pi^k}(\cdot)$, the third inequality holds since $2\widehat{\beta}_k \bar{\sigma}_{k,h} \ge \sqrt{d} \cdot H/\sqrt{d} \ge H$. By (2.8.10) we have

$$
\begin{aligned}
&V_{k,h}(s_h^k) - V_h^{\pi^k}(s_h^k) - [V_{k,h+1}(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)] \qquad (2.8.11)\\
&\quad \le 2\widehat{\beta}_k \bar{\sigma}_{k,h} \min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\Big\|_2\Big\} \\
&\qquad + \mathbb{P}_h[V_{k,h+1} - V_{h+1}^{\pi^k}](s_h^k, a_h^k) - [V_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k). \qquad (2.8.12)
\end{aligned}
$$

Summing up these inequalities for $k \in [K]$ and $h = h', \ldots, H$,

$$
\begin{aligned}
&\sum_{k=1}^{K}\Big[V_{k,h'}(s_{k,h'}) - V_{h'}^{\pi^k}(s_{k,h'})\Big] \\
&\le 2\sum_{k=1}^{K}\sum_{h=h'}^{H} \widehat{\beta}_k \bar{\sigma}_{k,h} \min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\Big\|_2\Big\} \\
&\quad + \sum_{k=1}^{K}\sum_{h=h'}^{H}\Big[[\mathbb{P}_h(V_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [V_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k)\Big] \\
&\le \underbrace{2\sum_{k=1}^{K}\sum_{h=1}^{H} \widehat{\beta}_k \bar{\sigma}_{k,h} \min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\Big\|_2\Big\}}_{I_1} + 4H\sqrt{2T\log(H/\delta)}, \qquad (2.8.13)
\end{aligned}
$$

where the first inequality holds by a telescoping argument and since $V_{k,H+1}(\cdot) = V_{h+1}^{\pi^k}(\cdot) = 0$,

the second inequality holds due to $\mathcal{E}_1$. To further bound I_1, we have

$$\begin{aligned} I_1 &\le \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\widehat{\beta}_k^2 \min\left\{1, \left\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\right\|_2^2\right\}} \\ &\le \widehat{\beta}_K\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{1, \left\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\right\|_2^2\right\}} \\ &\le \widehat{\beta}_K\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{2Hd\log(1+K/\lambda)}, \end{aligned} \tag{2.8.14}$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds since $\widehat{\beta}_k \le \widehat{\beta}_K$, the third inequality holds due to Lemma 2.7.2 with the fact that $\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\|_2 \le \|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)\|_2 \cdot \sqrt{d}/H \le \sqrt{d}$. Substituting (2.8.14) into (2.8.13) gives

$$\sum_{k=1}^{K}\left[V_{k,h'}(s_{k,h'}) - V_{h'}^{\pi^k}(s_{k,h'})\right] \le 2\widehat{\beta}_K\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{2Hd\log(1+K/\lambda)} + 4H\sqrt{2T\log(H/\delta)}\,. \tag{2.8.15}$$

Choosing $h' = 1$ here we get the first inequality that was to be proven. To get the second inequality, note that

$$\begin{aligned} &\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{P}_h[V_{k,h+1} - V_{h+1}^{\pi^k}](s_h^k, a_h^k) \\ &= \sum_{k=1}^{K}\sum_{h=2}^{H}[V_{k,h} - V_h^{\pi^k}](s_h^k) + \sum_{k=1}^{K}\sum_{h=1}^{H}\left[[\mathbb{P}_h(V_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - [V_{k,h+1} - V_{h+1}^{\pi^k}](s_{h+1}^k)\right] \\ &\le 2\widehat{\beta}_K\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{2dH^3\log(1+KH/(d\lambda))} + 4H^2\sqrt{2T\log(H/\delta)}, \end{aligned}$$

where to get the last inequality we sum up (2.8.15) for $h' = 2, \ldots, H$, and use the inequality that defines $\mathcal{E}_1$, which is followed by loosening the resulting bound. □

The next lemma is concerned with bounding $\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2$ on $\mathcal{E} \cap \mathcal{E}_2$:

Lemma 2.8.6. Let $V_{k,h}, \bar{\sigma}_{k,h}$ be defined in Algorithm 2. Then, on the event $\mathcal{E} \cap \mathcal{E}_2$, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \bar{\sigma}_{k,h}^2 \le H^2T/d + 3(HT + H^3\log(1/\delta)) + 2H\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{P}_h[V_{k,h+1} - V_{h+1}^{\pi^k}](s_h^k, a_h^k) + 2\widetilde{\beta}_K\sqrt{T}\sqrt{2dH\log(1 + KH^4/(d\lambda))} + 7\check{\beta}_K H^2\sqrt{T}\sqrt{2dH\log(1 + K/\lambda)}.$$

Proof. Assume that $\mathcal{E} \cap \mathcal{E}_2$ holds. Since we are on $\mathcal{E}$, by Lemma 2.8.4, for all k, h, $V_{k,h}(\cdot) \ge V_h^*(\cdot) \ge V_h^{\pi^k}(\cdot)$. Now, we calculate

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \bar{\sigma}_{k,h}^2 \\
&\quad\le \sum_{k=1}^{K}\sum_{h=1}^{H} \Big[H^2/d + [\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) + E_{k,h}\Big] \\
&\quad= H^2T/d + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \Big[[\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{h+1}^{\pi^k}](s_h^k, a_h^k)\Big]}_{I_1} + \underbrace{2\sum_{k=1}^{K}\sum_{h=1}^{H} E_{k,h}}_{I_2} \\
&\qquad+ \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} [\mathbb{V}_h V_{h+1}^{\pi^k}](s_h^k, a_h^k)}_{I_3} + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \Big[[\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k, a_h^k) - [\mathbb{V}_h V_{k,h+1}](s_h^k, a_h^k) - E_{k,h}\Big]}_{I_4},
\end{aligned} \tag{2.8.16}$$

where the first inequality holds due to the definition of $\bar{\sigma}_{k,h}$. To bound I_1, we have

$$\begin{aligned}
I_1 &\le \sum_{k=1}^{K}\sum_{h=1}^{H} [\mathbb{P}_h V_{k,h+1}^2](s_h^k, a_h^k) - [\mathbb{P}_h (V_{h+1}^{\pi^k})^2](s_h^k, a_h^k) \\
&\le 2H\sum_{k=1}^{K}\sum_{h=1}^{H} [\mathbb{P}_h(V_{k,h+1} - V_{h+1}^{\pi^k})](s_h^k, a_h^k),
\end{aligned}$$

where the first inequality holds since $V_{h+1}^{\pi^k}(\cdot) \le V_{h+1}^*(\cdot) \le V_{k,h+1}(\cdot)$, the second inequality holds since $V_{h+1}^{\pi^k}(\cdot), V_{k,h+1}(\cdot) \le H$. To bound I_2, we have

$$\begin{aligned}
I_2 \le\ & 2\sum_{k=1}^{K}\sum_{h=1}^{H} \widetilde{\beta}_k \min\Big\{1, \Big\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k, a_h^k)\Big\|_2\Big\} \\
&+ 4H\sum_{k=1}^{K}\sum_{h=1}^{H} \check{\beta}_k\bar{\sigma}_{k,h} \min\Big\{1, \Big\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k, a_h^k)/\bar{\sigma}_{k,h}\Big\|_2\Big\}
\end{aligned}$$

$$\leq 2\widetilde{\beta}_K\sqrt{T}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{1,\left\|\widetilde{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k)\right\|_2^2\right\}}$$
$$+ 7\check{\beta}_K H^2\sqrt{T}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{1,\left\|\widehat{\boldsymbol{\Sigma}}_{k,h}^{-1/2}\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)/\bar{\sigma}_{k,h}\right\|_2^2\right\}}$$
$$\leq 2\widetilde{\beta}_K\sqrt{T}\sqrt{2dH\log(1+KH^4/(d\lambda))} + 7\check{\beta}_K H^2\sqrt{T}\sqrt{2dH\log(1+K/\lambda)},$$

where the first inequality holds since $\widetilde{\beta}_k \geq H^2$ and $\check{\beta}_k\bar{\sigma}_{k,h} \geq \sqrt{d}\cdot H/\sqrt{d} = H$, the second inequality holds due to Cauchy-Schwarz inequality, $\widetilde{\beta}_k \leq \widetilde{\beta}_K$, $\check{\beta}_k \leq \check{\beta}_K$, and the following bound on $\bar{\sigma}_{k,h}$ due to the definitions of $\bar{\sigma}_{k,h}, [\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k)$ and $E_{k,h}$:

$$\bar{\sigma}_{k,h}^2 = \max\left\{H^2/d, [\bar{\mathbb{V}}_{k,h}V_{k,h+1}](s_h^k,a_h^k) + E_{k,h}\right\} \leq \max\left\{H^2/d, H^2+2H^2\right\} = 3H^2\,.$$

Finally, the third inequality holds due to Lemma 2.7.2 together with the facts that $\left\|\boldsymbol{\phi}_{V_{k,h+1}^2}(s_h^k,a_h^k)\right\|_2 \leq H^2$ and $\left\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)/\bar{\sigma}_{k,h}\right\|_2 \leq \left\|\boldsymbol{\phi}_{V_{k,h+1}}(s_h^k,a_h^k)\right\|_2\cdot\sqrt{d}/H \leq \sqrt{d}$. To bound I_3, since $\mathcal{E}_2$ holds, we have

$$I_3 \leq 3(HT + H^3\log(1/\delta)).$$

Finally, due to Lemma 2.5.1, we have $I_4 \leq 0$. Substituting I_1, I_2, I_3, I_4 into (2.8.16) ends our proof. □

With all above lemmas, we are ready to prove Theorem 2.5.2.

Proof of Theorem 2.5.2. By construction, taking a union bound, we have with probability $1-5\delta$ that $\mathcal{E}\cap\mathcal{E}_1\cap\mathcal{E}_2$ holds. In the remainder of the proof, assume that we are on this event. Thus, we can also use the conclusions of Lemmas 2.8.4, 2.8.5 and 2.8.6. We bound the regret as

$$\begin{aligned}\mathrm{Regret}(M_{\boldsymbol{\theta}^*},K) &\leq \sum_{k=1}^{K}\left[V_{k,1}(s_1^k) - V_1^{\pi^k}(s_1^k)\right]\\ &\leq 2\widehat{\beta}_K\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{2Hd\log(1+KH/(d\lambda))} + 4H\sqrt{2T\log(H/\delta)}\end{aligned}$$

$$= \widetilde{O}\bigg(\sqrt{dH}\sqrt{d}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2} + H\sqrt{T}\bigg), \tag{2.8.17}$$

where the first inequality holds due to Lemma 2.8.4, the second inequality holds due to Lemma 2.8.5, the equality holds since when $\lambda = 1/B^2$,

$$\widehat{\beta}_K = 8\sqrt{d\log(1+K/\lambda)\log(4K^2H/\delta)} + 4\sqrt{d}\log(4K^2H/\delta) + \sqrt{\lambda}B = \widetilde{\Theta}(\sqrt{d}).$$

It remains to bound $\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2$. For this we have

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2 &\le H^2T/d + 3(HT + H^3\log(1/\delta)) + 2H\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{P}_h[V_{k,h+1} - V_{h+1}^{\pi^k}](s_h^k, a_h^k) \\
&\quad + 2\widetilde{\beta}_K\sqrt{T}\sqrt{2dH\log(1+KH^4/(d\lambda))} + 7\check{\beta}_K H^2\sqrt{T}\sqrt{2dH\log(1+K/\lambda)} \\
&\le H^2T/d + 3(HT + H^3\log(1/\delta)) + 2H \\
&\quad \cdot \bigg(2\widehat{\beta}_K\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{2dH^3\log(1+K/\lambda)} + 4H^2\sqrt{2T\log(H/\delta)}\bigg) \\
&\quad + 2\widetilde{\beta}_K\sqrt{T}\sqrt{2dH\log(1+KH^4/(d\lambda))} + 7\check{\beta}_K H^2\sqrt{T}\sqrt{2dH\log(1+K/\lambda)} \\
&= \widetilde{O}\bigg(\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2}\sqrt{d^2H^5} + H^2T/d + TH + \sqrt{T}d^{1.5}H^{2.5} + \sqrt{T}H^3\bigg).
\end{aligned} \tag{2.8.18}$$

where the first inequality holds due to Lemma 2.8.6, the second inequality holds due to Lemma 2.8.5, the last equality holds due to the fact that $\widehat{\beta}_K = \widetilde{O}(\sqrt{d})$, $\lambda = 1/B^2$,

$$\begin{aligned}
\check{\beta}_K &= 8d\sqrt{\log(1+K/\lambda)\log(4k^2H/\delta)} + 4\sqrt{d}\log(4k^2H/\delta) + \sqrt{\lambda}B = \widetilde{\Theta}(d), \\
\widetilde{\beta}_K &= 8\sqrt{dH^4\log(1+KH^4/(d\lambda))\log(4k^2H/\delta)} + 4H^2\log(4k^2H/\delta) + \sqrt{\lambda}B = \widetilde{\Theta}(\sqrt{d}H^2).
\end{aligned}$$

Therefore, by the fact that $x \le a\sqrt{x} + b$ implies $x \le c(a^2 + b)$ with some $c > 0$, (2.8.18) yields that

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\sigma}_{k,h}^2 &\le \widetilde{O}\big(d^2H^5 + H^2T/d + TH + \sqrt{T}d^{1.5}H^{2.5} + \sqrt{T}H^3\big) \\
&= \widetilde{O}\big(d^2H^5 + d^4H^3 + TH + H^2T/d\big),
\end{aligned} \tag{2.8.19}$$

where the equality holds since $\sqrt{T}d^{1.5}H^{2.5} \leq (TH^2/d + d^4H^3)/2$ and $\sqrt{T}H^3 \leq (H^2T/d + H^4d)/2$. Substituting (2.8.19) into (2.8.17), we have

$$\text{Regret}\big(M_{\boldsymbol{\Theta}^*}, K\big) = \widetilde{O}\Big(\sqrt{d^2H^2 + dH^3}\sqrt{T} + d^2H^3 + d^3H^2\Big),$$

finishing the proof. □

Remark 2.8.7. To derive our upper bound of regret, we actually only need a weaker assumption on reward functions r_h such that for any policy π, we have $0 \leq \sum_{h=1}^{H} r_h(s_h, a_h) \leq H$, where $a_h = \pi_h(s_h),\ s_{h+1} \sim \mathbb{P}(\cdot|s_h, a_h)$. Therefore, under the assumption $0 \leq \sum_{h=1}^{H} r_h(s_h, a_h) \leq 1$ studied in Dann and Brunskill (2015); Jiang and Agarwal (2018); Wang et al. (2020a); Zhang et al. (2021a), by simply rescaling all parameters in Algorithm 2 by a factor of $1/H$, UCRL-VTR$^+$ achieves the regret $\widetilde{O}(\sqrt{d^2 + dH}\sqrt{T} + d^2H^2 + d^3H)$. Zhang et al. (2021a) has shown that in the tabular, *homogeneous* case with this normalization the regret is $\widetilde{O}(\sqrt{|\mathcal{S}||\mathcal{A}|K} + |\mathcal{S}|^2|\mathcal{A}|)$, regardless of the value of H. It remains an interesting open question whether this can be also achieved in homogeneous linear mixture MDPs.

2.9 Proof of Lower Bound Results in Section 2.5

2.9.1 Overview of the Lower Bound Construction

To prove the lower bound, we construct a hard instance $M(\mathcal{S}, \mathcal{A}, H, \{r_h\}, \{\mathbb{P}_h\})$ based on the hard-to-learn MDPs introduced in Zhou et al. (2021b). The transitions for stage h of the MDP are shown in Figure 2.1. The state space $\mathcal{S}$ consists of states $x_1, \ldots x_{H+2}$, where x_{H+1} and x_{H+2} are absorbing states. There are 2^{d-1} actions and $\mathcal{A} = \{-1, 1\}^{d-1}$. Regardless of the stage $h \in [H]$, no transition incurs a reward except transitions originating at x_{H+2}, which, as a result, can be regarded as the goal state. Under $\mathbb{P}_h$, the transition structure is as follows: As noted before, x_{H+1} and x_{H+2} are absorbing regardless of the action taken. If the state is x_i with $i \leq H$, under action $\boldsymbol{a} \in \{-1, 1\}^{d-1}$, the next state is either x_{H+2} or x_{i+1}, with respective probabilities $\delta + \langle \boldsymbol{\mu}_h, \boldsymbol{a} \rangle$ and $1 - (\delta + \langle \boldsymbol{\mu}_h, \boldsymbol{a} \rangle)$, where $\delta = 1/H$ and

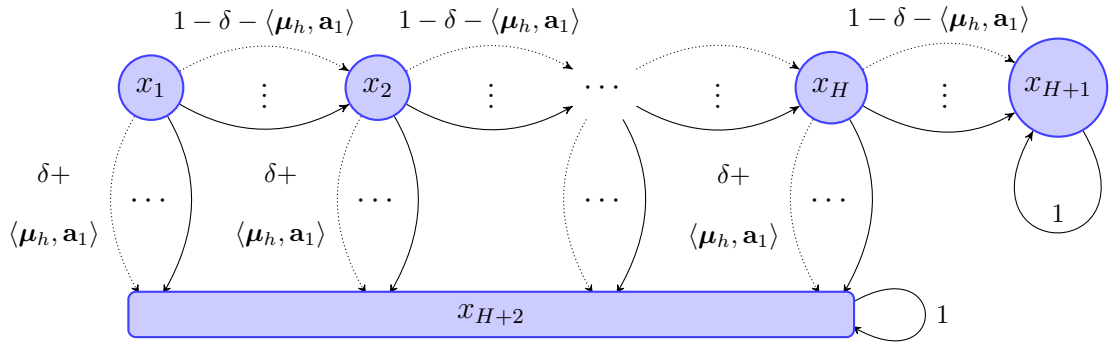


Figure 2.1: The transition kernel $\mathbb{P}_h$ of the class of hard-to-learn linear mixture MDPs. The kernel $\mathbb{P}_h$ is parameterized by $\boldsymbol{\mu}_h \in \{-\Delta, \Delta\}^{d-1}$ for some small Δ, $\delta = 1/H$ and the actions are from $\mathbf{a} \in \{+1, -1\}^{d-1}$. The learner knows this structure, but does not know $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_H)$.

$\boldsymbol{\mu}_h \in \{-\Delta, \Delta\}^{d-1}$ with $\Delta = \sqrt{\delta/K}/(4\sqrt{2})$ so that the probabilities are well-defined.

This is an inhomogeneous, linear mixture MDP. In particular, $\mathbb{P}_h(s'|s, \mathbf{a}) = \langle \boldsymbol{\phi}(s'|s, \mathbf{a}), \boldsymbol{\theta}_h \rangle$, with

$$\boldsymbol{\phi}(s'|s, \mathbf{a}) = \begin{cases} (\alpha(1-\delta), -\beta\mathbf{a}^\top)^\top, & s = x_h, s' = x_{h+1}, h \in [H]\,; \\ (\alpha\delta, \beta\mathbf{a}^\top)^\top, & s = x_h, s' = x_{H+2}, h \in [H]\,; \\ (\alpha, \mathbf{0}^\top)^\top, & s \in \{x_{H+1}, x_{H+2}\}, s' = s\,; \\ \mathbf{0}, & \text{otherwise}\,. \end{cases},$$

$$\boldsymbol{\theta}_h = (1/\alpha, \boldsymbol{\mu}_h^\top/\beta)^\top,\ h \in [H],$$

where $\alpha = \sqrt{1/(1+\Delta(d-1))}$, $\beta = \sqrt{\Delta/(1+\Delta(d-1))}$. It can be verified that $\boldsymbol{\phi}(\cdot|\cdot, \cdot)$ and $\{\boldsymbol{\theta}_h\}$ satisfy the requirements of a B-bounded linear mixture MDPs. In particular, (2.3.1) holds. Indeed, if we let $V : \mathcal{S} \to [0, 1]$ be any bounded function then for $s = x_{H+1}$ or $s = x_{H+2}$, $\boldsymbol{\phi}_V(s, \boldsymbol{a}) = \sum_{s'} \boldsymbol{\phi}(s'|s, \boldsymbol{a})V(s') = (\alpha V(s), \mathbf{0}^\top)^\top$ and hence $||\boldsymbol{\phi}_V(s, \boldsymbol{a})||_2 \le 1$, while for $s = x_h$ with $h \in [H]$, we have

$$||\boldsymbol{\phi}_V(s, \boldsymbol{a})||_2^2 = \alpha^2(V(x_{H+2})\delta + V(x_{h+1})(1-\delta))^2 + \beta^2(V(x_{H+2}) - V(x_{h+1}))^2||\boldsymbol{a}||_2^2$$

$$\leq \alpha^2 + (d-1)\beta^2$$
$$= 1. \tag{2.9.1}$$

Meanwhile, since $K \geq (d-1)/(32H(B-1))$, we have

$$\|\boldsymbol{\theta}_h\|_2^2 = \frac{1}{\alpha^2} + \frac{\|\boldsymbol{\mu}_h\|_2^2}{\beta^2} = (1 + \Delta(d-1))^2 = (1 + \sqrt{\delta/K}/4\sqrt{2} \cdot (d-1))^2 \leq B^2.$$

The initial state in each episode k is $s_{k,1} = x_1$. Note that if the agent transitions to x_{H+2} it remains there until the end of the episode. Due to the special structure of the MDP, at any stage $h \in [H]$, either the state is x_{H+2} or it is x_h. Further, state x_h can only be reached one way, through states $x_1, x_2, \ldots, x_{h-1}$. As such, knowing the current state is equivalent to knowing the history from the beginning of the episode and hence policies that simply decide at the beginning of the episode what actions to take upon reaching a state are as powerful as those that can use the "within episode" history.

Now, clearly, since the only rewarding transitions are those from x_{H+2}, the optimal strategy in stage h when in state x_h is to take action $\operatorname{argmax}_{\boldsymbol{a} \in \mathcal{A}} \langle \boldsymbol{\mu}_h, \boldsymbol{a} \rangle$. Intuitively, the learning problem is not *harder* than minimizing the regret on H linear bandit problems with a shared action set $\mathcal{A} = \{-1, +1\}^{d-1}$ and where the payoff on bandit $h \leq H/2$ of taking action $\boldsymbol{a} \in \mathcal{A}$ is $\Omega(H)Z$, where Z is drawn from a Bernoulli with parameter $\delta + \langle \boldsymbol{\mu}_h, \boldsymbol{a} \rangle$. Some calculation shows that the reverse is also true: Thanks to the choice of δ, $(1-\delta)^{H/2} \approx \text{const}$, hence there is sufficiently high probability of reaching all stages including stage $H/2$, even under the optimal policy. Hence, the MDP learning problem is not easier than solving the first $\Omega(H/2)$ bandit problems. Choosing $\Delta = \Theta(\sqrt{\delta/K})$, for K large enough, $(d-1)\Delta \leq \delta$ so the probabilities are well defined. Furthermore, on each of the bandit, the regret is at least $\Omega(dH\sqrt{K\delta})$. Since there are $\Omega(H/2)$ bandit problems, plugging in the choice of δ, we find that the total regret is $\Omega(dH\sqrt{KH})$ and the result follows by noting that $T = KH$.

Remark 2.9.1. Our lower bound analysis can be adapted to prove a lower bound for linear MDPs proposed in (Yang and Wang, 2019; Jin et al., 2020). In specific, based on our

constructed linear mixture MDP M in the proof sketch of Theorem 2.5.4, we can construct a linear MDP $\bar{M}(\mathcal{S}, \mathcal{A}, H, \{\bar{r}_h\}, \{\bar{\mathbb{P}}_h\})$ as follows. For each stage $h \in [H]$, the transition probability kernel $\bar{\mathbb{P}}_h$ and the reward function $\bar{r}_h$ are defined as $\bar{\mathbb{P}}_h(s'|s, \mathbf{a}) = \langle \boldsymbol{\phi}(s, \mathbf{a}), \boldsymbol{\mu}_h(s') \rangle$ and $\bar{r}_h(s, \mathbf{a}) = \langle \boldsymbol{\phi}(s, \mathbf{a}), \boldsymbol{\xi}_h \rangle$, where $\boldsymbol{\phi}(s, a), \boldsymbol{\mu}(s') \in \mathbb{R}^{d+1}$ are two feature mappings, and $\boldsymbol{\xi}_h \in \mathbb{R}^{d+1}$ is a parameter vector. Here, we choose $\boldsymbol{\phi}(s, \mathbf{a}), \boldsymbol{\mu}_h(s'), \boldsymbol{\xi}_h \in \mathbb{R}^{d+1}$ as follows:

$$\boldsymbol{\phi}(s, \mathbf{a}) = \begin{cases} (\alpha, \beta \mathbf{a}^\top, 0)^\top, & s = x_h,\ h \in [H+1]; \\ (0, \mathbf{0}^\top, 1)^\top, & s = x_{H+2}. \end{cases},$$

$$\boldsymbol{\mu}_h(s') = \begin{cases} ((1-\delta)/\alpha, -\boldsymbol{\mu}_h^\top/\beta, 0)^\top, & s' = x_{h+1}; \\ (\delta/\alpha, \boldsymbol{\mu}_h^\top/\beta, 1)^\top, & s' = x_{H+2}; \\ \mathbf{0}, & \text{otherwise}, \end{cases}$$

and $\boldsymbol{\xi}_h = (\mathbf{0}^\top, 1)^\top$. It can be verified that $\max\{\|\boldsymbol{\xi}_h\|_2, \|\boldsymbol{\mu}_h(\mathcal{S})\|_2\} \leq \sqrt{d+1}$, and $\|\boldsymbol{\phi}(s, \mathbf{a})\|_2 \leq 1$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. In addition, for any $h \in [H]$, we have $\mathbb{P}_h(s'|s, a) = \bar{\mathbb{P}}_h(s'|s, a)$ and $r_h(s, a) = \bar{r}_h(s, a)$ when $s = x_h$ or x_{H+2}. Since at stage h, s can be either x_h or x_{H+2}, we can show that the constructed linear MDP $\bar{M}$ has the same transition probability as the the linear mixture MDP M, which suggests the same lower bound $\Omega(dH\sqrt{T})$ in Theorem 2.5.4 also holds for linear MDP.

2.9.2 Proof of Theorem 2.5.4

We select $\delta = 1/H$ as suggested in Section 2.9.1. For brevity, with a slight abuse of notation, we will use $M_{\boldsymbol{\mu}}$ to denote the MDP described in Section 2.9.1 corresponding to the parameters $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_H)$. We will use $\mathbb{E}_{\boldsymbol{\mu}}$ denote the expectation underlying the distribution generated from the interconnection of a policy and MDP $M_{\boldsymbol{\mu}}$; since the policy is not denoted, we tacitly assume that the identity of the policy will always be clear from the context. We will similarly use $\mathbb{P}_{\boldsymbol{\mu}}$ to denote the corresponding probability measure.

We start with a lemma that will be the basis of our argument that shows that the regret

in our MDP can be lower bounded by the regret of $H/2$ bandit instances:

Lemma 2.9.2. Suppose $H \geq 3$ and $3(d-1)\Delta \leq \delta$. Fix $\boldsymbol{\mu} \in (\{-\Delta, \Delta\}^{d-1})^H$. Fix a possibly history dependent policy π and define $\bar{\boldsymbol{a}}_h^\pi = \mathbb{E}_{\boldsymbol{\mu}}[\boldsymbol{a}_h \,|\, s_h = x_h, s_1 = x_1]$: the expected action taken by the policy when it visits state x_h in stage h provided that the initial state is x_1. Then, letting V^* (V^π) be the optimal value function (the value function of policy π, respectively), we have

$$V_1^*(x_1) - V_1^\pi(x_1) \geq \frac{H}{10} \sum_{h=1}^{H/2} \Big(\max_{\mathbf{a} \in \mathcal{A}} \langle \boldsymbol{\mu}_h, \mathbf{a} \rangle - \langle \boldsymbol{\mu}_h, \bar{\boldsymbol{a}}_h^\pi \rangle \Big).$$

Proof. Fix $\boldsymbol{\mu}$. Since $\boldsymbol{\mu}$ is fixed, we drop the subindex from $\mathbb{P}$ and $\mathbb{E}$. Since $\mathcal{A} = \{+1, -1\}^{d-1}$ and $\boldsymbol{\mu}_h \in \{-\Delta, \Delta\}^{d-1}$, we have $(d-1)\Delta = \max_{\mathbf{a} \in \mathcal{A}} \langle \boldsymbol{\mu}_h, \mathbf{a} \rangle$. Recall the definition of the value of policy π in state x_1:

$$V_1^\pi(x_1) = \mathbb{E}\left[\sum_{h=1}^{H} r_h(s_h, a_h) \middle| s_1 = x_1, a_h \sim \pi_h(\cdot | s_1, a_1, \ldots, s_{h-1}, a_{h-1}, s_h) \right]. \tag{2.9.2}$$

Note that by the definition of our MDPs, only x_{H+2} satisfies that $r_h(x_{H+2}, \mathbf{a}) = 1$, all other rewards are zero. Also, once entered, the process does not leave x_{H+2}. Therefore,

$$V_1^\pi(x_1) = \sum_{h=1}^{H-1} (H-h) \mathbb{P}(N_h | s_1 = x_1). \tag{2.9.3}$$

where N_h is the event of visiting state x_h in stage h and then entering x_{H+2}:

$$N_h = \{ s_{h+1} = x_{H+2}, s_h = x_h \}. \tag{2.9.4}$$

By the law of total probability, the Markov property and the definition of $M_{\boldsymbol{\mu}}$,

$$\begin{aligned}
&\mathbb{P}(s_{h+1} = x_{H+2} | s_h = x_h, s_1 = x_1) \\
&\quad = \sum_{\boldsymbol{a} \in \mathcal{A}} \mathbb{P}(s_{h+1} = x_{H+2} | s_h = x_h, a_h = \boldsymbol{a}) \mathbb{P}(a_h = \boldsymbol{a} | s_h = x_h, s_1 = x_1) \\
&\quad = \sum_{\boldsymbol{a} \in \mathcal{A}} (\delta + \langle \boldsymbol{\mu}_h, \boldsymbol{a} \rangle) \mathbb{P}(a_h = \boldsymbol{a} | s_h = x_h, s_1 = x_1) \\
&\quad = \delta + \langle \boldsymbol{\mu}_h, \bar{\boldsymbol{a}}_h^\pi \rangle,
\end{aligned}$$

where the last equality used that by definition, $\bar{\boldsymbol{a}}_h^\pi = \sum_{\boldsymbol{a}\in\mathcal{A}} \mathbb{P}(a_h = \boldsymbol{a}|s_h = x_h, s_1 = x_1)\boldsymbol{a}$. It also follows that $\mathbb{P}(s_{h+1} = x_{h+1}|s_h = x_h, s_1 = x_1) = 1 - (\delta + \langle \boldsymbol{\mu}_h, \bar{\boldsymbol{a}}_h^\pi \rangle)$. Hence,

$$\mathbb{P}(N_h) = (\delta + \langle \boldsymbol{\mu}_h, \bar{\boldsymbol{a}}_h^\pi \rangle) \prod_{j=1}^{h-1} (1 - \delta - \langle \boldsymbol{\mu}_j, \bar{\boldsymbol{a}}_j^\pi \rangle)\,. \tag{2.9.5}$$

Defining $a_h = \langle \boldsymbol{\mu}_h, \bar{\boldsymbol{a}}_h^\pi \rangle$, we get that

$$V_1^\pi(x_1) = \sum_{h=1}^{H} (H-h)(a_h + \delta) \prod_{j=1}^{h-1} (1 - a_j - \delta)\,.$$

Working backwards, it is not hard to see that the optimal policy must take at stage the action that maximizes $\langle \boldsymbol{\mu}_h, \boldsymbol{a} \rangle$. Since $\max_{a\in\mathcal{A}} \langle \boldsymbol{\mu}_h, \boldsymbol{a} \rangle = (d-1)\Delta$, we get

$$V_1^*(x_1) = \sum_{h=1}^{H} (H-h)(1 - (d-1)\Delta - \delta)^{h-1}((d-1)\Delta + \delta).$$

For $i \in [H]$, introduce

$$S_i = \sum_{h=i}^{H} (H-h) \prod_{j=i}^{h-1} (1 - a_j - \delta)(a_h + \delta),\ T_i = \sum_{h=i}^{H} (H-h)(1 - (d-1)\Delta - \delta)^{h-i}((d-1)\Delta + \delta).$$

Then $V_1^*(x_1) - V_1^\pi(x_1) = T_1 - S_1$. To lower bound $T_1 - S_1$, first note that

$$S_i = (H-i)(a_i + \delta) + S_{i+1}(1 - a_i - \delta),\ T_i = (H-i)((d-1)\Delta + \delta) + T_{i+1}(1 - (d-1)\Delta - \delta),$$

which gives that

$$T_i - S_i = (H - i - T_{i+1})((d-1)\Delta - a_i) + (1 - a_i - \delta)(T_{i+1} - S_{i+1}). \tag{2.9.6}$$

Therefore by induction, we get that

$$T_1 - S_1 = \sum_{h=1}^{H-1} ((d-1)\Delta - a_h)(H - h - T_{h+1}) \prod_{j=1}^{h-1} (1 - a_j - \delta). \tag{2.9.7}$$

To further bound (2.9.7), first we note that T_h can be written as the following closed-form expression:

$$T_h = \frac{(1 - (d-1)\Delta - \delta)^{H-h} - 1}{(d-1)\Delta + \delta} + H - h + 1 - (1 - (d-1)\Delta - \delta)^{H-h},$$

Hence, for any $h \le H/2$,

$$\begin{aligned} H - h - T_{h+1} &= \frac{1 - (1 - (d-1)\Delta - \delta)^{H-h}}{(d-1)\Delta + \delta} + (1 - (d-1)\Delta - \delta)^{H-h} \\ &\ge \frac{1 - (1 - (d-1)\Delta - \delta)^{H/2}}{(d-1)\Delta + \delta} \ge H/3, \end{aligned} \tag{2.9.8}$$

where the last inequality holds since $3(d-1)\Delta \le \delta = 1/H$ and $H \ge 3$. Furthermore we have

$$\prod_{j=1}^{h-1}(1 - a_j - \delta) \ge (1 - 4\delta/3)^H \ge 1/3, \tag{2.9.9}$$

where the first inequality holds since $a_j \le (d-1)\Delta, 3(d-1)\Delta \le \delta$, the second one holds since $\delta = 1/H$ and $H \ge 3$. Therefore, substituting (2.9.8) and (2.9.9) into (2.9.7), we have

$$V_1^*(x_1) - V_1^\pi(x_1) = T_1 - S_1 \ge \frac{H}{10} \cdot \sum_{h=1}^{H/2}((d-1)\Delta - a_h),$$

which finishes the proof. □

We also need a lower bound on the regret on linear bandits with the hypercube action set $\mathcal{A} = \{-1, 1\}^{d-1}$, Bernoulli bandits with linear mean payoff. While the proof technique used is standard (cf. Lattimore and Szepesvári 2020), we give the full proof as the "scaling" of the reward parameters is nonstandard:

Lemma 2.9.3. Fix a positive real $0 < \delta \le 1/3$, and positive integers K, d and assume that $K \ge d^2/(2\delta)$. Let $\Delta = \sqrt{\delta/K}/(4\sqrt{2})$ and consider the linear bandit problems $\mathcal{L}_{\boldsymbol{\mu}}$ parameterized with a parameter vector $\boldsymbol{\mu} \in \{-\Delta, \Delta\}^d$ and action set $\mathcal{A} = \{-1, 1\}^d$ so that the reward distribution for taking action $\boldsymbol{a} \in \mathcal{A}$ is a Bernoulli distribution $B(\delta + \langle \boldsymbol{\mu}^*, \mathbf{a} \rangle)$. Then for any bandit algorithm $\mathcal{B}$, there exists a $\boldsymbol{\mu}^* \in \{-\Delta, \Delta\}^d$ such that the expected pseudo-regret of $\mathcal{B}$ over first K steps on bandit $\mathcal{L}_{\boldsymbol{\mu}^*}$ is lower bounded as follows:

$$\mathbb{E}_{\boldsymbol{\mu}^*}\text{Regret}(K) \ge \frac{d\sqrt{K\delta}}{8\sqrt{2}}.$$

Note that the expectation is with respect to a distribution that depends both on $\mathcal{B}$ and $\boldsymbol{\mu}^*$, but since $\mathcal{B}$ is fixed, this dependence is hidden.

Proof. Let $\mathbf{a}_k \in \mathcal{A} = \{-1, 1\}^d$ denote the action chosen in round k. Then for any $\boldsymbol{\mu} \in \{-\Delta, \Delta\}^d$, the expected pseudo regret $\mathbb{E}_{\boldsymbol{\mu}}\text{Regret}(K)$ corresponding to $\boldsymbol{\mu}$ satisfies

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu}}\text{Regret}(K) &= \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}(\max_{\mathbf{a}\in\mathcal{A}} \langle \boldsymbol{\mu}, \mathbf{a}\rangle - \langle \boldsymbol{\mu}, \mathbf{a}_k\rangle) \\
&= \Delta \sum_{k=1}^{K} \sum_{j=1}^{d} \mathbb{E}_{\boldsymbol{\mu}}\, \mathbb{1}\{\text{sgn}([\boldsymbol{\mu}]_j) \neq \text{sgn}([\mathbf{a}_k]_j)\} \\
&= \Delta \sum_{j=1}^{d} \underbrace{\sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}\, \mathbb{1}\{\text{sgn}([\boldsymbol{\mu}]_j) \neq \text{sgn}([\mathbf{a}_k]_j)\}}_{N_j(\boldsymbol{\mu})}, \qquad (2.9.10)
\end{aligned}$$

where for a vector $\boldsymbol{x}$, we use $[\boldsymbol{x}]_j$ to denote its jth entry. Let $\boldsymbol{\mu}^j \in \{-\Delta, \Delta\}^d$ denote the vector which differs from $\boldsymbol{\mu}$ at its jth coordinate only. Then, we have

$$\begin{aligned}
2\sum_{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{\mu}}\text{Regret}(K) &= \Delta \sum_{\boldsymbol{\mu}} \sum_{j=1}^{d} (\mathbb{E}_{\boldsymbol{\mu}} N_j(\boldsymbol{\mu}) + \mathbb{E}_{\boldsymbol{\mu}^j} N_j(\boldsymbol{\mu}^j)) \\
&= \Delta \sum_{\boldsymbol{\mu}} \sum_{j=1}^{d} (K + \mathbb{E}_{\boldsymbol{\mu}} N_j(\boldsymbol{\mu}) - \mathbb{E}_{\boldsymbol{\mu}^j} N_j(\boldsymbol{\mu})) \\
&\geq \Delta \sum_{\boldsymbol{\mu}} \sum_{j=1}^{d} (K - \sqrt{1/2} K \sqrt{\text{KL}(\mathcal{P}_{\boldsymbol{\mu}}, \mathcal{P}_{\boldsymbol{\mu}^j})}), \qquad (2.9.11)
\end{aligned}$$

where the inequality holds due to $N_j(\boldsymbol{\mu}) \in [0, K]$ and Pinsker's inequality (Exercise 14.4 and Eq. 14.12, Lattimore and Szepesvári 2020), $\mathcal{P}_{\boldsymbol{\mu}}$ denotes the joint distribution over the all possible reward sequences $(r_1, \ldots, r_K) \in \{0, 1\}^K$ of length K, induced by the interconnection of the algorithm and the bandit parameterized by $\boldsymbol{\mu}$. By the chain rule of relative entropy, $\text{KL}(\mathcal{P}_{\boldsymbol{\mu}}, \mathcal{P}_{\boldsymbol{\mu}^j})$ can be further decomposed as (cf. Exercise 14.11 of Lattimore and Szepesvári 2020),

$$\begin{aligned}
\text{KL}(\mathcal{P}_{\boldsymbol{\mu}}, \mathcal{P}_{\boldsymbol{\mu}^j}) &= \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[\text{KL}(\mathcal{P}_{\boldsymbol{\mu}}(r_k|\mathbf{r}_{1:k-1}), \mathcal{P}_{\boldsymbol{\mu}^j}(r_k|\mathbf{r}_{1:k-1}))] \\
&= \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\mu}}[\text{KL}(B(\delta + \langle \mathbf{a}_k, \boldsymbol{\mu}\rangle), (B(\delta + \langle \mathbf{a}_k, \boldsymbol{\mu}^j\rangle))] \\
&\leq \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\mu}} \left[\frac{2\langle \boldsymbol{\mu} - \boldsymbol{\mu}^j, \mathbf{a}_k\rangle^2}{\langle \boldsymbol{\mu}, \mathbf{a}_k\rangle + \delta}\right]
\end{aligned}$$

$$\leq \frac{16K\Delta^2}{\delta}, \tag{2.9.12}$$

where the second equality holds since the round k reward's distribution is the Bernoulli distribution $B(\delta + \langle \mathbf{a}_k, \boldsymbol{\mu} \rangle)$ in the environment parameterized by $\boldsymbol{\mu}$, the first inequality holds since for any two Bernoulli distribution $B(a)$ and $B(b)$, we have $\mathrm{KL}(B(a), B(b)) \leq 2(a-b)^2/a$ when $a \leq 1/2, a+b \leq 1$, the second inequality holds since $\boldsymbol{\mu}$ only differs from $\boldsymbol{\mu}^j$ at j-th coordinate, $\langle \boldsymbol{\mu}, \mathbf{a}_k \rangle \geq -d\Delta \geq -\delta/2$. It can be verified that these requirements hold when $\delta \leq 1/3$, $d\Delta \leq \delta/2$. Therefore, substituting (2.9.12) into (2.9.11), we have

$$2\sum_{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{\mu}} \mathrm{Regret}(K) \geq \sum_{\boldsymbol{\mu}} \Delta d(K - \sqrt{2}K^{3/2}\Delta/\sqrt{\delta}) = \sum_{\boldsymbol{\mu}} \frac{d\sqrt{K\delta}}{4\sqrt{2}},$$

where the equality holds since $\Delta = \sqrt{\delta/K}/(4\sqrt{2})$. Selecting $\boldsymbol{\mu}^*$ which maximizes $\mathbb{E}_{\boldsymbol{\mu}} \mathrm{Regret}(K)$ finishes the proof. □

With this, we are ready to prove Theorem 2.5.4.

Proof of Theorem 2.5.4. We can verify that the selection of K, d, H, δ satisfy the requirement of Lemma 2.9.2 and Lemma 2.9.3. Let π^k denote the possibly nonstationary policy that is executed in episode k given the history up to the beginning of the episode. Then, by Lemma 2.9.2, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu}} \mathrm{Regret}\Big(M_{\boldsymbol{\mu}}, K\Big) &= \mathbb{E}_{\boldsymbol{\mu}} \Bigg[\sum_{k=1}^{K} [V_1^*(x_1) - V_1^{\pi^k}(x_1)] \Bigg] \\
&\geq \frac{H}{10} \sum_{h=1}^{H/2} \underbrace{\mathbb{E}_{\boldsymbol{\mu}} \Bigg[\sum_{k=1}^{K} \Big(\max_{\mathbf{a} \in \mathcal{A}} \langle \boldsymbol{\mu}_h, \mathbf{a} \rangle - \langle \boldsymbol{\mu}_h, \bar{\boldsymbol{a}}_h^{\pi_k} \rangle \Big) \Bigg]}_{I_h(\boldsymbol{\mu}, \pi)}.
\end{aligned} \tag{2.9.13}$$

Let $\boldsymbol{\mu}^{-h} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{h-1}, \boldsymbol{\mu}_{h+1}, \ldots, \boldsymbol{\mu}_H)$. Now, every MDP policy π gives rise to a bandit algorithm $\mathcal{B}_{\pi,h,\boldsymbol{\mu}^{-h}}$ for the linear bandit $\mathcal{L}_{\boldsymbol{\mu}_h}$ of Lemma 2.9.3. This bandit algorithm is such that the distribution of action it plays in round k matches the distribution of action played by π in stage h of episode k conditioned on the event that $s_h^k = x_h$, i.e., $\mathbb{P}_{\mu,\pi}(a_h^k = \cdot | s_h^k = x_h)$ with the tacit assumption that the first state in every episode is x_1.

As the notation suggests, the bandit algorithm depends on $\boldsymbol{\mu}^{-h}$. In particular, to play in round k, the bandit algorithm feeds π with data from the MDP kernels up until the beginning of episode k: For $i \neq h$, this can be done by just following $\mathbb{P}_i$ since the parameters of these kernels is known to $\mathcal{B}_{\pi,h,\boldsymbol{\mu}^{-h}}$. When $i = h$, since $\mathbb{P}_h$ is not available to the bandit algorithm, every time it is on stage h, if the state is x_h, it feeds the action obtained from π to $\mathcal{L}_\mu$ and if the reward is 1, it feeds π with the next state x_{H+2}, otherwise it feeds it with next state x_{h+1}. When $i = h$ and the state is not x_h, it can only be x_{H+2}, in which case the next state fed to π is x_{H+2} regardless of the action it takes. At the beginning of episode k, to ensure that state x_h is "reached", π is fed with the states x_1, x_2, ..., x_h. Then, π is queried for its action, which is the action that the bandit plays in round k. Clearly, by this construction, the distribution of action played in round k by $\mathcal{B}_{\pi,h,\boldsymbol{\mu}^{-h}}$ matches the target.

Denoting by $\mathrm{BanditRegret}(\mathcal{B}_{\pi,h,\boldsymbol{\mu}^{-h}}, \boldsymbol{\mu}_h)$ the regret of this bandit algorithm on $\mathcal{L}_{\boldsymbol{\mu}}$, by our construction, $I_h(\boldsymbol{\mu}, \pi) = \mathrm{BanditRegret}(\mathcal{B}_{\pi,h,\boldsymbol{\mu}^{-h}}, \boldsymbol{\mu}_h)$ for all $h \in [H/2]$. Hence,

$$\begin{aligned}
\sup_{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{\mu}} \mathrm{Regret}\Big(M_{\boldsymbol{\mu}}, K\Big) &\geq \sup_{\boldsymbol{\mu}} \frac{H}{10} \sum_{h=1}^{H/2} \mathrm{BanditRegret}(\mathcal{B}_{\pi,h,\boldsymbol{\mu}^{-h}}, \boldsymbol{\mu}_h) \\
&\geq \sup_{\boldsymbol{\mu}} \frac{H}{10} \sum_{h=1}^{H/2} \inf_{\widetilde{\boldsymbol{\mu}}^{-h}} \mathrm{BanditRegret}(\mathcal{B}_{\pi,h,\widetilde{\boldsymbol{\mu}}^{-h}}, \boldsymbol{\mu}_h) \\
&= \frac{H}{10} \sum_{h=1}^{H/2} \sup_{\boldsymbol{\mu}^h} \inf_{\widetilde{\boldsymbol{\mu}}^{-h}} \mathrm{BanditRegret}(\mathcal{B}_{\pi,h,\widetilde{\boldsymbol{\mu}}^{-h}}, \boldsymbol{\mu}_h) \\
&\geq \frac{H^2}{20} \frac{(d-1)\sqrt{K\delta}}{8\sqrt{2}}\,,
\end{aligned}$$

where the last inequality follows by Lemma 2.9.3. The result follows by plugging in $\delta = 1/H$ and $T = KH$. □

CHAPTER 3

Computational Efficient Reinforcement Learning through Uncertainties

3.1 Introduction

In this chapter, we aim to develop computational-efficient RL through the use of uncertainties. Specifically, we aim to develop online RL algorithms with linear function approximation under adaptivity constraints, where only finite number of policy updates is allowed. In detail, we consider time-inhomogeneous episodic linear MDPs (Jin et al., 2020) where both the transition probability and the reward function are unknown to the agent. In terms of the limited adaptivity imposed on the agent, we consider two scenarios that have been previously studied in the online learning literature (Perchet et al., 2016; Abbasi-Yadkori et al., 2011): the batch learning model and the rare policy switch model. More specifically, in the batch learning model (Perchet et al., 2016), the agent is forced to pre-determine the number of batches (or equivalently batch size). Within each batch, the same policy is used to select actions, and the policy is updated only at the end of this batch. The amount of adaptivity in the batch learning model is measured by the number of batches, which is expected to be as small as possible. In contrast, in the rare policy switch model (Abbasi-Yadkori et al., 2011), the agent can adaptively choose when to switch the policy and therefore start a new batch in the learning process as long as the total number of policy updates does not exceed the given budget on the number of policy switches. The amount of adaptivity in the rare policy switch model can be measured by the number of policy switches, which turns out to be the

same as the global switching cost introduced in Bai et al. (2019). It is worth noting that for the same amount of adaptivity[1], the rare policy switch model can be seen as a relaxation of the batch learning model since the agent in the batch learning model can only change the policy at pre-defined time steps. In our work, for each of these limited adaptivity models, we propose a variant of the LSVI-UCB algorithm (Jin et al., 2020), which can be viewed as an RL algorithm with full adaptivity in the sense that it switches the policy at a per-episode scale. Our algorithms can attain the same regret as LSVI-UCB, yet with a substantially smaller number of batches/policy switches. This enables parallel learning and improves the large-scale deployment of RL algorithms with linear function approximation.

The main contributions of this chapter are summarized as follows:

- For the *batch learning* model, we propose an LSVI-UCB-Batch algorithm for linear MDPs and show that it enjoys an $\widetilde{O}(\sqrt{d^3H^3T} + dHT/B)$ regret, where d is the dimension of the feature mapping, H is the episode length, T is the number of interactions and B is the number of batches. Our result suggests that it suffices to use only $\sqrt{T/dH}$ batches, rather than T batches, to obtain the same regret $\widetilde{O}(\sqrt{d^3H^3T})$ achieved by LSVI-UCB (Jin et al., 2020) in the fully sequential decision model. We also prove a lower bound of the regret for this model, which suggests that the required number of batches $\widetilde{O}(\sqrt{T})$ is sharp.

- For the *rare policy switch* model, we propose an LSVI-UCB-RareSwitch algorithm for linear MDPs and show that it enjoys an $\widetilde{O}(\sqrt{d^3H^3T[1+T/(dH)]^{dH/B}})$ regret, where B is the number of policy switches. The policy update rule of LSVI-UCB-RareSwitch depends on a criterion defined by *epistemic uncertainty* of the estimated transition dynamic. Our result implies that $dH \log T$ policy switches are sufficient to obtain the same regret $\widetilde{O}(\sqrt{d^3H^3T})$ achieved by LSVI-UCB. The number of policy switches is much smaller than that[2] of the batch learning model when T is large.

[1]The number of batches in the batch learning model is comparable to the number of policy switches in the rare policy switch model.

[2]The number of policy switches is identical to the number of batches in the batch learning model.

The rest of this chapter is organized as follows. In Section 3.2 we discuss previous works related to this chapter, with a focus on RL with linear function approximation and online learning with limited adaptivity. In Section 3.3 we introduce necessary preliminaries for MDPs and adaptivity constraints. Sections 3.4 and 3.5 present our proposed algorithms and the corresponding theoretical results for the batch learning model and the rare policy switch model respectively. In Section 3.6 we present the numerical experiment which supports our theory. Finally, we conclude this chapter and point out a future direction in Section 3.7.

3.2 Related Works

Reinforcement Learning with Linear Function Approximation Recently, there have been many advances in RL with function approximation, especially the linear case. Jin et al. (2020) proposed an efficient algorithm for the first time for linear MDPs of which the transition probability and the rewards are both linear functions with respect to a feature mapping $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. Under similar assumptions, different settings (e.g., discounted MDPs) have also been studied in Yang and Wang (2019); Du et al. (2019b); Zanette et al. (2020b); Neu and Pike-Burke (2020) and He et al. (2021a). A parallel line of work studies linear mixture MDPs (a.k.a. linear kernel MDPs) based on a ternary feature mapping $\psi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ (see Jia et al. (2020); Zhou et al. (2021b); Cai et al. (2020); Zhou et al. (2021a)). For other function approximation settings, we refer readers to generalized linear model (Wang et al., 2020c), general function approximation with Eluder dimension (Wang et al., 2020b; Ayoub et al., 2020), kernel approximation (Yang et al., 2020), function approximation with disagreement coefficients (Foster et al., 2021) and bilinear classes (Du et al., 2021).

Online Learning with Limited Adaptivity As we mentioned before, online learning with limited adaptivity has been studied in two popular models of adaptivity constraints: the batch learning model and the rare policy switch model.

For the *batch learning model*, Altschuler and Talwar (2018) proved that the optimal

regret bound for prediction-from-experts (PFE) is $\widetilde{O}(\sqrt{T\log n})$ when the number of batches $B = \Omega(\sqrt{T\log n})$, and $\min(\widetilde{O}(T\log n/B), T)$ when $B = O(\sqrt{T\log n})$, exhibiting a phase-transition phenomenon[3]. Here T is the number of rounds and n is the number of actions. For general online convex optimization, Chen et al. (2020) showed that the minimax regret bound is $\widetilde{O}(T/\sqrt{B})$. Perchet et al. (2016) studied batched 2-arm bandits, and Gao et al. (2019) studied the batched multi-armed bandits (MAB). Dekel et al. (2014) proved a $\Omega(T/\sqrt{B})$ lower bound for batched MAB, and Altschuler and Talwar (2018) further characterized the dependence on the number of actions n and showed that the corresponding minimax regret bound is $\min(\widetilde{O}(T\sqrt{n}/\sqrt{B}), T)$. For batched linear bandits with adversarial contexts, Han et al. (2020) showed that the minimax regret bound is $\widetilde{O}(\sqrt{dT} + dT/B)$ where d is the dimension of the context vectors. Better rates can be achieved for batched linear bandits with stochastic contexts as shown in Esfandiari et al. (2021); Han et al. (2020); Ruan et al. (2021).

For the *rare policy switch model*, the minimax optimal regret bound for PFE is $O(\sqrt{T\log n})$ in terms of both the expected regret (Kalai and Vempala, 2005; Geulen et al., 2010; Cesa-Bianchi et al., 2013; Devroye et al., 2015) and high-probability guarantees (Altschuler and Talwar, 2018), where T is the number of rounds, and n is the number of possible actions. For MAB, the minimax regret bound has been shown to be $\widetilde{O}(T^{2/3}n^{1/3})$ by Arora et al. (2012); Dekel et al. (2014). For stochastic linear bandits, Abbasi-Yadkori et al. (2011) proposed a rarely switching OFUL algorithm achieving $\widetilde{O}(d\sqrt{T})$ regret with $\log(T)$ batches. Ruan et al. (2021) proposed an algorithm achieving $\widetilde{O}(\sqrt{dT})$ regret with less than $O(d\log d\log T)$ batches for stochastic linear bandits with adversarial contexts.

For episodic RL with finite state and action space, Bai et al. (2019) proposed an algorithm achieving $\widetilde{O}(\sqrt{H^3SAT})$ regret with $O(H^3SA\log(T/(AH)))$ local switching cost where S and A are the number of states and actions respectively. They also provided a $\Omega(HSA)$ lower bound on the local switching cost that is necessary for sublinear regret. For the global

[3]They call it B-switching budget setting, which is identical to the batch learning model.

switching cost, Zhang et al. (2021a) proposed an MVP algorithm with at most $O(SA\log(KH))$ global switching cost for time-homogeneous tabular MDPs.

3.3 Preliminaries

We adapt the similar problem setting as that in Chapter 2. In the online learning setting, at the beginning of k-th episode, the agent chooses a policy π^k and the environment selects an initial state s_1^k, then the agent interacts with environment following policy π^k and receives states s_h^k and rewards $r_h(s_h^k, a_h^k)$ for $h \in [H]$. Here the reward function is deterministic and known to the agent. To measure the performance of the algorithm, we adopt the following notion of the total regret, which is the summation of suboptimalities between policy π^k and optimal policy π^*:

Definition 3.3.1. We denote $T = KH$, and the regret $\mathrm{Regret}(T)$ is defined as

$$\mathrm{Regret}(T) = \sum_{k=1}^{K} \left[V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k) \right].$$

3.3.1 Linear Function Approximation

In this work, we consider a special class of MDPs called *linear MDPs* (Yang and Wang, 2019; Jin et al., 2020), where both the transition probability function and reward function can be represented as a linear function of a given feature mapping $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. Formally speaking, we have the following definition for linear MDPs.

Definition 3.3.2. $M(\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h\in[H]}, \{\mathbb{P}_h\}_{h\in[H]})$ is called a linear MDP if there exist a *known* feature mapping $\boldsymbol{\phi}(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, *unknown* measures $\{\boldsymbol{\mu}_h = (\mu_h^{(1)}, \cdots, \mu_h^{(d)})\}_{h\in[H]}$ over $\mathcal{S}$ and unknown vectors $\{\boldsymbol{\theta}_h \in \mathbb{R}^d\}_{h\in[H]}$ with $\max_{h\in[H]}\{\|\boldsymbol{\mu}_h(\mathcal{S})\|_2, \|\boldsymbol{\theta}_h\|\} \le \sqrt{d}$, such that the following holds for all $h \in [H]$:

- For any state-action-state triplet $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, $\mathbb{P}_h(s'|s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h(s') \rangle$.

- For any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, $r_h(s, a) = \langle \boldsymbol{\phi}(s, a), \boldsymbol{\theta}_h \rangle$.

Without loss of generality, we also assume that $\|\boldsymbol{\phi}(s, a)\|_2 \leq 1$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

With Definition 3.3.2, it is shown in Jin et al. (2020) that the action-value function can be written as a linear function of the features.

Proposition 3.3.3 (Proposition 2.3, Jin et al. 2020)**.** For a linear MDP, for any policy π, there exist weight vectors $\{\mathbf{w}_h^\pi\}_{h \in [H]}$ such that for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we have $Q_h^\pi(s, a) = \langle \boldsymbol{\phi}(s, a), \mathbf{w}_h^\pi \rangle$. Moreover, we have $\|\mathbf{w}_h^\pi\|_2 \leq 2H\sqrt{d}$ for all $h \in [H]$.

Therefore, with the known feature mapping $\boldsymbol{\phi}(\cdot, \cdot)$, it suffices to estimate the weight vectors $\{\mathbf{w}_h^\pi\}_{h \in [H]}$ in order to recover the action-value functions. This is the core idea behind almost all the algorithms and theoretical analyses for linear MDPs.

3.3.2 Models for Limited Adaptivity

In this work, we consider RL algorithms with limited adaptivity. There are two typical models for online learning with such limited adaptivity: *batch learning model* (Perchet et al., 2016) and *rare policy switch model* (Abbasi-Yadkori et al., 2011).

For the batch learning model, the agent pre-determines the batch grids $1 = t_1 < t_2 < \cdots < t_B < t_{B+1} = K + 1$ at the beginning of the algorithm, where B is the number of batches. The b-th batch consists of t_b-th to $(t_{b+1} - 1)$-th episodes, and the agent follows the same policy within each batch. The adaptivity is measured by the number of batches.

For the rare policy switch model, the agent can decide whether she wants to switch the current policy or not. The adaptivity is measured by the number of policy switches, which is defined as

$$N_{\text{switch}} = \sum_{k=1}^{K-1} \mathbb{1}\{\pi^k \neq \pi^{k+1}\},$$

where $\pi^k \neq \pi^{k+1}$ means that there exists some $(h, s) \in [H] \times \mathcal{S}$ such that $\pi_h^k(s) \neq \pi_h^{k+1}(s)$.

Given a budget on the number of batches or the number of policy switches, we aim to design RL algorithms with linear function approximation that can achieve the same regret as their full adaptivity counterpart, e.g., LSVI-UCB (Jin et al., 2020).

3.4 RL in the Batch Learning Model

In this section, we consider RL with linear function approximation in the batch learning model, where given the number of batches B, we need to pin down the batches before the agent starts to interact with the environment.

Algorithm We propose LSVI-UCB-Batch algorithm as displayed in Algorithm 3, which can be regarded as a variant of the LSVI-UCB algorithm proposed in Jin et al. (2020) yet with limited adaptivity. Algorithm 3 takes a series of batch grids $\{t_1, \ldots, t_{B+1}\}$ as input, where the i-th batch starts at t_i and ends at $t_{i+1} - 1$. LSVI-UCB-Batch takes the uniform batch grids as its selection of grids, i.e., $t_i = (i-1) \cdot \lfloor K/B \rfloor + 1, i \in [B]$. By Proposition 3.3.3, we know that for each $h \in [H]$, the optimal value function Q_h^* has the linear form $\langle \boldsymbol{\phi}(\cdot, \cdot), \mathbf{w}_h^* \rangle$. Therefore, to estimate the Q_h^*, it suffices to estimate $\mathbf{w}_h^*$. At the beginning of each batch, Algorithm 3 calculates $\mathbf{w}_h^k$ as an estimate of $\mathbf{w}_h^*$ by ridge regression (Line 8).

Epistemic Uncertainty Estimate The basic idea to compute $\mathbf{w}_h^k$ is to utilize the Bellman optimality equation with an estimated transition dynamic $\widehat{\mathbb{P}}_h$ and an estimated reward function $\widehat{r}_h$, similar to the strategy we have adapted in Chapter 2. Due to the finite number of samples we can use at k-th episode, there exists uncertainty between the estimate $\widehat{r}_h, \widehat{\mathbb{P}}_h$ and the ground truth $r_h, \mathbb{P}_h$, which is the *epistemic uncertainty*. The uncertainty estimate is built as $\Gamma_h^k(\cdot, \cdot)$ (Line 9). Similar to the uncertainty estimate we have established in Chapter 2, we have the following lemma which suggests the uncertainty estimate is valid:

Lemma 3.4.1 (Lemma B.4, Jin et al. 2020)**.** There exists some constant c such that if we set

$\beta = cdH\sqrt{\log(dT/\delta)}$, then for any fixed policy π we have for all $(s, a, h, k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$ that

$$\left|\boldsymbol{\phi}(s,a)^\top(\mathbf{w}_h^k - \mathbf{w}_h^\pi) - \left(\mathbb{P}_h(V_{h+1}^k - V_{h+1}^\pi)\right)(s,a)\right| \leq \Gamma_h^k(s,a)$$

with probability at least $1 - \delta$.

Then Algorithm 3 sets the estimate $Q_h^k(\cdot,\cdot)$ as the summation of the linear function $\langle \boldsymbol{\phi}(\cdot,\cdot), \mathbf{w}_h^k \rangle$ and a Hoeffding-type exploration bonus term (the epistemic uncertainty estimate) $\Gamma_h^k(\cdot,\cdot)$ (Line 10), which is calculated based on the confidence radius β. Then it sets the policy π_h^k as the greedy policy with respect to Q_h^k. Within each batch, Algorithm 3 simply maintains the policy used in the previous episode without updating (Line 13). Apparently, the number of batches of Algorithm 3 is B.

Here we would like to make a comparison between our LSVI-UCB-Batch and other related algorithms. The most related algorithm is LSVI-UCB proposed in Jin et al. (2020). The main difference between LSVI-UCB-Batch and LSVI-UCB is the introduction of batches. In detail, when $B = K$, LSVI-UCB-Batch degenerates to LSVI-UCB. Another related algorithm is the SBUCB algorithm proposed by Han et al. (2020). Both LSVI-UCB-Batch and SBUCB take uniform batch grids as the selection of batches. The difference is that SBUCB is designed for linear bandits, which is a special case of episodic MDPs with $H = 1$.

Regret Analysis The following theorem presents the regret bound of Algorithm 3.

Theorem 3.4.2. There exists a constant $c > 0$ such that for any $\delta \in (0,1)$, if we set $\lambda = 1$, $\beta = cdH\sqrt{\log(2dT/\delta)}$, then under Assumption 3.3.2, the total regret of Algorithm 3 is bounded by

$$\begin{aligned}\text{Regret}(T) \leq & 2H\sqrt{T\log\left(\frac{2dT}{\delta}\right)} + \frac{dHT}{2B\log 2}\log\left(\frac{T}{dH}+1\right) \\ & + 4c\sqrt{2d^3H^3T\log\left(\frac{2dT}{\delta}\right)\log\left(\frac{T}{dH}+1\right)}\end{aligned}$$

Algorithm 3 LSVI-UCB-Batch

Require: Number of batches B, confidence radius β, regularization parameter λ

Set $b \leftarrow 1$, $t_i \leftarrow (i-1) \cdot \lfloor K/B \rfloor + 1, i \in [B]$

for episode $k = 1, 2, \ldots, K$ **do**

  Receive the initial state s_1^k

  if $k = t_b$ **then**

    $b \leftarrow b + 1$, $Q_{H+1}^k(\cdot, \cdot) \leftarrow 0$

    for stage $h = H, H-1, \ldots, 1$ **do**

      $\mathbf{\Lambda}_h^k \leftarrow \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top + \lambda \mathbf{I}$

      $\mathbf{w}_h^k \leftarrow (\mathbf{\Lambda}_h^k)^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot [r_h(s_h^\tau, a_h^\tau) + \max_{a \in \mathcal{A}} Q_{h+1}^k(s_{h+1}^\tau, a)]$

      $\Gamma_h^k(\cdot, \cdot) \leftarrow \beta \cdot [\boldsymbol{\phi}(\cdot, \cdot)^\top (\mathbf{\Lambda}_h^k)^{-1} \boldsymbol{\phi}(\cdot, \cdot)]^{1/2}$

      $Q_h^k(\cdot, \cdot) \leftarrow \min\{\boldsymbol{\phi}(\cdot, \cdot)^\top \mathbf{w}_h^k + \Gamma_h^k(\cdot, \cdot), H - h + 1\}^+$, $\pi_h^k(\cdot) \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} Q_h^k(\cdot, a)$

    end for

  else

    $Q_h^k \leftarrow Q_h^{k-1}$, $\pi_h^k \leftarrow \pi_h^{k-1}$, $\forall h \in [H]$

  end if

  for stage $h = 1 \ldots, H$ **do**

    Take the action $a_h^k \leftarrow \pi_h^k(s_h^k)$, receive the reward $r_h(s_h^k, a_h^k)$ and the next state s_{h+1}^k

  end for

end for

with probability at least $1-\delta$.

Theorem 3.4.2 suggests that the total regret of Algorithm 3 is bounded by $\widetilde{O}(\sqrt{d^3H^3T} + dHT/B)$. When $B = \Omega(\sqrt{T/dH})$, the regret of Algorithm 3 is $\widetilde{O}(\sqrt{d^3H^3T})$, which is the same as that of LSVI-UCB in Jin et al. (2020). However, it is worth noting that LSVI-UCB needs K batches, while Algorithm 3 only requires $\sqrt{T/dH}$ batches, which can be much smaller than K.

Next, we present a lower bound to show the dependency of the total regret on the number of batches for the batch learning model.

Theorem 3.4.3. Suppose that $B \geq (d-1)H/2$. Then for any batch learning algorithm with B batches, there exists a linear MDP such that the regret over the first T rounds is lower bounded by

$$\text{Regret}(T) = \Omega(dH\sqrt{T} + dHT/B).$$

Theorem 3.4.3 suggests that in order to obtain a standard $\sqrt{T}$-regret, the number of batches B should be at least in the order of $\Omega(\sqrt{T})$, which is similar to its counterpart for batched linear bandits (Han et al., 2020).

3.5 RL in the Rare Policy Switch Model

In this section, we consider the rare policy switch model, where the agent can adaptively choose the batch sizes according to the information collected during the learning process.

Algorithm: Epistemic Uncertainty-Inspired Policy Update Rule We first present our second algorithm, LSVI-UCB-RareSwitch, as illustrated in Algorithm 4. Again, due to the nature of linear MDPs, we only need to estimate $\mathbf{w}_h^*$ by ridge regression, and then calculate the optimistic action-value function using the Hoeffding-type exploration bonus $\Gamma_h^k(\cdot,\cdot)$ along with the confidence radius β. Note that the size of the bonus term in Q_h^k is

determined by $\mathbf{\Lambda}_h^k$. Intuitively speaking, the matrix $\mathbf{\Lambda}_h^k$ in Algorithm 4 represents how much information has been learned about the underlying MDP, and the agent only needs to switch the policy after collecting a significant amount of additional information. This is reflected by the determinant of $\mathbf{\Lambda}_h^k$, and the upper confidence bound will become tighter (shrink) as $\det(\mathbf{\Lambda}_h^k)$ increases. The determinant based criterion is similar to the idea of doubling trick, which has been used in the rarely switching OFUL algorithm for stochastic linear bandits (Abbasi-Yadkori et al., 2011), UCRL2 algorithm for tabular MDPs (Jaksch et al., 2010), and UCLK/UCLK+ for linear mixture MDPs in the discounted setting (Zhou et al., 2021b).

As shown in Algorithm 4, for each stage $h \in [H]$ the algorithm maintains a matrix $\mathbf{\Lambda}_h$ which is updated at each policy switch (Line 10). For every $k \in [K]$, we denote by b_k the episode from which the policy π_k is computed. This is consistent with the one defined in Algorithm 3 in Section 3.4. At the start of each episode k, the algorithm computes $\{\mathbf{\Lambda}_h^k\}_{h\in[H]}$ (Line 5) and then compares them with $\{\mathbf{\Lambda}_h\}_{h\in[H]}$ using the determinant-based criterion (Line 7). The agent switches the policy if there exists some $h \in [H]$ such that $\det(\mathbf{\Lambda}_h^k)$ has increased by some pre-determined parameter $\eta > 1$, followed by policy evaluation (Lines 11-13). Otherwise, the algorithm retains the previous policy (Line 16). Here the hyperparameter η controls the frequency of policy switch, and the total number of policy switches can be bounded by a function of η. In detail, the relationship between the determinant and epistemic uncertainty estimate is characterized by the following lemma:

Lemma 3.5.1 (Lemma 12, Abbasi-Yadkori et al. 2011)**.** Suppose $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ are two positive definite matrices satisfying that $\mathbf{A} \succeq \mathbf{B}$, then for any $\mathbf{x} \in \mathbb{R}^d$, we have $\|\mathbf{x}\|_{\mathbf{A}} \leq \|\mathbf{x}\|_{\mathbf{B}} \cdot \sqrt{\det(\mathbf{A})/\det(\mathbf{B})}$.

By Lemma 3.5.1, we know that

$$\frac{\Gamma_h^k(\cdot,\cdot)}{\Gamma_h^{k'}(\cdot,\cdot)} = \frac{\|(\mathbf{\Lambda}_h^k)^{-1/2}\boldsymbol{\phi}(\cdot,\cdot)\|_2}{\|(\mathbf{\Lambda}_h^{k'})^{-1/2}\boldsymbol{\phi}(\cdot,\cdot)\|_2} \leq \sqrt{\frac{\det(\mathbf{\Lambda}_h^{k'})}{\det(\mathbf{\Lambda}_h^k)}}, \tag{3.5.1}$$

for any $k < k'$. (3.5.1) suggests that as long as the ratio between the determinants $\det(\mathbf{\Lambda}_h^{k'})/\det(\mathbf{\Lambda}_h^k)$ is upper bounded, the ratio between the epistemic uncertainty estimates

Algorithm 4 LSVI-UCB-RareSwitch

Require: Policy switch parameter η, confidence radius β, regularization parameter λ

Initialize $\mathbf{\Lambda}_h = \mathbf{\Lambda}_h^0 = \lambda \mathbf{I}_d$ for all $h \in [H]$
for episode $k = 1, 2, \ldots, K$ **do**
  Receive the initial state s_1^k
  for stage $h = 1, 2, \cdots, H$ **do**
    $\mathbf{\Lambda}_h^k \leftarrow \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top + \lambda \mathbf{I}_d$
  end for
  if $\exists h \in [H], \det(\mathbf{\Lambda}_h^k) > \eta \cdot \det(\mathbf{\Lambda}_h)$ **then**
    $Q_{H+1}^k(\cdot, \cdot) \leftarrow 0$
    for step $h = H, H-1, \cdots, 1$ **do**
      $\mathbf{\Lambda}_h \leftarrow \mathbf{\Lambda}_h^k$
      $\mathbf{w}_h^k \leftarrow (\mathbf{\Lambda}_h^k)^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau) \cdot [r_h(s_h^\tau, a_h^\tau) + \max_{a \in \mathcal{A}} Q_{h+1}^k(s_{h+1}^\tau, a)]$
      $\Gamma_h^k(\cdot, \cdot) \leftarrow \beta \cdot [\boldsymbol{\phi}(\cdot, \cdot)^\top (\mathbf{\Lambda}_h^k)^{-1} \phi(\cdot, \cdot)]^{1/2}$
      $Q_h^k(\cdot, \cdot) \leftarrow \min\{\phi(\cdot, \cdot)^\top \mathbf{w}_h^k + \Gamma_h^k(\cdot, \cdot), H - h + 1\}^+$, $\pi_h^k(\cdot) \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} Q_h^k(\cdot, a)$
    end for
  else
    $Q_h^k \leftarrow Q_h^{k-1}$, $\pi_h^k \leftarrow \pi_h^{k-1}$, $\forall h \in [H]$
  end if
  for stage $h = 1 \ldots, H$ **do**
    Take the action $a_h^k \leftarrow \pi_h^k(s_h^k)$, receive the reward $r_h(s_h^k, a_h^k)$ and the next state s_{h+1}^k
  end for
end for

$\Gamma_h^k(\cdot,\cdot)/\Gamma_h^{k'}(\cdot,\cdot)$ is also upper bounded. That suggests we can use an 'older' epistemic uncertainty estimate instead of a newer one to reduce the computational cost caused by the policy update.

Algorithm 4 is also a variant of LSVI-UCB proposed in Jin et al. (2020). Compared with LSVI-UCB-Batch in Algorithm 3 for the batch learning model, LSVI-UCB-RareSwitch adaptively decides when to switch the policy and can be tuned by the hyperparameter η and therefore fits into the rare policy switch model.

Regret Analysis We present the regret bound of Algorithm 4 in the following theorem.

Theorem 3.5.2. There exists some constant $c > 0$ such that for any $\delta \in (0,1)$, if we set $\lambda = 1$, $\beta = cdH\sqrt{\log(2dT/\delta)}$ and $\eta = (1+K/d)^{dH/B}$, then the number of policy switches N_{switch} in Algorithm 4 will not exceed B. Moreover, the total regret of Algorithm 4 is bounded by

$$\text{Regret}(T) \leq 2H\sqrt{T\log\left(\frac{2dT}{\delta}\right)} + 2c\sqrt{2d^3H^3T}\cdot\sqrt{\left(\frac{T}{dH}+1\right)^{\frac{dH}{B}}\log\left(\frac{T}{dH}+1\right)\log\left(\frac{2dT}{\delta}\right)} \tag{3.5.2}$$

with probability at least $1-\delta$.

A few remarks are in order.

Remark 3.5.3. Algorithm 4 needs to update the value of each $\det(\mathbf{\Lambda}_h^k)$, and thanks to the special structure of $\mathbf{\Lambda}_h^k$, this can be done efficiently by applying the matrix determinant lemma along with the Sherman Morrison formula for efficiently updating each $(\mathbf{\Lambda}_h^k)^{-1}$. For simplicity and clarity of the presentation, we do not include these details in the pseudo-code.

Remark 3.5.4. By ignoring the non-dominating term, Theorem 3.5.2 suggests that the total regret of Algorithm 4 is bounded by $\widetilde{O}(\sqrt{d^3H^3T[1+T/(dH)]^{dH/B}})$. Also, if we are allowed to choose B, we can choose $B = \Omega(dH\log T)$ to achieve $\widetilde{O}(\sqrt{d^3H^3T})$ regret, which is the same

as that of LSVI-UCB in Jin et al. (2020). This also significantly improves upon Algorithm 3 when T is sufficiently large since previously we need $B = \Omega(\sqrt{T/dH})$. Our result exhibits a trade-off between the total regret bound and the number of policy switches, i.e., as the adaptivity budget B increases, the regret bound decreases. This will also be reflected by the numerical results later in Section 3.6.

Remark 3.5.5. Gao et al. (2021) proposed an algorithm with $B = \Omega(dH \log T)$ policy switches. Note that $B = \Omega(dH \log T)$ corresponds to choosing η to be a constant, which can be viewed as a special case of our algorithm. Their algorithm does not adapt to different values of budget B. Also, they did not study the batch learning model (Section 3.4) which we think is of equally important practical interest.

Remark 3.5.6. Gao et al. (2021) established a lower bound, which claims that any rare policy switch RL algorithm suffers a linear regret when $B = \widetilde{o}(dH)$. However, unlike our lower bound for the batch learning model (Theorem 3.4.3), their result does not provide a fine-grained regret lower bound for arbitrary adaptivity constraint B. It remains an open problem to establish such kind of lower bound for the rare policy switch model.

3.6 Numerical Experiment

In this section, we provide numerical experiments to support our theory. We run our algorithms, LSVI-UCB-Batch and LSVI-UCB-RareSwitch, on a synthetic linear MDP given in Example 3.6.1, and compare them with the fully adaptive baseline, LSVI-UCB (Jin et al., 2020).

Example 3.6.1 (Hard-to-learn linear MDP, Zhou et al. 2021b)**.** Let $d > 0$ be some integer and $\delta \in (0, 1)$ be a constant. The state space $\mathcal{S} = \{0, 1\}$ consists of two states, and the action space $\mathcal{A} = \{\pm 1\}^{d-3}$ contains 2^{d-3} actions where each action is represented by a $(d-3)$-dimensional vector $\mathbf{a}$. For each state-action pair $(s, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$, the feature vector is

given by

$$\boldsymbol{\phi}(s,a) = \begin{cases} (-\mathbf{a}^\top, 1-\delta, \delta)^\top & s = 0, \\ (0, \ldots, 0, \delta, 1-\delta) & s = 1. \end{cases} \tag{3.6.1}$$

For each $h \in [H]$, let $\boldsymbol{\gamma}_h \in \{\pm\delta/(d-2)\}^{d-2}$ and define the corresponding vector-valued measure as

$$\boldsymbol{\mu}_h(s) = \begin{cases} (\boldsymbol{\gamma}_h^\top, 1, 0)^\top & s = 0 \\ (-\boldsymbol{\gamma}_h^\top, 0, 1)^\top & s = 1 \end{cases}. \tag{3.6.2}$$

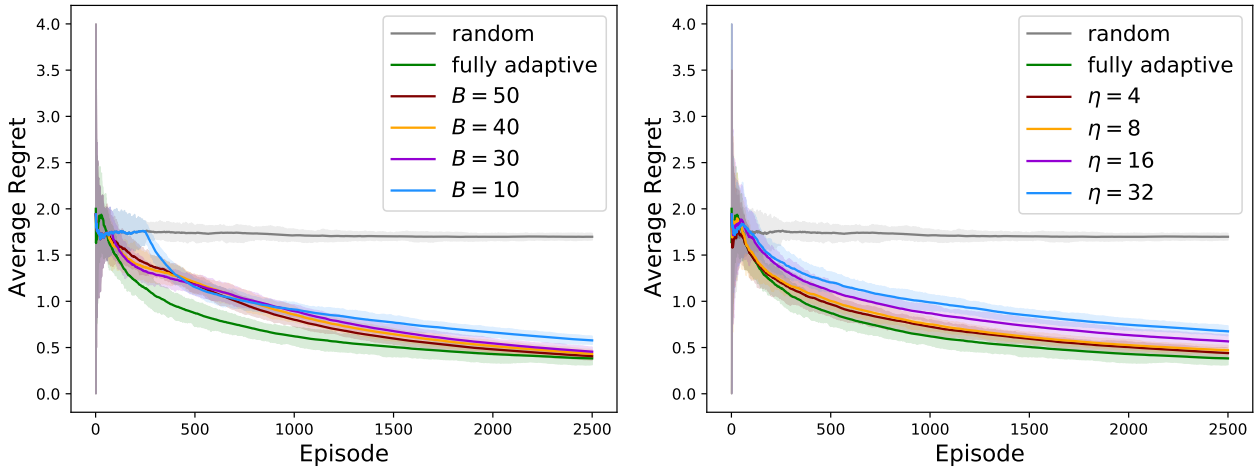
Finally, we set $\boldsymbol{\theta}_h \equiv (0, \ldots, 0, -\delta/(1-2\delta), (1-\delta)/(1-2\delta)) \in \mathbb{R}^d$ for all $h \in [H]$.

It is straightforward to verify that the feature vectors in (3.6.1) and the vector-valued measures in (3.6.2) constitute a valid linear MDP such that, for all $\mathbf{a} \in \mathcal{A}$ and $h \in [H]$,

$$r_h(s, \mathbf{a}) = \mathbb{1}\{s = 1\}, \qquad \mathbb{P}_h(s'|s, \mathbf{a}) = \begin{cases} 1 - \delta - \langle \mathbf{a}, \gamma_h \rangle & (s, s') = (0, 0), \\ \delta + \langle \mathbf{a}, \boldsymbol{\gamma}_h \rangle & (s, s') = (0, 1), \\ \delta & (s, s') = (1, 0), \\ 1 - \delta & (s, s') = (1, 1). \end{cases}$$

In our experiment[4], we set $H = 10$, $K = 2500$, $\delta = 0.35$ and $d = 13$, thus $\mathcal{A}$ contains 1024 actions. Now we apply our algorithms, LSVI-UCB-Batch and LSVI-UCB-RareSwitch, to this linear MDP instance, and compare their performance with the fully adaptive baseline LSVI-UCB (Jin et al., 2020) under different parameter settings. In detail, for LSVI-UCB-Batch, we run the algorithm for $B = 10, 20, 30, 40, 50$ respectively; for LSVI-UCB-RareSwitch, we set $\eta = 2, 4, 8, 16, 32$. We plot the average regret ($\text{Regret}(T)/K$) against the number of episodes in Figure 3.1. In addition to the regret of the proposed algorithms, we also plot the regret of a uniformly random policy (i.e., choosing actions uniformly randomly in each step) as a baseline.

[4]All experiments are performed on a PC with Intel i7-9700K CPU.



(a) LSVI-UCB-Batch

(b) LSVI-UCB-RareSwitch

Figure 3.1: Plot of average regret ($\text{Regret}(T)/K$) v.s. the number of episodes. The results are averaged over 50 rounds of each algorithm, and the error bars are chosen to be $[20\%, 80\%]$ empirical confidence intervals.

From Figure 3.1, we can see that for LSVI-UCB-Batch, when $B \approx \sqrt{K}$, it achieves a similar regret as the fully adaptive LSVI-UCB as it collects more and more trajectories. For LSVI-UCB-RareSwitch, a constant value of η yields a similar order of regret compared with LSVI-UCB as suggested by Theorem 3.5.2. By comparing Figure 3.1(a) and 3.1(b), we can see that the performance of LSVI-UCB-RareSwitch is consistently close to that of the fully-adaptive LSVI-UCB throughout the learning process, while the performance gap between LSVI-UCB-Batch and LSVI-UCB is small only when k is large. This suggests a better adaptivity of LSVI-UCB-RareSwitch than LSVI-UCB-Batch, which only updates the policy at prefixed time steps, thus being not adaptive enough.

Moreover, we can also see the trade-off between the regret and the adaptivity level: with more limited adaptivity (smaller B or larger η) the regret gap between our algorithms and the fully adaptive LSVI-UCB becomes larger. These results indicate that our algorithms can indeed achieve comparable performance as LSVI-UCB, even under adaptivity constraints.

This corroborates our theory.

3.7 Conclusion

In this chapter, we study online RL with linear function approximation under the adaptivity constraints. We consider both the batch learning model and the rare policy switch models and propose two new algorithms LSVI-UCB-Batch and LSVI-UCB-RareSwitch for each setting. We show that LSVI-UCB-Batch enjoys an $\widetilde{O}(\sqrt{d^3H^3T} + dHT/B)$ regret and LSVI-UCB-RareSwitch enjoys an $\widetilde{O}(\sqrt{d^3H^3T[1+T/(dH)]^{dH/B}})$ regret. Compared with the fully adaptive LSVI-UCB algorithm (Jin et al., 2020), our algorithms can achieve the same regret with a much fewer number of batches/policy switches. We also prove the regret lower bound for the batch learning learning model, which suggests that the dependency on B in LSVI-UCB-Batch is tight.

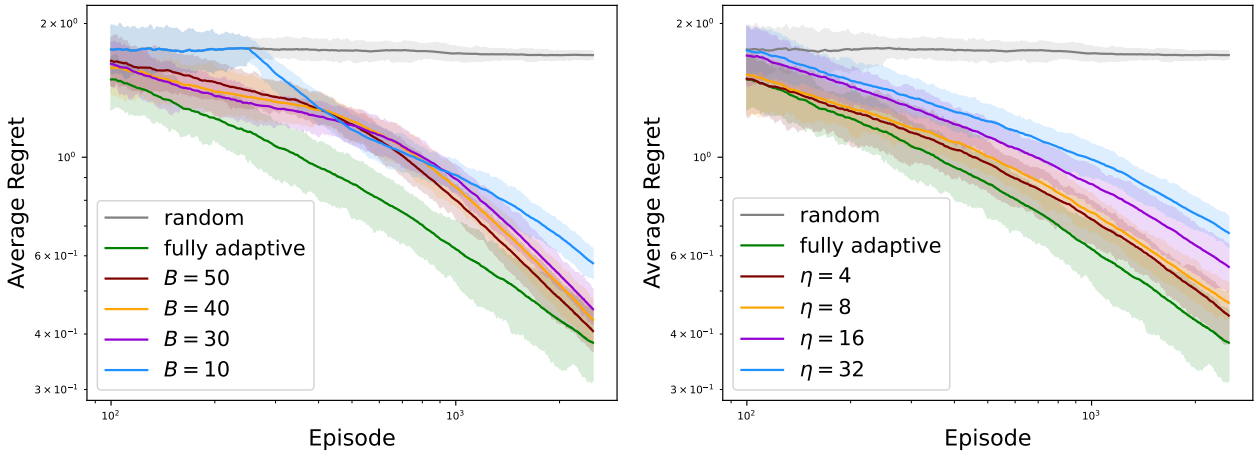
3.8 Additional Details on the Numerical Experiments

3.8.1 Log-scaled Plot of the Average Regret

We also provide log-scaled plot of the average regret in Figure 3.2. We can see that the slope of the average regret curves for our proposed algorithms is similar to that of the fully adaptive LSVI-UCB, all indicating an $\widetilde{O}(1/\sqrt{T})$ scaling.

3.8.2 Misspecified Linear MDP

We also empirically evaluate our algorithms on linear MDP with different levels of misspecification. In particular, based on the linear MDP instance constructed in Example 3.6.1, we follow the definition of ζ-approximate linear MDP in Jin et al. (2020), and consider a



(a) LSVI-UCB-Batch

(b) LSVI-UCB-RareSwitch

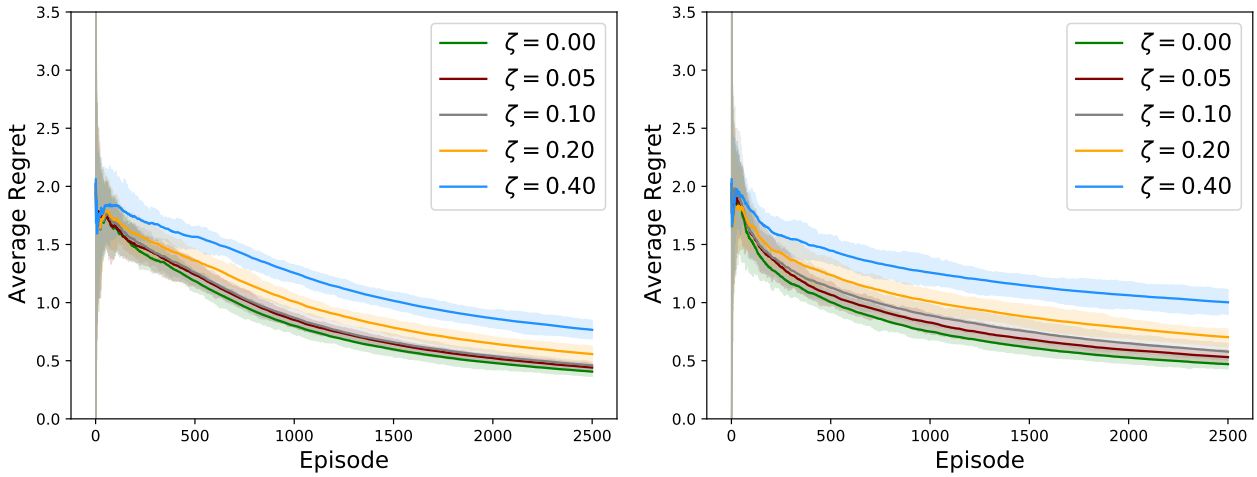
Figure 3.2: Plot of average regret ($\mathrm{Regret}(T)/K$) v.s. the number of episodes in log-scale. The results are averaged over 50 rounds of each algorithm, and the error bars are chosen to be $[20\%, 80\%]$ empirical confidence intervals.

corrupted transition given by

$$\mathbb{P}_h(s'|0,a) = (1 - f(a))\boldsymbol{\phi}(0,a)^\top \boldsymbol{\mu}_h(s') + f(a)\,\mathbb{1}\{s' = g(a)\}$$

where $f : \mathcal{A} \to [0, \zeta]$, $\zeta \in (0,1)$ and $g : \mathcal{A} \to \mathcal{S}$ are unknown. The two additional functions, f and g, can be constructed by random sampling before running the algorithms, and the magnitude of $\zeta \in (0,1)$ characterizes the level of model misspecification. All the other components of the model and the experiment configurations remain the same as those in Section 3.6.

Under this misspecified model with levels $\zeta = 0.05, 0.1, 0.2, 0.4$, we run LSVI-UCB-Batch with $B = 50$ and LSVI-UCB-RareSwitch with $\eta = 8$ respectively. We plot the average regret of the algorithms in Figure 3.3. We can see that our algorithms can still achieve a reasonably good performance under considerable levels of model misspecification.



(a) LSVI-UCB-Batch ($B = 50$)

(b) LSVI-UCB-RareSwitch ($\eta = 8$)

Figure 3.3: Plot of average regret ($\mathrm{Regret}(T)/K$) v.s. the number of episodes for a misspecified linear MDP. The results are averaged over 50 rounds of each algorithm, and the error bars are chosen to be $[20\%, 80\%]$ empirical confidence intervals.

3.9 Proofs of Theorem 3.4.2

In this section we prove Theorem 3.4.2. For simplicity, we use b_k to denote the batch t_b satisfying $t_b \le k < t_{b+1}$. Let $\Gamma_h^k(\cdot,\cdot)$ be $\beta \cdot [\boldsymbol{\phi}(\cdot,\cdot)^\top (\boldsymbol{\Lambda}_h^k)^{-1} \boldsymbol{\phi}(\cdot,\cdot)]^{1/2}$ for any $h \in [H], k \in [K]$. First, we need the following lemma which gives $\mathrm{Regret}(T)$ a high probability upper bound that depends on the summation of bonuses.

Lemma 3.9.1. With probability at least $1-\delta$, the total regret of Algorithm 3 satisfies

$$\mathrm{Regret}(T) \le \sum_{k=1}^{K} \sum_{h=1}^{H} \min\left\{H, 2\Gamma_h^{b_k}(s_h^k, a_h^k)\right\} + 2H\sqrt{T \log\left(\frac{2dT}{\delta}\right)}.$$

Lemma 3.9.1 suggests that in order to bound the total regret, it suffices to bound the summation of the 'delayed' bonuses $\Gamma_h^{b_k}(s_h^k, a_h^k)$, in contrast to the per-episode bonuses $\Gamma_h^k(s_h^k, a_h^k)$ for all $k \in [K]$. The superscript b_k suggests that instead of using all the information up to the current episode k, Algorithm 3 can only use the information before the current

batch b_k due to its batch learning nature. How to control the error induced by batch learning is the main difficulty in our analysis. To tackle this difficulty, we first need an upper bound for the summation of per-episode bonuses $\Gamma_h^k(s_h^k, a_h^k)$.

Lemma 3.9.2. Let β be selected as Theorem 3.4.2 suggests. Then the summation of all the per-episode bonuses is bounded by

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\Gamma_h^k(s_h^k, a_h^k) \leq \beta\sqrt{2dHT\log\left(\frac{T}{dH}+1\right)}.$$

It is worth noting that the per-episode bonuses are not generated from our algorithm, but instead are some virtual terms that we introduce to facilitate our analysis. Equipped with Lemma 3.9.2, we only need to bound the difference between delayed bonuses and per-episode bonuses. We consider all the indices $(k,h) \in [K] \times [H]$. The next lemma suggests that considering the ratio between delayed bonuses and per-episode bonuses, the 'bad' indices, where the ratio is large, only appear few times. This is also the key lemma of our analysis.

Lemma 3.9.3. Define the set $\mathcal{C}$ as follows

$$\mathcal{C} = \{(k,h) : \Gamma_h^{b_k}(s_h^k, a_h^k)/\Gamma_h^k(s_h^k, a_h^k) > 2\},$$

then we have $|\mathcal{C}| \leq dHK\log(K/d+1)/(2B\log 2)$.

With all the above lemmas, we now begin to prove our main theorem.

Proof of Theorem 3.4.2. Suppose the event defined in Lemma 3.9.1 holds. Then by Lemma 3.9.1 we have that

$$\text{Regret}(T) \leq \underbrace{\sum_{h=1}^{H}\sum_{k=1}^{K}\min\left\{H, 2\Gamma_h^{b_k}(s_h^k, a_h^k)\right\}}_{I} + 2H\sqrt{T\log\left(\frac{2dT}{\delta}\right)} \tag{3.9.1}$$

holds with probability at least $1-\delta$. Next, we are going to bound I. Let $\mathcal{C}$ be the set defined in Lemma 3.9.3. Then we have

$$I = \sum_{(k,h)\in\mathcal{C}}\min\left\{H, 2\Gamma_h^{b_k}(s_h^k, a_h^k)\right\} + \sum_{(k,h)\notin\mathcal{C}}\min\left\{H, 2\Gamma_h^{b_k}(s_h^k, a_h^k)\right\}$$

$$\leq H|\mathcal{C}| + 4 \sum_{(k,h)\notin\mathcal{C}} \Gamma_h^k(s_h^k, a_h^k)$$
$$\leq H|\mathcal{C}| + 4 \sum_{h=1}^{H}\sum_{k=1}^{K} \Gamma_h^k(s_h^k, a_h^k), \tag{3.9.2}$$

where the first inequality holds due to the definition of $\mathcal{C}$, and the second one holds trivially. Therefore, substituting (3.9.2) into (3.9.1), the regret can be bounded by

$$\begin{aligned}\text{Regret}(T) &\leq 2H\sqrt{T\log\left(\frac{2dT}{\delta}\right)} + H|\mathcal{C}| + 4\sum_{h=1}^{H}\sum_{k=1}^{K}\Gamma_h^k(s_h^k, a_h^k) \\ &\leq 2H\sqrt{T\log\left(\frac{2dT}{\delta}\right)} + \frac{dHT}{2B\log 2}\log\left(\frac{T}{dH}+1\right) \\ &\quad + 4c\sqrt{2d^3H^3T\log\left(\frac{2dT}{\delta}\right)\log\left(\frac{T}{dH}+1\right)},\end{aligned} \tag{3.9.3}$$

where the second inequality holds due to Lemmas 3.9.2 and 3.9.3 and the fact that $T = KH$. This completes the proof. □

3.9.1 Proof of Lemma 3.9.1

The following two lemmas in Jin et al. (2020) characterize the quality of the estimates given by the LSVI-UCB-type algorithms.

Lemma 3.9.4 (Lemma B.5, Jin et al. 2020)**.** With probability at least $1-\delta$, we have $Q_h^k(s,a) \geq Q_h^*(s,a)$ for all $(s,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$.

Proof of Lemma 3.9.1. By Lemma 3.9.4, we have $Q_h^k(s,a) \geq Q_h^*(s,a)$ for all $(s,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$ on some event $\mathcal{E}$ such that $\mathbb{P}(\mathcal{E}) \geq 1-\delta/2$. In the following argument, all statements would be conditioned on the event $\mathcal{E}$. Then by the definition of V_1^k we know that $V_1^k(s) = \max_{a\in\mathcal{A}} Q_1^k(s,a) \geq \max_{a\in\mathcal{A}} Q_1^*(s,a) = V_1^*(s)$ for all $(s,k) \in \mathcal{S}\times[K]$. Therefore, we have

$$\text{Regret}(T) = \sum_{k=1}^{K}\left[V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)\right] \leq \sum_{k=1}^{K}\left[V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\right] = \sum_{k=1}^{K}\left[V_1^{b_k}(s_1^k) - V_1^{\pi^k}(s_1^k)\right].$$

Note that

$$V_h^{b_k}(s_h^k) - V_h^{\pi^k}(s_h^k) = Q_h^{b_k}(s_h^k, a_h^k) - Q_h^{\pi^k}(s_h^k, a_h^k),$$

which together with the definition of $Q_h^{b_k}$ and Lemma 3.4.1 implies that

$$\begin{aligned} V_h^{b_k}(s_h^k) - V_h^{\pi^k}(s_h^k) &\leq \boldsymbol{\phi}(s_h^k, a_h^k)^\top \mathbf{w}_h^{b_k} - \boldsymbol{\phi}(s_h^k, a_h^k)^\top \mathbf{w}_h^{\pi^k} + \Gamma_h^{b_k}(s_h^k, a_h^k) \\ &\leq \left[\mathbb{P}_h\left(V_{h+1}^{b_k} - V_{h+1}^{\pi^k}\right)\right](s_h^k, a_h^k) + 2\Gamma_h^{b_k}(s_h^k, a_h^k), \end{aligned}$$

where the first inequality holds due to the algorithm design, the second one holds due to Lemma 3.4.1. Meanwhile, notice that $0 \leq V_h^{b_k}(s_h^k) - V_h^*(s_h^k) \leq V_h^{b_k}(s_h^k) - V_h^{\pi^k}(s_h^k) \leq H$, then we have

$$\begin{aligned} V_h^{b_k}(s_h^k) - V_h^{\pi^k}(s_h^k) &\leq \min\left\{H, \left[\mathbb{P}_h\left(V_{h+1}^{b_k} - V_{h+1}^{\pi^k}\right)\right](s_h^k, a_h^k) + 2\Gamma_h^{b_k}(s_h^k, a_h^k)\right\} \\ &\leq \left[\mathbb{P}_h\left(V_{h+1}^{b_k} - V_{h+1}^{\pi^k}\right)\right](s_h^k, a_h^k) + \min\left\{H, 2\Gamma_h^{b_k}(s_h^k, a_h^k)\right\} \\ &= V_{h+1}^{b_k}(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k) + \min\left\{H, 2\Gamma_h^{b_k}(s_h^k, a_h^k)\right\} \\ &\quad + \left[\mathbb{P}_h\left(V_{h+1}^{b_k} - V_{h+1}^{\pi^k}\right)\right](s_h^k, a_h^k) - \left(V_{h+1}^{b_k}(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)\right), \end{aligned}$$

where the second inequality holds since $V_{h+1}^{b_k} - V_{h+1}^{\pi^k} \geq 0$. Recursively expand the above inequality, and we have

$$\begin{aligned} V_1^{b_k}(s_1^k) - V_1^{\pi^k}(s_1^k) = &\sum_{h=1}^H \left\{\left[\mathbb{P}_h\left(V_{h+1}^{b_k} - V_{h+1}^{\pi^k}\right)\right](s_h^k, a_h^k) - \left(V_{h+1}^{b_k}(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)\right)\right\} \\ &+ \sum_{h=1}^H \min\left\{H, 2\Gamma_h^{b_k}(s_h^k, a_h^k)\right\}. \end{aligned}$$

Therefore, the total regret can be bounded as follows

$$\begin{aligned} \text{Regret}(T) \leq &\sum_{k=1}^K \sum_{h=1}^H \left\{\left[\mathbb{P}_h\left(V_{h+1}^{b_k} - V_{h+1}^{\pi^k}\right)\right](s_h^k, a_h^k) - \left(V_{h+1}^{b_k} - V_{h+1}^{\pi^k}\right)(s_{h+1}^k)\right\} \\ &+ \sum_{k=1}^K \sum_{h=1}^H \min\left\{H, 2\Gamma_h^{b_k}(s_h^k, a_h^k)\right\}. \end{aligned}$$

Note that conditional on $\mathcal{F}_{k,h,1}$, $V_{h+1}^{b_k}$ and $V_{h+1}^{\pi^k}$ are both deterministic, while s_{h+1}^k follows the distribution $\mathbb{P}_h(\cdot|s_h^k, a_h^k)$. Therefore, the first term on the RHS is a sum of a martingale

difference sequence such that each summand has absolute value at most $2H$. Applying Azuma-Hoeffding inequaliy yields

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left\{\left[\mathbb{P}_h\left(V_{h+1}^{b_k}-V_{h+1}^{\pi^k}\right)\right](s_h^k,a_h^k)-\left(V_{h+1}^{b_k}-V_{h+1}^{\pi^k}\right)(s_{h+1}^k)\right\}\leq 2H\sqrt{T\log\left(\frac{2dT}{\delta}\right)},$$

with probability at least $1-\delta/2$. By a union bound over the event $\mathcal{E}$ and the convergence of the martingale, with probability at least $1-\delta$, we have

$$\mathrm{Regret}(T)\leq 2H\sqrt{T\log\left(\frac{2dT}{\delta}\right)}+\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{H,2\Gamma_h^{b_k}(s_h^k,a_h^k)\right\}.$$

□

3.9.2 Proof of Lemma 3.9.2

We need the following lemma to bound the sum of the bonus terms.

Lemma 3.9.5 (Lemma 11, Abbasi-Yadkori et al. 2011)**.** Let $\{\boldsymbol{\phi}_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d-$valued sequence. Meanwhile, let $\boldsymbol{\Lambda}_0\in\mathbb{R}^{d\times d}$ be a positive-definite matrix and $\boldsymbol{\Lambda}_t=\boldsymbol{\Lambda}_0+\sum_{i=1}^{t-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top$. It holds for any $t\in\mathbb{Z}_+$ that

$$\sum_{i=1}^{t}\min\{1,\boldsymbol{\phi}_i^\top\boldsymbol{\Lambda}_i^{-1}\boldsymbol{\phi}_i\}\leq 2\log\left(\frac{\det(\boldsymbol{\Lambda}_{t+1})}{\det(\boldsymbol{\Lambda}_1)}\right).$$

Moreover, assuming that $\|\boldsymbol{\phi}_i\|_2\leq 1$ for all $i\in\mathbb{Z}_+$ and $\lambda_{\min}(\boldsymbol{\Lambda}_0)\geq 1$, it holds for any $t\in\mathbb{Z}_+$ that

$$\log\left(\frac{\det(\boldsymbol{\Lambda}_{t+1})}{\det(\boldsymbol{\Lambda}_1)}\right)\leq\sum_{i=1}^{t}\boldsymbol{\phi}_i^\top\boldsymbol{\Lambda}_i^{-1}\boldsymbol{\phi}_i\leq 2\log\left(\frac{\det(\boldsymbol{\Lambda}_{t+1})}{\det(\boldsymbol{\Lambda}_1)}\right).$$

Proof of Lemma 3.9.2. We can bound the summation of $\Gamma_h^k(s_h^k,a_h^k)$ as follows:

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\Gamma_h^k(s_h^k,a_h^k)\leq\sum_{h=1}^{H}\sqrt{K\cdot\sum_{k=1}^{K}[\Gamma_h^k(s_h^k,a_h^k)]^2}=\beta\sqrt{K}\sum_{h=1}^{H}\sqrt{\sum_{k=1}^{K}\boldsymbol{\phi}(s_h^k,a_h^k)^\top[\boldsymbol{\Lambda}_h^k]^{-1}\boldsymbol{\phi}(s_h^k,a_h^k)},$$

where the inequality holds due to Cauchy-Schwarz inequality. Furthermore, by Lemma 3.9.5, we have

$$\sum_{k=1}^{K} \boldsymbol{\phi}(s_h^k, a_h^k)^\top [\boldsymbol{\Lambda}_h^k]^{-1} \boldsymbol{\phi}(s_h^k, a_h^k) \leq 2 \log \left(\frac{\det \boldsymbol{\Lambda}_h^{K+1}}{\det \boldsymbol{\Lambda}_h^1} \right) \leq 2d \log(K/d + 1),$$

where the second inequality holds due to Lemma 3.10.1. That finishes our proof. □

3.9.3 Proof of Lemma 3.9.3

Proof of Lemma 3.9.3. First, let $\mathcal{C}_h$ denote the indices k where $(k, h) \in \mathcal{C}$, then we have $|\mathcal{C}| = \sum_{h=1}^{H} |\mathcal{C}_h|$. Next we bound $|\mathcal{C}_h|$ for each h. For each $k \in \mathcal{C}_h$, suppose $t_b \leq k < t_{b+1}$, then we have $b_k = t_b$ and

$$\log \det(\boldsymbol{\Lambda}_h^{t_{b+1}}) - \log \det(\boldsymbol{\Lambda}_h^{t_b}) \geq \log \det(\boldsymbol{\Lambda}_h^k) - \log \det(\boldsymbol{\Lambda}_h^{b_k}) \geq 2 \log(\Gamma_h^{b_k}(s_h^k, a_h^k)/\Gamma_h^k(s_h^k, a_h^k)) > 2 \log 2,$$

where the first inequality holds since $\boldsymbol{\Lambda}_h^{t_{b+1}} \succeq \boldsymbol{\Lambda}_h^k$, the second inequality holds due to Lemma 3.5.1, the third one holds due to the definition of $\mathcal{C}_h$. Thus, let $\widehat{\mathcal{C}}_h$ denote the set

$$\widehat{\mathcal{C}}_h = \{b \in [B] : \log \det(\boldsymbol{\Lambda}_h^{t_{b+1}}) - \log \det(\boldsymbol{\Lambda}_h^{t_b}) > 2 \log 2\},$$

we have $|\mathcal{C}_h| \leq \lfloor K/B \rfloor \cdot |\widehat{\mathcal{C}}_h|$. In the following we bound $|\widehat{\mathcal{C}}_h|$. Now we consider the sequence $\{\log \det(\boldsymbol{\Lambda}_h^{t_{b+1}}) - \log \det(\boldsymbol{\Lambda}_h^{t_b})\}$. It is easy to see $\log \det(\boldsymbol{\Lambda}_h^{t_{b+1}}) - \log \det(\boldsymbol{\Lambda}_h^{t_b}) \geq 0$, therefore

$$2 \log 2 |\widehat{\mathcal{C}}_h| \leq \sum_{b \in \widehat{\mathcal{C}}_h} [\log \det(\boldsymbol{\Lambda}_h^{t_{b+1}}) - \log \det(\boldsymbol{\Lambda}_h^{t_b})] \leq \sum_{b=1}^{B} [\log \det(\boldsymbol{\Lambda}_h^{t_{b+1}}) - \log \det(\boldsymbol{\Lambda}_h^{t_b})]. \quad (3.9.4)$$

Meanwhile, we have

$$\sum_{b=1}^{B} [\log \det(\boldsymbol{\Lambda}_h^{t_{b+1}}) - \log \det(\boldsymbol{\Lambda}_h^{t_b})] = \log \det(\boldsymbol{\Lambda}_h^{t_{B+1}}) = \log \det(\boldsymbol{\Lambda}_h^{K+1}) \leq d \log(K/d + 1), \quad (3.9.5)$$

where the last inequality holds due to Lemma 3.10.1. Therefore, (3.9.4) and (3.9.5) suggest that $|\widehat{\mathcal{C}}_h| \leq d \log(K/d + 1)/(2 \log 2)$. Finally, we bound $|\mathcal{C}|$ as follows, which ends our proof.

$$|\mathcal{C}| = \sum_{h=1}^{H} |\mathcal{C}_h| \leq \sum_{h=1}^{H} K/B \cdot |\widehat{\mathcal{C}}_h| \leq dHK \log(K/d + 1)/(2B \log 2).$$

□

3.10 Proof of Theorem 3.5.2

Now we provide the proof of Theorem 3.5.2. We continue to use the notions that have been introduced in Section 3.4. We first give an upper bound on the determinant of $\mathbf{\Lambda}_h^k$.

Lemma 3.10.1. Let $\{\mathbf{\Lambda}_h^k, (k,h) \in [K] \times [H]\}$ be as defined in Algorithms 3 and 4. Then for all $h \in [H]$ and $k \in [K]$, we have $\det(\mathbf{\Lambda}_h^k) \leq (\lambda + (k-1)/d)^d$.

Proof. Note that

$$\mathrm{tr}(\mathbf{\Lambda}_h^k) = \mathrm{tr}(\lambda \mathbf{I}_d) + \sum_{\tau=1}^{k-1} \mathrm{tr}\left(\boldsymbol{\phi}(s_h^\tau, a_h^\tau)\boldsymbol{\phi}(s_h^\tau, a_h^\tau)^\top\right) = \lambda d + \sum_{\tau=1}^{k-1} \|\boldsymbol{\phi}(s_h^\tau, a_h^\tau)\|_2^2 \leq \lambda d + k - 1,$$

where the inequality follows from the assumption that $\|\boldsymbol{\phi}(s,a)\|_2 \leq 1$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Since $\mathbf{\Lambda}_h^k$ is positive semi-definite, by inequality of arithmetic and geometric means, we have

$$\det(\mathbf{\Lambda}_h^k) \leq \left(\frac{\mathrm{tr}(\mathbf{\Lambda}_h^k)}{d}\right)^d \leq \left(\lambda + \frac{k-1}{d}\right)^d.$$

This finishes the proof. □

The switching cost of Algorithm 4 is characterized in the following lemma.

Lemma 3.10.2. For any $\eta > 1$ and $\lambda > 0$, the global switching cost of Algorithm 4 is bounded by

$$N_{\mathrm{switch}} \leq \frac{dH}{\log \eta} \log\left(1 + \frac{K}{\lambda d}\right).$$

Proof. Let $\{k_1, k_2, \cdots, k_{N_{\mathrm{switch}}}\}$ be the episodes where the algorithm updates the policy, and we also define $k_0 = 0$. Then by the determinant-based criterion (Line 7), for each $i \in [N_{\mathrm{switch}}]$ there exists at least one $h \in [H]$ such that

$$\det(\mathbf{\Lambda}_h^{k_i}) > \eta \cdot \det(\mathbf{\Lambda}_h^{k_{i-1}}).$$

By the definition of $\mathbf{\Lambda}_h^k$ (Line 5), we know that $\mathbf{\Lambda}_h^{j_1} \succeq \mathbf{\Lambda}_h^{j_2}$ for all $j_1 \geq j_2$ and $h \in [H]$. Thus we further have

$$\prod_{h=1}^{H} \det(\mathbf{\Lambda}_h^{k_i}) > \eta \cdot \prod_{h=1}^{H} \det(\mathbf{\Lambda}_h^{k_{i-1}}).$$

Applying the above inequality for all $i \in [N_{\text{switch}}]$ yields

$$\prod_{h=1}^{H} \det\left(\mathbf{\Lambda}_h^{k_{N_{\text{switch}}}}\right) > \eta^{N_{\text{switch}}} \cdot \prod_{h=1}^{H} \det(\mathbf{\Lambda}_h^0) = \eta^{N_{\text{switch}}} \lambda^{dH},$$

as we initialize $\mathbf{\Lambda}_h^0$ to be $\lambda \mathbf{I}_d$. While by Lemma 3.10.1, we have

$$\prod_{h=1}^{H} \det\left(\mathbf{\Lambda}_h^{k_{N_{\text{switch}}}}\right) \leq \prod_{h=1}^{H} \det(\mathbf{\Lambda}_h^K) \leq \left(\lambda + \frac{K}{d}\right)^{dH}.$$

Therefore, combining the above two inequalities, we obtain that

$$N_{\text{switch}} \leq \frac{dH}{\log \eta} \log\left(1 + \frac{K}{\lambda d}\right).$$

This completes the proof. □

We now begin to prove our main theorem.

Proof of Theorem 3.5.2. First, substituting the choice of η and $\lambda = 1$ into the bound in Lemma 3.10.2 yields that $N_{\text{switch}} \leq B$.

Next, we bound the regret of Algorithm 4. The result of Lemma 3.9.1 still holds here, thus it suffices to bound the summation of the bonus terms $\Gamma_h^{b_k}(s_h^k, a_h^k)$. Note that $b_k \leq k$, and thus $\mathbf{\Lambda}_h^k \succeq \mathbf{\Lambda}_h^{b_k}$ for all $(h, k) \in [H] \times [K]$. Then by Lemma 3.5.1 we have

$$\frac{\Gamma_h^{b_k}(s_h^k, a_h^k)}{\Gamma_h^k(s_h^k, a_h^k)} \leq \sqrt{\frac{\det(\mathbf{\Lambda}_h^k)}{\det(\mathbf{\Lambda}_h^{b_k})}} \leq \sqrt{\eta} \tag{3.10.1}$$

for all $(h, k) \in [H] \times [K]$, where the second inequality holds due to the algorithm design. Hence, we have

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \Gamma_h^{b_k}(s_h^k, a_h^k) \leq \sqrt{\eta} \sum_{k=1}^{K} \sum_{h=1}^{H} \Gamma_h^k(s_h^k, a_h^k) \leq \beta \sqrt{2\eta dHT \log\left(\frac{T}{dH} + 1\right)},$$

where the second inequality follows from Lemma 3.9.2. Therefore, we conclude by Lemma 3.9.1 that

$$\text{Regret}(T) \leq 2c\sqrt{2\eta d^3H^3T\log\left(\frac{T}{dH}+1\right)\log\left(\frac{2dT}{\delta}\right)} + 2H\sqrt{T\log\left(\frac{2dT}{\delta}\right)} \tag{3.10.2}$$

holds with probability at least $1-\delta$. Finally, substituting the choice of η into (3.10.2) finishes our proof. □

3.11 Proofs of Theorem 3.4.3

In this section, we prove the lower bound for the batch learning model.

Proof of Theorem 3.4.3. We prove the $\Omega(dH\sqrt{T})$ and $\Omega(dHT/B)$ lower bounds separately. The first term has been proved in Remark 2.9.1, Chapter 2. In the remaining of this proof, we prove the second term. We consider a class of MDPs parameterized by $\boldsymbol{\gamma} \in \Gamma \subset \mathbb{R}^{2dH}$, where Γ is defined as follows

$$\Gamma = \big\{(\mathbf{b}_{1,1}^\top, \cdots, \mathbf{b}_{H,d}^\top)^\top : \mathbf{b}_{i,j} \in \{(0,1)^\top, (1,0)^\top\}\big\}.$$

The MDP is defined as follows. The states space $\mathcal{S}$ consist of has $d+1$ states $x_0, \cdots, x_d$, and the action space $\mathcal{A}$ contains two actions $\mathbf{a}_1 = (0,1)^\top, \mathbf{a}_2 = (1,0)^\top$. For any $\boldsymbol{\gamma} = (\mathbf{b}_{1,1}^\top, \cdots, \mathbf{b}_{H,d}^\top)^\top$, the feature mapping is defined as

$$\boldsymbol{\phi}(x_0, \mathbf{a}_j) = (1, \underbrace{0, \cdots, 0}_{2d})^\top, \qquad \boldsymbol{\phi}(x_i, \mathbf{a}_j) = (1, \underbrace{0, \cdots, 0}_{2i-2}, \mathbf{a}_j^\top, \underbrace{0, \cdots, 0}_{2d-2i})^\top \in \mathbb{R}^{2d+1}$$

for every $i \in [d]$ and $j \in \{1,2\}$. We further define the vector-valued measures as

$$\boldsymbol{\mu}_h^{\boldsymbol{\gamma}}(x_0) = (1, -\mathbf{b}_{h,1}^\top, \cdots, -\mathbf{b}_{h,d}^\top)^\top, \qquad \boldsymbol{\mu}_h^{\boldsymbol{\gamma}}(x_i) = (\underbrace{0, \cdots, 0}_{2i-1}, \mathbf{b}_{h,i}^\top, \underbrace{0, \cdots, 0}_{2d-2i})^\top$$

for every $i \in [d]$, $j \in \{1,2\}$ and $h \in [H]$. Finally, for each $h \in [H]$, we define

$$\boldsymbol{\theta}_h = (0, \underbrace{1, \cdots, 1}_{2d})^\top \in \mathbb{R}^{2d+1}.$$

Thereby, for each $h \in [H]$, the transition $\mathbb{P}_h^{\boldsymbol{\gamma}}$ is defined as $\mathbb{P}_h^{\boldsymbol{\gamma}}(s'|s, \mathbf{a}) = \langle \boldsymbol{\phi}(s, \mathbf{a}), \boldsymbol{\mu}_h^{\boldsymbol{\gamma}}(s') \rangle$, and the reward function is $r_h(s, \mathbf{a}) = \langle \boldsymbol{\phi}(s, \mathbf{a}), \boldsymbol{\theta}_h \rangle$ for all $(s, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$. It is straightforward to see that the reward satisfies $r_h(x_0, \mathbf{a}) = 0$ and $r_h(x_i, \mathbf{a}) = 1$ for $i \in [d]$ and all $\mathbf{a} \in \mathcal{A}$. In addition, the starting state can be x_0 or x_i.

Based on the above definition, we have the following transition dynamic:

- x_0 is an absorbing state.
- For any $i \in [d]$, x_i can only transit to x_0 or x_i.
- For any episode starting from x_0, there is no regret.
- For any episode starting from some x_i with $i \in [d]$, suppose h is the first stage where the agent did not choose the "right" action $\mathbf{a} = \mathbf{b}_{h,i}$, then the regret for this episode is $H - h$.

Now we show that for any deterministic algorithm[5], there exists a $\boldsymbol{\gamma} \in \Gamma$ such that the regret is lower bounded by dHT/B. Suppose $1 = t_1 < \cdots < t_{B+1} = K + 1$. We can treat all episodes in the same batch as copies of one episode, because all actions taken by the agent, transitions and rewards are the same. When $B \geq dH$, there exists $\mathcal{C} = \{c_{1,1}, \cdots, c_{H,d}\} \subset [B]$ with $|\mathcal{C}| = dH$ such that

$$\sum_{h \in [H]} \sum_{j \in [d]} (t_{c_{h,j}+1} - t_{c_{h,j}}) \geq \frac{dHK}{B}.$$

For simplicity, we denote the i-th batch as the collection of episodes $\{t_i, \cdots, t_{i+1} - 1\}$. Now we carefully pick the starting state s_0^i for the episodes in the i-th batch.

- For any batch whose starting episode does not belong to $\mathcal{C}$, we set the starting states of the episodes in this batch as x_0. In other words, for $i \notin \mathcal{C}$, we set $s_0^{t_i} = \cdots = s_0^{t_{i+1}-1} = x_0$.

[5]The lower bound of random algorithms is lower bounded by the lower bound of deterministic algorithms according to Yao's minimax principle.

- For any batch whose starting episode lies in $\mathcal{C}$, for $i = c_{h,j} \in \mathcal{C}$, we set $s_0^{t_{c_{h,j}}} = \cdots = s_0^{t_{c_{h,j}+1}-1} = x_j$.

We consider the regret over batches $c_{1,i}, \cdots, c_{H,i}$. Since the algorithm, transition and reward are all deterministic, then the environment can predict the agent's selection. Specifically, suppose the agent will always take action $\mathbf{a}$ at h-th stage in the episodes belonging to the $c_{h,j}$-th batch, where $h \leq H/2$. Then the environment selects $\mathbf{b}_{h,j}$ as $(1,1)^\top - \mathbf{a}$, i.e., the other action. Therefore, the agent will always pick the "wrong" action when she firstly visits state x_j at h-th stage, which occurs at least $H - h \geq H/2$ regret. Moreover, since for the batch learning model, all the actions are decided at the beginning of each batch, then the $H/2$ regret will last $(t_{c_{h,j}+1} - t_{c_{h,j}})$ episodes. Taking the summation, we have

$$\text{Regret}(T) \geq \frac{H}{2} \cdot \sum_{h\in[H]} \sum_{j\in[d]} (t_{c_{h,j}+1} - t_{c_{h,j}}) \geq \frac{dHT}{2B}.$$

Finally, replacing d by $(d-1)/2$, we can convert our feature mapping from a $(2d+1)$-dimensional vector to a d-dimensional vector and complete the proof. □

CHAPTER 4

Efficient Uncertainty Estimation for Neural Contextual Bandits

4.1 Introduction

In this chapter, we propose efficient uncertainty estimation methods for neural contextual bandit problem. The contextual bandit problem is defined as follows: at round $t \in \{1, 2, \ldots, T\}$, an agent is presented with a set of K actions, each of which is associated with a d-dimensional feature vector. After choosing an action, the agent will receive a stochastic reward generated from some unknown distribution conditioned on the action's feature vector. The goal of the agent is to maximize the expected cumulative rewards over T rounds. Recently, deep neural networks (DNNs) (Goodfellow et al., 2016) have been introduced to learn the underlying reward function in contextual bandit problem, thanks to their strong representation power. We call these approaches collectively as *neural contextual bandit algorithms*. Given the fact that DNNs enable the agent to make use of nonlinear models with less domain knowledge, existing work (Riquelme et al., 2018; Zahavy and Mannor, 2019) study *neural-linear bandits*. That is, they use all but the last layers of a DNN as a feature map, which transforms contexts from the raw input space to a low-dimensional space, usually with better representation and less frequent updates. Then they learn a linear exploration policy on top of the last hidden layer of the DNN with more frequent updates. These attempts have achieved great empirical success, but no regret guarantees are provided. Starting from here, we propose a new algorithm called NeuralUCB, uses a neural network to learn the unknown reward

function, and follows a UCB strategy for exploration. At the core of the algorithm is the novel use of DNN-based random feature mappings to construct the UCB. Its regret analysis is built on recent advances on optimization and generalization of deep neural networks (Jacot et al., 2018; Arora et al., 2019; Cao and Gu, 2019). Crucially, the analysis makes no modeling assumptions about the reward function, other than that it be bounded. While the main focus of this chapter is theoretical, we also show in a few benchmark problems the effectiveness of NeuralUCB, and demonstrate its benefits against several representative baselines.

Our main contributions are as follows:

- We propose a neural contextual bandit algorithm that can be regarded as an extension of existing (generalized) linear bandit algorithms (Abbasi-Yadkori et al., 2011; Filippi et al., 2010; Li et al., 2010, 2017) to the case of arbitrary bounded reward functions. They key technique of our proposed algorithm is a *gradient-based* epistemic uncertainty estimate. Our proposed estimate is computationally efficient, and we show that it is indeed a valid epistemic uncertainty estimate.

- We prove that, under standard assumptions, our algorithm is able to achieve $\widetilde{O}(\widetilde{d}\sqrt{T})$ regret, where $\widetilde{d}$ is the effective dimension of a neural tangent kernel matrix and T is the number of rounds. The bound recovers the existing $\widetilde{O}(d\sqrt{T})$ regret for linear contextual bandit as a special case (Abbasi-Yadkori et al., 2011), where d is the dimension of context.

- We demonstrate empirically the effectiveness of the algorithm in both synthetic and benchmark problems.

4.2 Related Work

Contextual Bandits There is a line of extensive work on linear bandits (e.g., Abe et al., 2003; Auer, 2002; Abe et al., 2003; Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010; Li et al., 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011). Many of these algorithms are

based on the idea of upper confidence bounds, and are shown to achieve near-optimal regret bounds. Our algorithm is also based on UCB exploration, and the regret bound reduces to that of Abbasi-Yadkori et al. (2011) in the linear case.

To deal with nonlinearity, a few authors have considered generalized linear bandits (Filippi et al., 2010; Li et al., 2017; Jun et al., 2017), where the reward function is a composition of a linear function and a (nonlinear) link function. Such models are special cases of what we study in this work.

More general nonlinear bandits without making strong modeling assumptions have also be considered. One line of work is the family of expert learning algorithms (Auer et al., 2002; Beygelzimer et al., 2011) that typically have a time complexity linear in the number of experts (which in many cases can be exponential in the number of parameters).

A second approach is to reduce a bandit problem to supervised learning, such as the epoch-greedy algorithm (Langford and Zhang, 2008) that has a non-optimal $O(T^{2/3})$ regret. Later, Agarwal et al. (2014) develop an algorithm that enjoys a near-optimal regret, but relies on an oracle, whose implementation still requires proper modeling assumptions.

A third approach uses nonparametric modeling, such as perceptrons (Kakade et al., 2008), random forests (Féraud et al., 2016), Gaussian processes and kernels (Kleinberg et al., 2008; Srinivas et al., 2010; Krause and Ong, 2011; Bubeck et al., 2011). The most relevant is by Valko et al. (2013), who assumed that the reward function lies in an RKHS with bounded RKHS norm and developed a UCB-based algorithm. They also proved an $\widetilde{O}(\sqrt{\widetilde{d}T})$ regret, where $\widetilde{d}$ is a form of effective dimension similar to ours. Compared with these interesting works, our neural network-based algorithm avoids the need to carefully choose a good kernel or metric, and can be computationally more efficient in large-scale problems. Recently, Foster and Rakhlin (2020) proposed contextual bandit algorithms with regression oracles which achieve a dimension-independent $O(T^{3/4})$ regret. Compared with Foster and Rakhlin (2020), NeuralUCB achieves a dimension-dependent $\widetilde{O}(\widetilde{d}\sqrt{T})$ regret with a better dependence on the time horizon.

Neural Networks Substantial progress has been made to understand the expressive power of DNNs, in connection to the network depth (Telgarsky, 2015, 2016; Liang and Srikant, 2017; Yarotsky, 2017, 2018; Hanin, 2019), as well as network width (Lu et al., 2017; Hanin and Sellke, 2017). The present paper on neural contextual bandit algorithms is inspired by these theoretical justifications and empirical evidence in the literature.

Our regret analysis for NeuralUCB makes use of recent advances in optimizing a DNN. A series of works show that (stochastic) gradient descent can find global minima of the training loss (Li and Liang, 2018; Du et al., 2019c; Allen-Zhu et al., 2019; Du et al., 2019a; Zou et al., 2019; Zou and Gu, 2019). For the generalization of DNNs, a number of authors (Daniely, 2017; Cao and Gu, 2019, 2020; Arora et al., 2019; Chen et al., 2021) show that by using (stochastic) gradient descent, the parameters of a DNN are located in a particular regime and the generalization bound of DNNs can be characterized by the best function in the corresponding neural tangent kernel space (Jacot et al., 2018).

4.3 Problem Setting

We consider the stochastic K-armed contextual bandit problem, where the total number of rounds T is known. At round $t \in [T]$, the agent observes the context consisting of K feature vectors: $\{\mathbf{x}_{t,a} \in \mathbb{R}^d \mid a \in [K]\}$. The agent selects an action a_t and receives a reward r_{t,a_t}. For brevity, we denote by $\{\mathbf{x}^i\}_{i=1}^{TK}$ the collection of $\{\mathbf{x}_{1,1}, \mathbf{x}_{1,2}, \ldots, \mathbf{x}_{T,K}\}$. Our goal is to maximize the following *pseudo regret* (or *regret* for short):

$$R_T = \mathbb{E}\bigg[\sum_{t=1}^{T}(r_{t,a_t^*} - r_{t,a_t})\bigg], \tag{4.3.1}$$

where $a_t^* = \text{argmax}_{a \in [K]}\, \mathbb{E}[r_{t,a}]$ is the optimal action at round t that maximizes the expected reward.

This work makes the following assumption about reward generation: for any round t,

$$r_{t,a_t} = h(\mathbf{x}_{t,a_t}) + \xi_t, \tag{4.3.2}$$

where h is an unknown function satisfying $0 \leq h(\mathbf{x}) \leq 1$ for any $\mathbf{x}$, and ξ_t is ν-sub-Gaussian noise conditioned on $\mathbf{x}_{1,a_1}, \ldots, \mathbf{x}_{t-1,a_{t-1}}$ satisfying $\mathbb{E}\xi_t = 0$. The ν-sub-Gaussian assumption for ξ_t is standard in the stochastic bandit literature (e.g., Abbasi-Yadkori et al., 2011; Li et al., 2017), and is satisfied by, for example, any bounded noise. The bounded h assumption holds true when h belongs to linear functions, generalized linear functions, Gaussian processes, and kernel functions with bounded RKHS norm over a bounded domain, among others. The contextual bandit can be regarded as a simplified MDP with the planning horizon $H = 1$. The state and action information is summarized by the context vector $(s_1^t, a_1^t) = \mathbf{x}_{t,a_1^t}$, and the reward function $r(s_1^t, a_1^t) = h(\mathbf{x}_{t,a_1^t})$. Unlike Chapter 2 and 3, we assume the reward is stochastic in this chapter.

In order to learn the reward function h in (4.3.2), we propose to use a fully connected neural networks with depth $L \geq 2$:

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sqrt{m}\mathbf{W}_L \sigma\Big(\mathbf{W}_{L-1}\sigma\big(\cdots \sigma(\mathbf{W}_1\mathbf{x})\big)\Big), \tag{4.3.3}$$

where $\sigma(x) = \max\{x, 0\}$ is the rectified linear unit (ReLU) activation function, $\mathbf{W}_1 \in \mathbb{R}^{m\times d}, \mathbf{W}_i \in \mathbb{R}^{m\times m}, 2 \leq i \leq L-1, \mathbf{W}_L \in \mathbb{R}^{m\times 1}$, and $\boldsymbol{\theta} = [\mathrm{vec}(\mathbf{W}_1)^\top, \ldots, \mathrm{vec}(\mathbf{W}_L)^\top]^\top \in \mathbb{R}^p$ with $p = m + md + m^2(L-1)$. Without loss of generality, we assume that the width of each hidden layer is the same (i.e., m) for convenience in analysis. We denote the gradient of the neural network function by $\mathbf{g}(\mathbf{x}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} f(\mathbf{x}; \boldsymbol{\theta}) \in \mathbb{R}^p$.

4.4 The NeuralUCB Algorithm

The key idea of NeuralUCB (Algorithm 5) is to use a neural network $f(\mathbf{x}; \boldsymbol{\theta})$ to predict the reward of context $\mathbf{x}$, and upper confidence bounds computed from the network to guide exploration (Auer, 2002).

Initialization It initializes the network by randomly generating each entry of $\boldsymbol{\theta}$ from an appropriate Gaussian distribution: for $1 \leq l \leq L-1$, $\mathbf{W}_l$ is set to be $\begin{pmatrix} \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{pmatrix}$, where each entry of $\mathbf{W}$ is generated independently from $N(0, 4/m)$; $\mathbf{W}_L$ is set to $(\mathbf{w}^\top, -\mathbf{w}^\top)$, where each entry of $\mathbf{w}$ is generated independently from $N(0, 2/m)$.

Epistemic Uncertainty Estimate For the contextual bandit, we want to learn the unknown reward function $h(\mathbf{x})$ by the neural network function $f(\mathbf{x}; \boldsymbol{\theta})$. Thus the epistemic uncertainty is defined as $|f(\mathbf{x}; \boldsymbol{\theta}) - h(\mathbf{x})|$ since contextual bandit is a simplified version of MDP. To estimate the epistemic uncertainty, we first assume that $h(\mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta}^*)$, which suggests h can be fully approximated by the neural network function with an unknown parameter $\boldsymbol{\theta}^*$. Then by the first-order Taylor expansion, we can further bound $|f(\mathbf{x}; \boldsymbol{\theta}_t) - h(\mathbf{x})|$ as follows:

$$|f(\mathbf{x}; \boldsymbol{\theta}_t) - h(\mathbf{x})| = |f(\mathbf{x}; \boldsymbol{\theta}_t) - f(\mathbf{x}; \boldsymbol{\theta}^*)| \approx |\langle \mathbf{g}(\mathbf{x}; \boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta}^* \rangle| \leq \gamma_t \cdot \|\mathbf{g}(\mathbf{x}; \boldsymbol{\theta}_t)\|_{\mathbf{Z}_t^{-1}},$$

where $\boldsymbol{\theta}_t$ is the current neural network parameter, and γ_{t-1} is a positive scaling factor. $\gamma_t \cdot \|\mathbf{g}(\mathbf{x}; \boldsymbol{\theta}_t)\|_{\mathbf{Z}_t^{-1}}$ is then selected as the epistemic uncertainty estimate of the reward h. The effectiveness of the above approximation will be validated by Lemma 4.6.1, 4.6.2 and 4.6.3.

Learning At round t, Algorithm 5 observes the contexts for all actions, $\{\mathbf{x}_{t,a}\}_{a=1}^K$. $U_{t,a}$ is defined as the summation of the estimated reward function $f(\mathbf{x}; \boldsymbol{\theta}_{t-1})$ and the epistemic uncertainty estimate $\gamma_{t-1} \cdot \|\mathbf{g}(\mathbf{x}; \boldsymbol{\theta}_{t-1})\|_{\mathbf{Z}_{t-1}^{-1}}$. It then chooses action a_t with the largest $U_{t,a}$, and receives the corresponding reward r_{t,a_t}. At the end of round t, NeuralUCB updates $\boldsymbol{\theta}_t$ by applying Algorithm 6 to (approximately) minimize $L(\boldsymbol{\theta})$ using gradient descent, and updates γ_t. We choose gradient descent in Algorithm 6 for the simplicity of analysis, although the training method can be replaced by stochastic gradient descent with a more involved analysis (Allen-Zhu et al., 2019; Zou et al., 2019).

Algorithm 5 NeuralUCB

Input: Number of rounds T, regularization parameter λ, exploration parameter ν, confidence parameter δ, norm parameter S, step size η, number of gradient descent steps J, network width m, network depth L.

Initialization: Randomly initialize $\boldsymbol{\theta}_0$ as described in the text

Initialize $\mathbf{Z}_0 = \lambda \mathbf{I}$

for $t = 1, \ldots, T$ **do**

  Observe $\{\mathbf{x}_{t,a}\}_{a=1}^K$

  for $a = 1, \ldots, K$ **do**

    Compute $U_{t,a} = f(\mathbf{x}_{t,a}; \boldsymbol{\theta}_{t-1}) + \gamma_{t-1} \sqrt{\mathbf{g}(\mathbf{x}_{t,a}; \boldsymbol{\theta}_{t-1})^\top \mathbf{Z}_{t-1}^{-1} \mathbf{g}(\mathbf{x}_{t,a}; \boldsymbol{\theta}_{t-1})/m}$

    Let $a_t = \text{argmax}_{a \in [K]} U_{t,a}$

  end for

  Play a_t and observe reward r_{t,a_t}

  Compute $\mathbf{Z}_t = \mathbf{Z}_{t-1} + \mathbf{g}(\mathbf{x}_{t,a_t}; \boldsymbol{\theta}_{t-1}) \mathbf{g}(\mathbf{x}_{t,a_t}; \boldsymbol{\theta}_{t-1})^\top / m$

  Let $\boldsymbol{\theta}_t = \text{TrainNN}(\lambda, \eta, J, m, \{\mathbf{x}_{i,a_i}\}_{i=1}^t, \{r_{i,a_i}\}_{i=1}^t, \boldsymbol{\theta}_0)$

  Compute

$$\begin{aligned}\gamma_t = &\sqrt{1 + C_1 m^{-1/6} \sqrt{\log m} L^4 t^{7/6} \lambda^{-7/6}} \\ &\cdot \left(\nu \sqrt{\log \frac{\det \mathbf{Z}_t}{\det \lambda \mathbf{I}} + C_2 m^{-1/6} \sqrt{\log m} L^4 t^{5/3} \lambda^{-1/6} - 2 \log \delta} + \sqrt{\lambda} S \right) \\ &+ (\lambda + C_3 t L) \Big[(1 - \eta m \lambda)^{J/2} \sqrt{t/\lambda} + m^{-1/6} \sqrt{\log m} L^{7/2} t^{5/3} \lambda^{-5/3} (1 + \sqrt{t/\lambda}) \Big].\end{aligned}$$

end for

Algorithm 6 TrainNN($\lambda, \eta, U, m, \{\mathbf{x}_{i,a_i}\}_{i=1}^t, \{r_{i,a_i}\}_{i=1}^t, \boldsymbol{\theta}^{(0)}$)

1: **Input:** Regularization parameter λ, step size η, number of gradient descent steps U, network width m, contexts $\{\mathbf{x}_{i,a_i}\}_{i=1}^t$, rewards $\{r_{i,a_i}\}_{i=1}^t$, initial parameter $\boldsymbol{\theta}^{(0)}$.
2: Define $\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^t (f(\mathbf{x}_{i,a_i}; \boldsymbol{\theta}) - r_{i,a_i})^2/2 + m\lambda\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}\|_2^2/2$.
3: **for** $j = 0, \ldots, J-1$ **do**
4: $\quad \boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} - \eta \nabla \mathcal{L}(\boldsymbol{\theta}^{(j)})$
5: **end for**
6: **Return** $\boldsymbol{\theta}^{(J)}$.

Comparison with Existing Algorithms We compare NeuralUCB with other neural contextual bandit algorithms. Allesiardo et al. (2014) proposed NeuralBandit which consists of K neural networks. It uses a committee of networks to compute the score of each action and chooses an action with the ϵ-greedy strategy. In contrast, our NeuralUCB uses upper confidence bound-based exploration, which is more effective than ϵ-greedy. In addition, our algorithm only uses one neural network instead of K networks, thus can be computationally more efficient.

Lipton et al. (2018) used Thompson sampling on deep neural networks (through variational inference) in reinforcement learning; a variant is proposed by Azizzadenesheli et al. (2018) that works well on a set of Atari benchmarks. Riquelme et al. (2018) proposed NeuralLinear, which uses the first $L-1$ layers of a L-layer DNN to learn a representation, then applies Thompson sampling on the last layer to choose action. Zahavy and Mannor (2019) proposed a NeuralLinear with limited memory (NeuralLinearLM), which also uses the first $L-1$ layers of a L-layer DNN to learn a representation and applies Thompson sampling on the last layer. Instead of computing the exact mean and variance in Thompson sampling, NeuralLinearLM only computes their approximation. Unlike NeuralLinear and NeuralLinearLM, NeuralUCB uses the entire DNN to learn the representation and constructs the upper confidence bound based on the random feature mapping defined by the neural network gradient. Finally, Kveton et al. (2020) studied the use of reward perturbation for exploration in neural network-based

bandit algorithms.

A Variant of NeuralUCB called NeuralUCB$_0$ is described in Section 4.13. It can be viewed as a simplified version of NeuralUCB where only the first-order Taylor approximation of the neural network around the initialized parameter is updated through online ridge regression. In this sense, NeuralUCB$_0$ can be seen as KernelUCB (Valko et al., 2013) specialized to the Neural Tangent Kernel (Jacot et al., 2018), or LinUCB (Li et al., 2010) with Neural Tangent Random Features (Cao and Gu, 2019).

While this variant has a comparable regret bound as NeuralUCB, we expect the latter to be stronger in practice. Indeed, as shown by Allen-Zhu and Li (2019), the Neural Tangent Kernel does not seem to completely realize the representation power of neural networks in supervised learning. A similar phenomenon will be demonstrated for contextual bandit learning in Section 4.7.

4.5 Regret Analysis

This section analyzes the regret of NeuralUCB. Recall that $\{\mathbf{x}^i\}_{i=1}^{TK}$ is the collection of all $\{\mathbf{x}_{t,a}\}$. Our regret analysis is built upon the recently proposed neural tangent kernel matrix (Jacot et al., 2018):

Definition 4.5.1 (Jacot et al. (2018); Cao and Gu (2019))**.** Let $\{\mathbf{x}^i\}_{i=1}^{TK}$ be a set of contexts. Define

$$\widetilde{\mathbf{H}}_{i,j}^{(1)} = \mathbf{\Sigma}_{i,j}^{(1)} = \langle \mathbf{x}^i, \mathbf{x}^j \rangle, \qquad \mathbf{A}_{i,j}^{(l)} = \begin{pmatrix} \mathbf{\Sigma}_{i,i}^{(l)} & \mathbf{\Sigma}_{i,j}^{(l)} \\ \mathbf{\Sigma}_{i,j}^{(l)} & \mathbf{\Sigma}_{j,j}^{(l)} \end{pmatrix},$$

$$\mathbf{\Sigma}_{i,j}^{(l+1)} = 2\mathbb{E}_{(u,v)\sim N(\mathbf{0},\mathbf{A}_{i,j}^{(l)})} \left[\sigma(u)\sigma(v)\right],$$

$$\widetilde{\mathbf{H}}_{i,j}^{(l+1)} = 2\widetilde{\mathbf{H}}_{i,j}^{(l)} \mathbb{E}_{(u,v)\sim N(\mathbf{0},\mathbf{A}_{i,j}^{(l)})} \left[\sigma'(u)\sigma'(v)\right] + \mathbf{\Sigma}_{i,j}^{(l+1)}.$$

Then, $\mathbf{H} = (\widetilde{\mathbf{H}}^{(L)} + \mathbf{\Sigma}^{(L)})/2$ is called the *neural tangent kernel (NTK)* matrix on the context set.

In the above definition, the Gram matrix $\mathbf{H}$ of the NTK on the contexts $\{\mathbf{x}^i\}_{i=1}^{TK}$ for L-layer neural networks is defined recursively from the input layer all the way to the output layer of the network. Interested readers are referred to Jacot et al. (2018) for more details about neural tangent kernels.

With Definition 4.5.1, we may state the following assumption on the contexts: $\{\mathbf{x}^i\}_{i=1}^{TK}$.

Assumption 4.5.2. $\mathbf{H} \succeq \lambda_0 \mathbf{I}$. Moreover, for any $1 \leq i \leq TK$, $\|\mathbf{x}^i\|_2 = 1$ and $[\mathbf{x}^i]_j = [\mathbf{x}^i]_{j+d/2}$.

The first part of the assumption says that the neural tangent kernel matrix is non-singular, a mild assumption commonly made in the related literature (Du et al., 2019a; Arora et al., 2019; Cao and Gu, 2019). It can be satisfied as long as *no* two contexts in $\{\mathbf{x}^i\}_{i=1}^{TK}$ are parallel. The second part is also mild and is just for convenience in analysis: for any context $\mathbf{x}$, $\|\mathbf{x}\|_2 = 1$, we can always construct a new context $\mathbf{x}' = [\mathbf{x}^\top, \mathbf{x}^\top]^\top / \sqrt{2}$ to satisfy Assumption 4.5.2. It can be verified that if $\boldsymbol{\theta}_0$ is initialized as in NeuralUCB, then $f(\mathbf{x}^i; \boldsymbol{\theta}_0) = 0$ for any $i \in [TK]$.

Next we define the effective dimension of the neural tangent kernel matrix.

Definition 4.5.3. The effective dimension $\widetilde{d}$ of the neural tangent kernel matrix on contexts $\{\mathbf{x}^i\}_{i=1}^{TK}$ is defined as

$$\widetilde{d} = \frac{\log \det(\mathbf{I} + \mathbf{H}/\lambda)}{\log(1 + TK/\lambda)}. \tag{4.5.1}$$

Remark 4.5.4. The notion of effective dimension was first introduced by Valko et al. (2013) for analyzing kernel contextual bandits, which was defined by the eigenvalues of any kernel matrix restricted to the given contexts. We adapt a similar but different definition of Yang and Wang (2020), which was used for the analysis of kernel-based Q-learning. Suppose the dimension of the reproducing kernel Hilbert space induced by the given kernel is $\widehat{d}$ and the feature mapping $\boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^{\widehat{d}}$ induced by the given kernel satisfies $\|\boldsymbol{\psi}(\mathbf{x})\|_2 \leq 1$ for any $\mathbf{x} \in \mathbb{R}^d$. Then, it can be verified that $\widetilde{d} \leq \widehat{d}$, as shown in Section 4.9.1. Intuitively, $\widetilde{d}$ measures how quickly the eigenvalues of $\mathbf{H}$ diminish, and only depends on T logarithmically in several special cases (Valko et al., 2013).

Now we are ready to present the main result, which provides the regret bound R_T of Algorithm 5.

Theorem 4.5.5. Let $\widetilde{d}$ be the effective dimension, and $\mathbf{h} = [h(\mathbf{x}^i)]_{i=1}^{TK} \in \mathbb{R}^{TK}$. There exist constant $C_1, C_2 > 0$, such that for any $\delta \in (0,1)$, if

$$m \geq \mathrm{poly}(T, L, K, \lambda^{-1}, \lambda_0^{-1}, S^{-1}, \log(1/\delta)),\ \eta = C_1(mTL + m\lambda)^{-1},$$

$\lambda \geq \max\{1, S^{-2}\}$, and $S \geq \sqrt{2\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{h}}$, then with probability at least $1-\delta$, the regret of Algorithm 5 satisfies

$$\begin{aligned} R_T \leq 3\sqrt{T}\sqrt{\widetilde{d}\log(1+TK/\lambda)+2} \cdot \Bigg[\nu\sqrt{\widetilde{d}\log(1+TK/\lambda)+2-2\log\delta} \\ + (\lambda + C_2TL)(1-\lambda/(TL))^{J/2}\sqrt{T/\lambda} + 2\sqrt{\lambda}S\Bigg] + 1. \end{aligned} \tag{4.5.2}$$

Remark 4.5.6. It is worth noting that, simply applying results for linear bandits to our algorithm would lead to a linear dependence of p or $\sqrt{p}$ in the regret. Such a bound is vacuous since in our setting p would be very large compared with the number of rounds T and the input context dimension d. In contrast, our regret bound only depends on $\widetilde{d}$, which can be much smaller than p.

Remark 4.5.7. Our regret bound (4.5.2) has a term $(\lambda + C_2TL)(1-\lambda/(TL))^{J/2}\sqrt{T/\lambda}$, which characterizes the optimization error of Algorithm 6 after J iterations. Setting

$$J = 2\log\frac{\lambda S}{\sqrt{T}(\lambda + C_2TL)}\frac{TL}{\lambda} = \widetilde{O}(TL/\lambda), \tag{4.5.3}$$

which is independent of m, we have $(\lambda + C_2TL)(1-\lambda/(TL))^{J/2}\sqrt{T/\lambda} \leq \sqrt{\lambda}S$, so the optimization error is dominated by $\sqrt{\lambda}S$. Hence, the order of the regret bound is not affected by the error of optimization.

Remark 4.5.8. With ν and λ treated as constants, $S = \sqrt{2\mathbf{h}^\top\mathbf{H}^{-1}\mathbf{h}}$, and J given in (4.5.3), the regret bound (4.5.2) becomes $R_T = \widetilde{O}\Big(\sqrt{\widetilde{d}T}\sqrt{\max\{\widetilde{d}, \mathbf{h}^\top\mathbf{H}^{-1}\mathbf{h}\}}\Big)$. Specifically, if h belongs to the RKHS $\mathcal{H}$ induced by the neural tangent kernel with bounded RKHS norm

$\|h\|_{\mathcal{H}}$, we have $\|h\|_{\mathcal{H}} \geq \sqrt{\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{h}}$; see Section 4.9.2 for more details. Thus our regret bound can be further written as

$$R_T = \widetilde{O}\Big(\sqrt{\widetilde{d}T}\sqrt{\max\{\widetilde{d}, \|h\|_{\mathcal{H}}\}}\Big). \tag{4.5.4}$$

The high-probability result in Theorem 4.5.5 can be used to obtain a bound on the expected regret.

Corollary 4.5.9. Under the same conditions in Theorem 4.5.5, there exists a positive constant C such that

$$\begin{aligned}\mathbb{E}[R_T] \leq 2 + 3\sqrt{T}\sqrt{\widetilde{d}\log(1+TK/\lambda)+2} \cdot \Bigg[\nu\sqrt{\widetilde{d}\log(1+TK/\lambda)+2+2\log T} \\ + 2\sqrt{\lambda}S + (\lambda + CTL)(1-\lambda/(TL))^{J/2}\sqrt{T/\lambda}\Bigg].\end{aligned}$$

4.6 Proof of Main Result

This section outlines the proof of Theorem 4.5.5, which has to deal with the following technical challenges:

- We do not make parametric assumptions on the reward function as some previous work (Filippi et al., 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011).
- To avoid strong parametric assumptions, we use overparameterized neural networks, which implies m (and thus p) is very large. Therefore, we need to make sure the regret bound is independent of m.
- Unlike the *static* feature mapping used in kernel bandit algorithms (Valko et al., 2013), NeuralUCB uses a neural network $f(\mathbf{x};\boldsymbol{\theta}_t)$ and its gradient $\mathbf{g}(\mathbf{x};\boldsymbol{\theta}_t)$ as a *dynamic* feature mapping depending on $\boldsymbol{\theta}_t$. This difference makes the analysis of NeuralUCB more difficult.

These challenges are addressed by the following technical lemmas, whose proofs are gathered in Section 4.10.1.

Lemma 4.6.1. There exists a positive constant $\bar{C}$ such that for any $\delta \in (0,1)$, if $m \geq \bar{C}T^4K^4L^6\log(T^2K^2L/\delta)/\lambda_0^4$, then with probability at least $1-\delta$, there exists a $\boldsymbol{\theta}^* \in \mathbb{R}^p$ such that

$$h(\mathbf{x}^i) = \langle \mathbf{g}(\mathbf{x}^i;\boldsymbol{\theta}_0), \boldsymbol{\theta}^* - \boldsymbol{\theta}_0 \rangle,\ \sqrt{m}\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_2 \leq \sqrt{2\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{h}}, \tag{4.6.1}$$

for all $i \in [TK]$.

Lemma 4.6.1 suggests that with high probability, the reward function restricted to $\{\mathbf{x}^i\}_{i=1}^{TK}$ can be regarded as a linear function of $\mathbf{g}(\mathbf{x}^i;\boldsymbol{\theta}_0)$ parameterized by $\boldsymbol{\theta}^* - \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}^*$ lies in a ball centered at $\boldsymbol{\theta}_0$. Note that here $\boldsymbol{\theta}^*$ is not a ground truth parameter for the reward function. Instead, it is introduced only for the sake of analysis. Equipped with Lemma 4.6.1, we can utilize existing results on linear bandits (Abbasi-Yadkori et al., 2011) to show that with high probability, $\boldsymbol{\theta}^*$ lies in the sequence of confidence sets.

Lemma 4.6.2. There exist positive constants $\bar{C}_1$ and $\bar{C}_2$ such that for any $\delta \in (0,1)$, if $\eta \leq \bar{C}_1(TmL + m\lambda)^{-1}$ and

$$m \geq \bar{C}_2 \max\big\{T^7\lambda^{-7}L^{21}(\log m)^3, \lambda^{-1/2}L^{-3/2}(\log(TKL^2/\delta))^{3/2}\big\},$$

then with probability at least $1-\delta$, we have $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 \leq 2\sqrt{t/(m\lambda)}$ and $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|_{\mathbf{Z}_t} \leq \gamma_t/\sqrt{m}$ for all $t \in [T]$, where γ_t is defined in Algorithm 5.

Lemma 4.6.3. There exists positive constants $\bar{C}_1, \bar{C}_2$ such that for any $\delta \in (0,1)$, if η and m satisfy the same conditions as in Lemma 4.6.2, then with probability at least $1-\delta$, for all $\mathbf{x} \in [\mathbf{x}^1, \ldots, \mathbf{x}^{TK}]$, we have

$$\begin{aligned}|h(\mathbf{x}) - f(\mathbf{x};\boldsymbol{\theta}_{t-1})| \leq \gamma_{t-1}\|\mathbf{g}(\mathbf{x};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}} + \bar{C}_1(m^{-1/6}\sqrt{\log m}t^{2/3}\lambda^{-2/3}L^3 \\ + \sqrt{\mathbf{h}^\top\mathbf{H}^{-1}\mathbf{h}}m^{-1/6}\sqrt{\log m}t^{1/6}\lambda^{-1/6}L^{7/2}).\end{aligned} \tag{4.6.2}$$

Furthermore, let $a_t^* = \mathrm{argmax}_{a\in[K]} h(\mathbf{x}_{t,a})$, we have

$$h\big(\mathbf{x}_{t,a_t^*}\big) - h\big(\mathbf{x}_{t,a_t}\big) \leq 2\gamma_{t-1}\min\left\{\|\mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}}, 1\right\}$$

$$+ \bar{C}\big(Sm^{-1/6}\sqrt{\log m}T^{7/6}\lambda^{-1/6}L^{7/2} + m^{-1/6}\sqrt{\log m}T^{5/3}\lambda^{-2/3}L^3\big). \tag{4.6.3}$$

Lemma 4.6.3 gives upper bounds for $h(\mathbf{x}) - f(\mathbf{x};\boldsymbol{\theta}_{t-1})$ and $h\big(\mathbf{x}_{t,a_t^*}\big) - h\big(\mathbf{x}_{t,a_t}\big)$. The first bound shows that up to some constants, the term $\gamma_{t-1}\|\mathbf{g}(\mathbf{x};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}}$ serves as an estimator of the *epistemic uncertainty*, which is the difference between the true reward $h(\mathbf{x})$ and the estimated reward $f(\mathbf{x};\boldsymbol{\theta}_{t-1})$ through neural networks. The second bound can be used to bound the regret R_T. It is worth noting that γ_t has a term $\log\det\mathbf{Z}_t$. A trivial upper bound of $\log\det\mathbf{Z}_t$ would result in a quadratic dependence on the network width m, since the dimension of $\mathbf{Z}_t$ is $p = md + m^2(L-2) + m$. Instead, we use the next lemma to establish an m-independent upper bound. The dependence on $\widetilde{d}$ is similar to Valko et al. (2013, Lemma 4), but the proof is different as our notion of effective dimension is different.

Lemma 4.6.4. There exist positive constants $\{\bar{C}_i\}_{i=1}^3$ such that for any $\delta \in (0,1)$, if $m \geq \bar{C}_1 \max\big\{T^7\lambda^{-7}L^{21}(\log m)^3, T^6K^6L^6(\log(TKL^2/\delta))^{3/2}\big\}$ and $\eta \leq \bar{C}_2(TmL + m\lambda)^{-1}$, then with probability at least $1-\delta$, we have

$$\begin{aligned}
&\sqrt{\sum_{t=1}^{T}\gamma_{t-1}^2 \min\left\{\|\mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}}^2, 1\right\}} \\
&\leq \sqrt{\widetilde{d}\log(1+TK/\lambda)+\Gamma_1}\cdot\bigg[\Gamma_2\bigg(\nu\sqrt{\widetilde{d}\log(1+TK/\lambda)+\Gamma_1-2\log\delta}+\sqrt{\lambda}S\bigg) \\
&\quad + (\lambda+\bar{C}_3tL)\Big[(1-\eta m\lambda)^{J/2}\sqrt{T/\lambda}+\Gamma_3(1+\sqrt{T/\lambda})\Big]\bigg],
\end{aligned}$$

where

$$\begin{aligned}
\Gamma_1 &= 1 + \bar{C}_3 m^{-1/6}\sqrt{\log m}L^4T^{5/3}\lambda^{-1/6}, \\
\Gamma_2 &= \sqrt{1 + \bar{C}_3 m^{-1/6}\sqrt{\log m}L^4T^{7/6}\lambda^{-7/6}}, \\
\Gamma_3 &= m^{-1/6}\sqrt{\log m}L^{7/2}T^{5/3}\lambda^{-5/3}.
\end{aligned}$$

We are now ready to prove the main result.

Proof of Theorem 4.5.5. Lemma 4.6.3 implies that the total regret R_T can be bounded as follows with a constant $C_1 > 0$:

$$\begin{aligned}
R_T &= \sum_{t=1}^{T} \big[h\big(\mathbf{x}_{t,a_t^*}\big) - h\big(\mathbf{x}_{t,a_t}\big)\big] \\
&\leq 2\sum_{t=1}^{T} \gamma_{t-1} \min\left\{\|\mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}}, 1\right\} + C_1\big(Sm^{-1/6}\sqrt{\log m}T^{13/6}\lambda^{-1/6}L^{7/2} \\
&\quad + m^{-1/6}\sqrt{\log m}T^{8/3}\lambda^{-2/3}L^3\big).
\end{aligned}$$

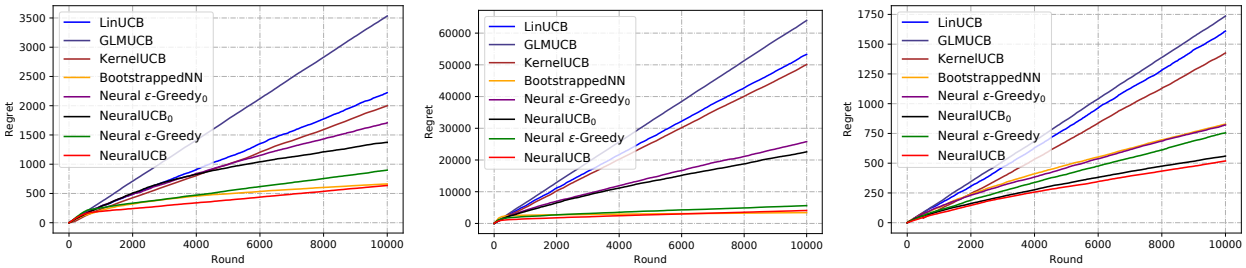
It can be further bounded as follows:

$$\begin{aligned}
R_T &\leq 2\sqrt{T\sum_{t=1}^{T} \gamma_{t-1}^2 \min\left\{\|\mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}}^2, 1\right\}} \\
&\quad + C_1\big(Sm^{-1/6}\sqrt{\log m}T^{13/6}\lambda^{-1/6}L^{7/2} + m^{-1/6}\sqrt{\log m}T^{8/3}\lambda^{-2/3}L^3\big) \\
&\leq 2\sqrt{T}\cdot\sqrt{\widetilde{d}\log(1+TK/\lambda) + \Gamma_1} \\
&\quad \bigg[\Gamma_2\bigg(\nu\sqrt{\widetilde{d}\log(1+TK/\lambda) + \Gamma_1 - 2\log\delta} + \sqrt{\lambda}S\bigg) + (\lambda + C_2TL)\Big[(1-\eta m\lambda)^{J/2}\sqrt{T/\lambda} \\
&\quad + \Gamma_3(1+\sqrt{T/\lambda})\Big]\bigg] + C_1\big(Sm^{-1/6}\sqrt{\log m}T^{13/6}\lambda^{-1/6}L^{7/2} + m^{-1/6}\sqrt{\log m}T^{8/3}\lambda^{-2/3}L^3\big) \\
&\leq 3\sqrt{T}\sqrt{\widetilde{d}\log(1+TK/\lambda) + 2}\cdot\bigg[\nu\sqrt{\widetilde{d}\log(1+TK/\lambda) + 2 - 2\log\delta} \\
&\quad + (\lambda + C_3TL)(1-\eta m\lambda)^{J/2}\sqrt{T/\lambda} + 2\sqrt{\lambda}S\bigg] + 1,
\end{aligned}$$

where C_1, C_2, C_3 are positive constants, the first inequality is due to Cauchy-Schwarz inequality, the second inequality due to Lemma 4.6.4, and the third inequality holds for sufficiently large m. This completes our proof. □

4.7 Experiments

In this section, we evaluate NeuralUCB empirically and compare it with seven representative baselines: (1) LinUCB, which is also based on UCB but adopts a linear representation;



(a) $h_1(\mathbf{x}) = 10(\mathbf{x}^\top \mathbf{a})^2$ (b) $h_2(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x}$ (c) $h_3(\mathbf{x}) = \cos(3\mathbf{x}^\top \mathbf{a})$

Figure 4.1: Comparison of NeuralUCB and baseline algorithms on synthetic datasets.

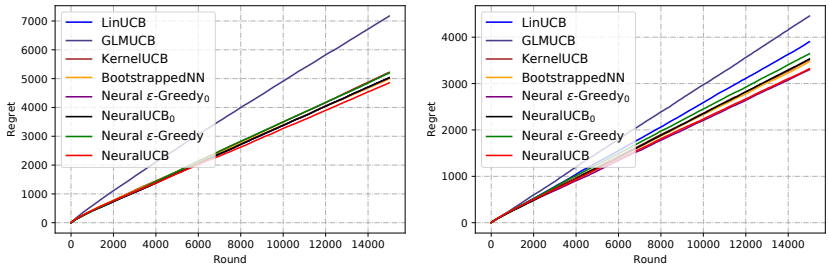
(2) GLMUCB (Filippi et al., 2010), which applies a nonlinear link function over a linear function; (3) KernelUCB (Valko et al., 2013), a kernelised UCB algorithm which makes use of a predefined kernel function; (4) BootstrappedNN (Efron, 1982; Riquelme et al., 2018), which simultaneously trains a set of neural networks using bootstrapped samples and at every round chooses an action based on the prediction of a randomly picked model; (5) Neural ϵ-Greedy, which replaces the UCB-based exploration in Algorithm 5 by ϵ-greedy; (6) NeuralUCB$_0$, as described in Section 4.4; and (7) Neural ϵ-Greedy$_0$, same as NeuralUCB$_0$ but with ϵ-greedy exploration. We use the cumulative regret as the performance metric.

4.7.1 Synthetic Datasets

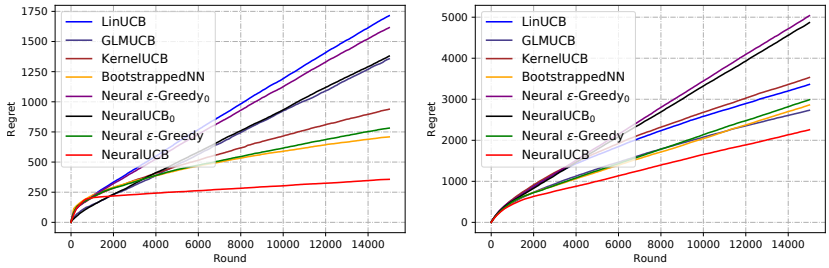
In the first set of experiments, we use contextual bandits with context dimension $d = 20$ and $K = 4$ actions. The number of rounds $T = 10\ 000$. The contextual vectors $\{\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{T,K}\}$ are chosen uniformly at random from the unit ball. The reward function h is one of the following:

$$h_1(\mathbf{x}) = 10(\mathbf{x}^\top \mathbf{a})^2, h_2(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x}, h_3(\mathbf{x}) = \cos(3\mathbf{x}^\top \mathbf{a}),$$

where each entry of $\mathbf{A} \in \mathbb{R}^{d \times d}$ is randomly generated from $N(0, 1)$, $\mathbf{a}$ is randomly generated from uniform distribution over unit ball. For each $h_i(\cdot)$, the reward is generated by $r_{t,a} = h_i(\mathbf{x}_{t,a}) + \xi_t$, where $\xi_t \sim N(0, 1)$.



(a) `covertype` (b) `magic`



(c) `statlog` (d) `mnist`

Figure 4.2: Comparison of NeuralUCB and baseline algorithms on real-world datasets.

Following Li et al. (2010), we implement LinUCB using a constant α (for the variance term in the UCB). We do a grid search for α over $\{0.01, 0.1, 1, 10\}$. For GLMUCB, we use the sigmoid function as the link function and adapt the online Newton step method to accelerate the computation (Zhang et al., 2016; Jun et al., 2017). We do grid searches over $\{0.1, 1, 10\}$ for regularization parameter, $\{1, 10, 100\}$ for step size, $\{0.01, 0.1, 1\}$ for exploration parameter. For KernelUCB, we use the radial basis function (RBF) kernel with parameter σ, and set the regularization parameter to 1. Grid searches over $\{0.1, 1, 10\}$ for σ and $\{0.01, 0.1, 1, 10\}$ for the exploration parameter are done. To accelerate the calculation, we stop adding contexts to KernelUCB after 1000 rounds, following the same setting for Gaussian Process in Riquelme et al. (2018). For all five neural algorithms, we choose a two-layer neural network $f(\mathbf{x}; \boldsymbol{\theta}) = \sqrt{m}\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x})$ with network width $m = 20$, where $\boldsymbol{\theta} = [\text{vec}(\mathbf{W}_1)^\top, \text{vec}(\mathbf{W}_2)^\top] \in \mathbb{R}^p$ and $p = md + m = 420$.[1] Moreover, we set $\gamma_t = \gamma$

[1]Note that the bound on the required network width m is likely not tight. Therefore, in experiments we choose m to be relatively large, but not as large as theory suggests.

Table 4.1: Dataset statistics

Dataset	Covertype	Magic	Statlog	mnist
feature dimension	54	10	8	784
number of classes	7	2	7	10
number of instances	581012	19020	58000	60000

in NeuralUCB, and do a grid search over $\{0.01, 0.1, 1, 10\}$. For NeuralUCB$_0$, we do grid searches for ν over $\{0.1, 1, 10\}$, for λ over $\{0.1, 1, 10\}$, for δ over $\{0.01, 0.1, 1\}$, for S over $\{0.01, 0.1, 1, 10\}$. For Neural ϵ-Greedy and Neural ϵ-Greedy$_0$, we do a grid search for ϵ over $\{0.001, 0.01, 0.1, 0.2\}$. For BootstrappedNN, we follow Riquelme et al. (2018) to set the number of models to be 10 and the transition probability to be 0.8. To accelerate the training process, for BootstrappedNN, NeuralUCB and Neural ϵ-Greedy, we update the parameter $\boldsymbol{\theta}_t$ by TrainNN every 50 rounds. We use stochastic gradient descent with batch size 50, $J = t$ at round t, and do a grid search for step size η over $\{0.001, 0.01, 0.1\}$. For all grid-searched parameters, we choose the best of them for the comparison. All experiments are repeated 10 times, and the averaged results reported for comparison.

4.7.2 Real-world Datasets

We evaluate our algorithms on real-world datasets from the UCI Machine Learning Repository (Dua and Graff, 2017): `covertype`, `magic`, and `statlog`. We also evaluate our algorithms on `mnist` dataset (LeCun et al., 1998). These are all K-class classification datasets (Table 4.1), and are converted into K-armed contextual bandits (Beygelzimer and Langford, 2009). The number of rounds is set as $T = 15000$. Following Riquelme et al. (2018), we

create contextual bandit problems based on the prediction accuracy. In detail, to transform a classification problem with k-classes into a bandit problem, we adapts the disjoint model (Li et al., 2010) which transforms each contextual vector $\mathbf{x} \in \mathbb{R}^d$ into k vectors $\mathbf{x}^{(1)} = (\mathbf{x}, \mathbf{0}, \ldots, \mathbf{0}), \ldots, \mathbf{x}^{(k)} = (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{x}) \in \mathbb{R}^{dk}$. The agent received regret 0 if he classifies the context correctly, and 1 otherwise. For all the algorithms, We reshuffle the order of contexts and repeat the experiment for 10 runs. Averaged results are reported for comparison.

For LinUCB, GLMUCB and KernelUCB, we tune their parameters as Section 4.7.1 suggests. For BootstrappedNN, NeuralUCB, NeuralUCB$_0$, Neural ϵ-Greedy and Neural ϵ-Greedy$_0$, we choose a two-layer neural network with width $m = 100$. For NeuralUCB and NeuralUCB$_0$, since it is computationally expensive to store and compute a whole matrix $\mathbf{Z}_t$, we use a diagonal matrix which consists of the diagonal elements of $\mathbf{Z}_t$ to approximate $\mathbf{Z}_t$. To accelerate the training process, for BootstrappedNN, NeuralUCB and Neural ϵ-Greedy, we update the parameter $\boldsymbol{\theta}_t$ by TrainNN every 100 rounds starting from round 2000. We do grid searches for λ over $\{10^{-i}\}, i = 1, 2, 3, 4$, for η over $\{2 \times 10^{-i}, 5 \times 10^{-i}\}, i = 1, 2, 3, 4$. We set $J = 1000$ and use stochastic gradient descent with batch size 500 to train the networks. For the rest of parameters, we tune them as those in Section 4.7.1 and choose the best of them for comparison.

4.7.3 Results

Figures 4.1 and 4.2 show the cumulative regret of all algorithms. First, due to the nonlinearity of reward functions h, LinUCB fails to learn them for nearly all tasks. GLMUCB is only able to learn the true reward functions for certain tasks due to its simple link function. In contrast, thanks to the neural network representation and efficient exploration, NeuralUCB achieves a substantially lower regret. The performance of Neural ϵ-Greedy is between the two. This suggests that while Neural ϵ-Greedy can capture the nonlinearity of the underlying reward function, ϵ-Greedy based exploration is not as effective as UCB based exploration. This confirms the effectiveness of NeuralUCB for contextual bandit problems with nonlinear

reward functions. Second, it is worth noting that NeuralUCB and Neural ϵ-Greedy outperform NeuralUCB$_0$ and Neural ϵ-Greedy$_0$. This suggests that using deep neural networks to predict the reward function is better than using a fixed feature mapping associated with the Neural Tangent Kernel, which mirrors similar findings in supervised learning (Allen-Zhu and Li, 2019). Furthermore, we can see that KernelUCB is not as good as NeuralUCB, which suggests the limitation of simple kernels like RBF compared to flexible neural networks. What's more, BootstrappedNN can be competitive, approaching the performance of NeuralUCB in some datasets. However, it requires to maintain and train multiple neural networks, so is computationally more expensive than our approach, especially in large-scale problems.

4.8 Conclusion

In this chapter, we proposed NeuralUCB, a new algorithm for stochastic contextual bandits based on neural networks and upper confidence bounds. Building on recent advances in optimization and generalization of deep neural networks, we showed that for an arbitrary bounded reward function, our algorithm achieves an $\widetilde{O}(\widetilde{d}\sqrt{T})$ regret bound. Promising empirical results on both synthetic and real-world data corroborated our theoretical findings, and suggested the potential of the algorithm in practice.

We conclude this chapter with a suggested direction for future research. Given the focus on UCB exploration in this work, a natural open question is provably efficient exploration based on randomized strategies, when DNNs are used. These methods are effective in practice, but existing regret analyses are mostly for shallow (i.e., linear or generalized linear) models (Chapelle and Li, 2011; Agrawal and Goyal, 2013; Russo et al., 2018; Kveton et al., 2020). Extending them to DNNs will be interesting.

4.9 Proof of Additional Results in Section 4.5

4.9.1 Verification of Remark 4.5.4

Suppose there exists a mapping $\boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^{\widehat{d}}$ satisfying $\|\boldsymbol{\psi}(\mathbf{x})\|_2 \leq 1$ which maps any context $\mathbf{x} \in \mathbb{R}^d$ to the Hilbert space $\mathcal{H}$ associated with the Gram matrix $\mathbf{H} \in \mathbb{R}^{TK \times TK}$ over contexts $\{\mathbf{x}^i\}_{i=1}^{TK}$. Then $\mathbf{H} = \boldsymbol{\Psi}^\top \boldsymbol{\Psi}$, where $\boldsymbol{\Psi} = [\boldsymbol{\psi}(\mathbf{x}^1), \ldots, \boldsymbol{\psi}(\mathbf{x}^{TK})] \in \mathbb{R}^{\widehat{d} \times TK}$. Thus, we can bound the effective dimension $\widetilde{d}$ as follows

$$\widetilde{d} = \frac{\log\det[\mathbf{I} + \mathbf{H}/\lambda]}{\log(1 + TK/\lambda)} = \frac{\log\det\left[\mathbf{I} + \boldsymbol{\Psi}\boldsymbol{\Psi}^\top/\lambda\right]}{\log(1 + TK/\lambda)} \leq \widehat{d} \cdot \frac{\log\left\|\mathbf{I} + \boldsymbol{\Psi}\boldsymbol{\Psi}^\top/\lambda\right\|_2}{\log(1 + TK/\lambda)}\,.$$

where the second equality holds due to the fact that $\det(\mathbf{I} + \mathbf{A}^\top\mathbf{A}/\lambda) = \det(\mathbf{I} + \mathbf{A}\mathbf{A}^\top/\lambda)$ holds for any matrix $\mathbf{A}$, and the inequality holds since $\det\mathbf{A} \leq \|\mathbf{A}\|_2^{\widehat{d}}$ for any $\mathbf{A} \in \mathbb{R}^{\widehat{d} \times \widehat{d}}$. Clearly, $\widetilde{d} \leq \widehat{d}$ as long as $\left\|\mathbf{I} + \boldsymbol{\Psi}\boldsymbol{\Psi}^\top/\lambda\right\|_2 \leq 1 + TK/\lambda$. Indeed,

$$\left\|\mathbf{I} + \boldsymbol{\Psi}\boldsymbol{\Psi}^\top/\lambda\right\|_2 \leq 1 + \left\|\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\right\|_2/\lambda \leq 1 + \sum_{i=1}^{TK}\left\|\boldsymbol{\psi}(\mathbf{x}^i)\boldsymbol{\psi}(\mathbf{x}^i)^\top\right\|_2/\lambda \leq 1 + TK/\lambda\,,$$

where the first inequality is due to triangle inequality and the fact $\lambda \geq 1$, the second inequality holds due to the definition of $\boldsymbol{\Psi}$ and triangle inequality, and the last inequality is by $\|\boldsymbol{\psi}(\mathbf{x}^i)\|_2 \leq 1$ for any $1 \leq i \leq TK$.

4.9.2 Verification of Remark 4.5.8

Let $K(\cdot,\cdot)$ be the NTK kernel, then for $i, j \in [TK]$, we have $\mathbf{H}_{i,j} = K(\mathbf{x}^i, \mathbf{x}^j)$. Suppose that $h \in \mathcal{H}$, then h can be decomposed as $h = h_{\mathbf{H}} + h_\perp$, where $h_{\mathbf{H}}(\mathbf{x}) = \sum_{i=1}^{TK}\alpha_i K(\mathbf{x}, \mathbf{x}^i)$ is the projection of h to the function space spanned by $\{K(\mathbf{x}, \mathbf{x}^i)\}_{i=1}^{TK}$ and $h_\perp$ is the orthogonal part. By definition we have $h(\mathbf{x}^i) = h_{\mathbf{H}}(\mathbf{x}^i)$ for $i \in [TK]$, thus

$$\begin{aligned}
\mathbf{h} &= [h(\mathbf{x}^1), \ldots, h(\mathbf{x}^{TK})]^\top \\
&= [h_{\mathbf{H}}(\mathbf{x}^1), \ldots, h_{\mathbf{H}}(\mathbf{x}^{TK})]^\top \\
&= \left[\sum_{i=1}^{TK}\alpha_i K(\mathbf{x}^1, \mathbf{x}^i), \ldots, \sum_{i=1}^{TK}\alpha_i K(\mathbf{x}^{TK}, \mathbf{x}^i)\right]^\top
\end{aligned}$$

$$= \mathbf{H}\boldsymbol{\alpha},$$

which implies that $\boldsymbol{\alpha} = \mathbf{H}^{-1}\mathbf{h}$. Thus, we have

$$\|h\|_{\mathcal{H}} \geq \|h_{\mathbf{H}}\|_{\mathcal{H}} = \sqrt{\boldsymbol{\alpha}^\top \mathbf{H}\boldsymbol{\alpha}} = \sqrt{\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{H}\mathbf{H}^{-1}\mathbf{h}} = \sqrt{\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{h}}.$$

4.9.3 Proof of Corollary 4.5.9

Proof of Corollary 4.5.9. Notice that $R_T \leq T$ since $0 \leq h(\mathbf{x}) \leq 1$. Thus, with the fact that with probability at least $1-\delta$, (4.5.2) holds, we can bound $\mathbb{E}[R_T]$ as

$$\begin{aligned}\mathbb{E}[R_T] \leq (1-\delta)\Bigg(3\sqrt{T}\sqrt{\widetilde{d}\log(1+TK/\lambda)+2}\Bigg[\nu\sqrt{\widetilde{d}\log(1+TK/\lambda)+2-2\log\delta} \\ + 2\sqrt{\lambda}S + (\lambda + C_2TL)(1-\eta m\lambda)^{J/2}\sqrt{T/\lambda}\Bigg] + 1\Bigg) + \delta T. \end{aligned} \tag{4.9.1}$$

Taking $\delta = 1/T$ completes the proof. □

4.10 Proof of Lemmas in Section 4.6

4.10.1 Proof of Lemma 4.6.1

We start with the following lemma:

Lemma 4.10.1. Let $\mathbf{G} = [\mathbf{g}(\mathbf{x}^1;\boldsymbol{\theta}_0),\ldots,\mathbf{g}(\mathbf{x}^{TK};\boldsymbol{\theta}_0)]/\sqrt{m} \in \mathbb{R}^{p\times(TK)}$. Let $\mathbf{H}$ be the NTK matrix as defined in Definition 4.5.1. For any $\delta \in (0,1)$, if

$$m = \Omega\bigg(\frac{L^6\log(TKL/\delta)}{\epsilon^4}\bigg),$$

then with probability at least $1-\delta$, we have

$$\|\mathbf{G}^\top\mathbf{G} - \mathbf{H}\|_F \leq TK\epsilon.$$

We begin to prove Lemma 4.6.1.

Proof of Lemma 4.6.1. By Assumption 4.5.2, we know that $\lambda_0 > 0$. By the choice of m, we have $m \geq \Omega(L^6 \log(TKL/\delta)/\epsilon^4)$, where $\epsilon = \lambda_0/(2TK)$. Thus, due to Lemma 4.10.1, with probability at least $1-\delta$, we have $\|\mathbf{G}^\top\mathbf{G} - \mathbf{H}\|_F \leq TK\epsilon = \lambda_0/2$. That leads to

$$\mathbf{G}^\top\mathbf{G} \succeq \mathbf{H} - \|\mathbf{G}^\top\mathbf{G} - \mathbf{H}\|_F\mathbf{I} \succeq \mathbf{H} - \lambda_0\mathbf{I}/2 \succeq \mathbf{H}/2 \succ 0, \tag{4.10.1}$$

where the first inequality holds due to the triangle inequality, the third and fourth inequality holds due to $\mathbf{H} \succeq \lambda_0\mathbf{I} \succ 0$. Thus, suppose the singular value decomposition of $\mathbf{G}$ is $\mathbf{G} = \mathbf{PAQ}^\top$, $\mathbf{P} \in \mathbb{R}^{p\times TK}, \mathbf{A} \in \mathbb{R}^{TK\times TK}, \mathbf{Q} \in \mathbb{R}^{TK\times TK}$, we have $\mathbf{A} \succ 0$. Now we are going to show that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0 + \mathbf{PA}^{-1}\mathbf{Q}^\top\mathbf{h}/\sqrt{m}$ satisfies (4.6.1). First, we have

$$\mathbf{G}^\top\sqrt{m}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0) = \mathbf{QAP}^\top\mathbf{PA}^{-1}\mathbf{Q}^\top\mathbf{h} = \mathbf{h},$$

which suggests that for any i, $\langle \mathbf{g}(\mathbf{x}^i;\boldsymbol{\theta}_0), \boldsymbol{\theta}^* - \boldsymbol{\theta}_0\rangle = h(\mathbf{x}^i)$. We also have

$$m\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_2^2 = \mathbf{h}^\top\mathbf{QA}^{-2}\mathbf{Q}^\top\mathbf{h} = \mathbf{h}^\top(\mathbf{G}^\top\mathbf{G})^{-1}\mathbf{h} \leq 2\mathbf{h}^\top\mathbf{H}^{-1}\mathbf{h},$$

where the last inequality holds due to (4.10.1). This completes the proof. □

4.10.2 Proof of Lemma 4.6.2

In this section we prove Lemma 4.6.2. For simplicity, we define $\bar{\mathbf{Z}}_t, \bar{\mathbf{b}}_t, \bar{\gamma}_t$ as follows:

$$\begin{aligned}
\bar{\mathbf{Z}}_t &= \lambda\mathbf{I} + \sum_{i=1}^{t}\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_0)\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_0)^\top/m,\\
\bar{\mathbf{b}}_t &= \sum_{i=1}^{t} r_{i,a_i}\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_0)/\sqrt{m},\\
\bar{\gamma}_t &= \nu\sqrt{\log\frac{\det\bar{\mathbf{Z}}_t}{\det\lambda\mathbf{I}} - 2\log\delta} + \sqrt{\lambda}S.
\end{aligned}$$

We need the following lemmas. The first lemma shows that the network parameter $\boldsymbol{\theta}_t$ at round t can be well approximated by $\boldsymbol{\theta}_0 + \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t/\sqrt{m}$.

Lemma 4.10.2. There exist constants $\{\bar{C}_i\}_{i=1}^5 > 0$ such that for any $\delta > 0$, if for all $t \in [T]$, η, m satisfy

$$
\begin{aligned}
&2\sqrt{t/(m\lambda)} \geq \bar{C}_1 m^{-3/2} L^{-3/2} [\log(TKL^2/\delta)]^{3/2}, \\
&2\sqrt{t/(m\lambda)} \leq \bar{C}_2 \min\big\{L^{-6}[\log m]^{-3/2}, \big(m(\lambda\eta)^2 L^{-6} t^{-1} (\log m)^{-1}\big)^{3/8}\big\}, \\
&\eta \leq \bar{C}_3 (m\lambda + tmL)^{-1}, \\
&m^{1/6} \geq \bar{C}_4 \sqrt{\log m} L^{7/2} t^{7/6} \lambda^{-7/6} (1 + \sqrt{t/\lambda}),
\end{aligned}
$$

then with probability at least $1 - \delta$, we have that $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 \leq 2\sqrt{t/(m\lambda)}$ and

$$
\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1} \bar{\mathbf{b}}_t / \sqrt{m}\|_2 \leq (1 - \eta m\lambda)^{J/2} \sqrt{t/(m\lambda)} + \bar{C}_5 m^{-2/3} \sqrt{\log m} L^{7/2} t^{5/3} \lambda^{-5/3} (1 + \sqrt{t/\lambda}).
$$

Next lemma shows the error bounds for $\bar{\mathbf{Z}}_t$ and $\mathbf{Z}_t$.

Lemma 4.10.3. There exist constants $\{\bar{C}_i\}_{i=1}^5 > 0$ such that for any $\delta > 0$, if m satisfies that

$$
\bar{C}_1 m^{-3/2} L^{-3/2} [\log(TKL^2/\delta)]^{3/2} \leq 2\sqrt{t/(m\lambda)} \leq \bar{C}_2 L^{-6} [\log m]^{-3/2}, \ \forall t \in [T],
$$

then with probability at least $1 - \delta$, for any $t \in [T]$, we have

$$
\begin{aligned}
&\|\mathbf{Z}_t\|_2 \leq \lambda + \bar{C}_3 tL, \\
&\|\bar{\mathbf{Z}}_t - \mathbf{Z}_t\|_F \leq \bar{C}_4 m^{-1/6} \sqrt{\log m} L^4 t^{7/6} \lambda^{-1/6}, \\
&\bigg| \log \frac{\det(\bar{\mathbf{Z}}_t)}{\det(\lambda \mathbf{I})} - \log \frac{\det(\mathbf{Z}_t)}{\det(\lambda \mathbf{I})} \bigg| \leq \bar{C}_5 m^{-1/6} \sqrt{\log m} L^4 t^{5/3} \lambda^{-1/6}.
\end{aligned}
$$

With above lemmas, we prove Lemma 4.6.2 as follows.

Proof of Lemma 4.6.2. By Lemma 4.10.2 we know that $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 \leq 2\sqrt{t/(m\lambda)}$. By Lemma 4.6.1, with probability at least $1 - \delta$, there exists $\boldsymbol{\theta}^*$ such that for any $1 \leq t \leq T$,

$$
h(\mathbf{x}_{t,a_t}) = \langle \mathbf{g}(\mathbf{x}_{t,a_t}; \boldsymbol{\theta}_0)/\sqrt{m}, \ \sqrt{m}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0) \rangle, \tag{4.10.2}
$$

$$
\sqrt{m}\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_2 \leq \sqrt{2\mathbf{h}^\top \mathbf{H}^{-1} \mathbf{h}} \leq S, \tag{4.10.3}
$$

where the second inequality holds since $S \geq \sqrt{2\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{h}}$ in the statement of Lemma 4.6.2. Thus, conditioned on (4.10.2) and (4.10.3), by Theorem 2 in Abbasi-Yadkori et al. (2011), with probability at least $1-\delta$, for any $1 \leq t \leq T$, $\boldsymbol{\theta}^*$ satisfies that

$$\|\sqrt{m}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0) - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t\|_{\bar{\mathbf{Z}}_t} \leq \bar{\gamma}_t. \tag{4.10.4}$$

We now prove that $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|_{\mathbf{Z}_t} \leq \gamma_t/\sqrt{m}$. From the triangle inequality,

$$\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|_{\mathbf{Z}_t} \leq \underbrace{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t/\sqrt{m}\|_{\mathbf{Z}_t}}_{I_1} + \underbrace{\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t/\sqrt{m}\|_{\mathbf{Z}_t}}_{I_2}. \tag{4.10.5}$$

We bound I_1 and I_2 separately. For I_1, we have

$$\begin{aligned}
I_1^2 &= (\boldsymbol{\theta}^* - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t/\sqrt{m})^\top \mathbf{Z}_t(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t/\sqrt{m}) \\
&= (\boldsymbol{\theta}^* - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t/\sqrt{m})^\top \bar{\mathbf{Z}}_t(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t/\sqrt{m}) \\
&\quad + (\boldsymbol{\theta}^* - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t/\sqrt{m})^\top (\mathbf{Z}_t - \bar{\mathbf{Z}}_t)(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t/\sqrt{m}) \\
&\leq (\boldsymbol{\theta}^* - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t/\sqrt{m})^\top \bar{\mathbf{Z}}_t(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t/\sqrt{m}) \\
&\quad + \frac{\|\mathbf{Z}_t - \bar{\mathbf{Z}}_t\|_2}{\lambda}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t/\sqrt{m})^\top \bar{\mathbf{Z}}_t(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t/\sqrt{m}) \\
&\leq (1 + \|\mathbf{Z}_t - \bar{\mathbf{Z}}_t\|_2/\lambda)\bar{\gamma}_t^2/m,
\end{aligned} \tag{4.10.6}$$

where the first inequality holds due to the fact that $\mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \mathbf{x}^\top \mathbf{B}\mathbf{x} \cdot \|\mathbf{A}\|_2/\lambda_{\min}(\mathbf{B})$ for some $\mathbf{B} \succ 0$ and the fact that $\lambda_{\min}(\bar{\mathbf{Z}}_t) \geq \lambda$, the second inequality holds due to (4.10.4). We have

$$\big\|\bar{\mathbf{Z}}_t - \mathbf{Z}_t\big\|_2 \leq \big\|\bar{\mathbf{Z}}_t - \mathbf{Z}_t\big\|_F \leq C_1 m^{-1/6}\sqrt{\log m} L^4 t^{7/6}\lambda^{-1/6}, \tag{4.10.7}$$

where the first inequality holds due to the fact that $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F$, the second inequality holds due to Lemma 4.10.3. We also have

$$\begin{aligned}
\bar{\gamma}_t &= \nu\sqrt{\log\frac{\det \bar{\mathbf{Z}}_t}{\det \lambda\mathbf{I}} - 2\log\delta} + \sqrt{\lambda}S \\
&= \nu\sqrt{\log\frac{\det \mathbf{Z}_t}{\det \lambda\mathbf{I}} + \log\frac{\det \bar{\mathbf{Z}}_t}{\det \lambda\mathbf{I}} - \log\frac{\det \mathbf{Z}_t}{\det \lambda\mathbf{I}} - 2\log\delta} + \sqrt{\lambda}S \\
&\leq \nu\sqrt{\log\frac{\det \mathbf{Z}_t}{\det \lambda\mathbf{I}} + C_2 m^{-1/6}\sqrt{\log m} L^4 t^{5/3}\lambda^{-1/6} - 2\log\delta} + \sqrt{\lambda}S,
\end{aligned} \tag{4.10.8}$$

where $C_1, C_2 > 0$ are two constants, the inequality holds due to Lemma 4.10.3. Substituting (4.10.7) and (4.10.8) into (4.10.6), we have

$$\begin{aligned} I_1 &\leq \sqrt{1 + \|\mathbf{Z}_t - \bar{\mathbf{Z}}_t\|_2/\lambda}\bar{\gamma}_t/\sqrt{m} \\ &\leq \sqrt{1 + C_1 m^{-1/6}\sqrt{\log m} L^4 t^{7/6}\lambda^{-7/6}}/\sqrt{m} \\ &\quad \cdot \left(\nu\sqrt{\log\frac{\det \mathbf{Z}_t}{\det \lambda\mathbf{I}} + C_2 m^{-1/6}\sqrt{\log m}L^4 t^{5/3}\lambda^{-1/6} - 2\log\delta} + \sqrt{\lambda}S\right). \end{aligned} \tag{4.10.9}$$

For I_2, we have

$$\begin{aligned} I_2 &= \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t/\sqrt{m}\|_{\mathbf{Z}_t} \\ &\leq \|\mathbf{Z}_t\|_2 \cdot \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t/\sqrt{m}\|_2 \\ &\leq (\lambda + C_3 tL)\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0 - \bar{\mathbf{Z}}_t^{-1}\bar{\mathbf{b}}_t/\sqrt{m}\|_2 \\ &\leq (\lambda + C_3 tL)\Big[(1-\eta m\lambda)^{J/2}\sqrt{t/(m\lambda)} + m^{-2/3}\sqrt{\log m}L^{7/2}t^{5/3}\lambda^{-5/3}(1+\sqrt{t/\lambda})\Big], \end{aligned} \tag{4.10.10}$$

where $C_3 > 0$ is a constant, the first inequality holds since for any vector $\mathbf{a}$, the second inequality holds due to $\|\mathbf{Z}_t\|_2 \leq \lambda + C_3 tL$ by Lemma 4.10.3, the third inequality holds due to Lemma 4.10.2. Substituting (4.10.9) and (4.10.10) into (4.10.5), we obtain $\big\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\big\|_{\mathbf{Z}_t} \leq \gamma_t/\sqrt{m}$. This completes the proof. □

4.10.3 Proof of Lemma 4.6.3

The proof starts with three lemmas that bound the error terms of the function value and gradient of neural networks.

Lemma 4.10.4 (Lemma 4.1, Cao and Gu (2019))**.** There exist constants $\{\bar{C}_i\}_{i=1}^3 > 0$ such that for any $\delta > 0$, if τ satisfies that

$$\bar{C}_1 m^{-3/2}L^{-3/2}[\log(TKL^2/\delta)]^{3/2} \leq \tau \leq \bar{C}_2 L^{-6}[\log m]^{-3/2},$$

then with probability at least $1-\delta$, for all $\widetilde{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}$ satisfying $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 \leq \tau, \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 \leq \tau$ and

$j \in [TK]$ we have

$$\left|f(\mathbf{x}^j;\widetilde{\boldsymbol{\theta}}) - f(\mathbf{x}^j;\widehat{\boldsymbol{\theta}}) - \langle \mathbf{g}(\mathbf{x}^j;\widehat{\boldsymbol{\theta}}), \widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\rangle\right| \leq \bar{C}_3\tau^{4/3}L^3\sqrt{m\log m}.$$

Lemma 4.10.5 (Theorem 5, Allen-Zhu et al. (2019))**.** There exist constants $\{\bar{C}_i\}_{i=1}^3 > 0$ such that for any $\delta \in (0,1)$, if τ satisfies that

$$\bar{C}_1 m^{-3/2}L^{-3/2}\max\{\log^{-3/2} m, \log^{3/2}(TK/\delta)\} \leq \tau \leq \bar{C}_2 L^{-9/2}\log^{-3} m,$$

then with probability at least $1-\delta$, for all $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leq \tau$ and $j \in [TK]$ we have

$$\|\mathbf{g}(\mathbf{x}^j;\boldsymbol{\theta}) - \mathbf{g}(\mathbf{x}^j;\boldsymbol{\theta}_0)\|_2 \leq \bar{C}_3\sqrt{\log m}\tau^{1/3}L^3\|\mathbf{g}(\mathbf{x}^j;\boldsymbol{\theta}_0)\|_2.$$

Lemma 4.10.6 (Lemma B.3, Cao and Gu (2019))**.** There exist constants $\{\bar{C}_i\}_{i=1}^3 > 0$ such that for any $\delta > 0$, if τ satisfies that

$$\bar{C}_1 m^{-3/2}L^{-3/2}[\log(TKL^2/\delta)]^{3/2} \leq \tau \leq \bar{C}_2 L^{-6}[\log m]^{-3/2},$$

then with probability at least $1-\delta$, for any $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2 \leq \tau$ and $j \in [TK]$ we have $\|\mathbf{g}(\mathbf{x}^j;\boldsymbol{\theta})\|_F \leq \bar{C}_3\sqrt{mL}$.

Proof of Lemma 4.6.3. We follow the regret bound analysis in Abbasi-Yadkori et al. (2011); Valko et al. (2013). Denote $a_t^* = \mathrm{argmax}_{a\in[K]} h(\mathbf{x}_{t,a})$ and $\mathcal{C}_t = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|_{\mathbf{Z}_t} \leq \gamma_t/\sqrt{m}\}$. By Lemma 4.6.2, for all $1 \leq t \leq T$, we have $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 \leq 2\sqrt{t/(m\lambda)}$ and $\boldsymbol{\theta}^* \in \mathcal{C}_t$. By the choice of m, Lemmas 4.10.4, 4.10.5 and 4.10.6 hold.

We denote $\mathcal{D} := [\mathbf{x}^1, \ldots, \mathbf{x}^{TK}]$. Then by Lemma 4.6.1, we have for all $\mathbf{x} \in \mathcal{D}$,

$$h(\mathbf{x}) = \langle \mathbf{g}(\mathbf{x};\boldsymbol{\theta}_0), \boldsymbol{\theta}^* - \boldsymbol{\theta}_0\rangle. \tag{4.10.11}$$

By Lemma 4.10.4, we also have for all $\mathbf{x} \in \mathcal{D}$,

$$|f(\mathbf{x};\boldsymbol{\theta}_{t-1}) - f(\mathbf{x};\boldsymbol{\theta}_0) - \langle \mathbf{g}(\mathbf{x};\boldsymbol{\theta}_0), \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}_0\rangle| \leq C_1 m^{-1/6}\sqrt{\log m}t^{2/3}\lambda^{-2/3}L^3. \tag{4.10.12}$$

Then for all $\mathbf{x} \in \mathcal{D}$, we have

$$
\begin{aligned}
&|f(\mathbf{x};\boldsymbol{\theta}_{t-1}) - h(\mathbf{x})| \\
&\leq |\langle \mathbf{g}(\mathbf{x};\boldsymbol{\theta}_0), \boldsymbol{\theta}^* - \boldsymbol{\theta}_{t-1}\rangle| + C_1 m^{-1/6}\sqrt{\log mt}\, t^{2/3}\lambda^{-2/3}L^3 \\
&\leq |\langle \mathbf{g}(\mathbf{x};\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}^* - \boldsymbol{\theta}_{t-1}\rangle| + |\langle \mathbf{g}(\mathbf{x};\boldsymbol{\theta}_{t-1}) - \mathbf{g}(\mathbf{x};\boldsymbol{\theta}_0), \boldsymbol{\theta}^* - \boldsymbol{\theta}_{t-1}\rangle| + C_1 m^{-1/6}\sqrt{\log mt}\, t^{2/3}\lambda^{-2/3}L^3 \\
&\leq \big\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_{t-1}\big\|_{\mathbf{Z}_{t-1}} \|\mathbf{g}(\mathbf{x};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}} + C_1 m^{-1/6}\sqrt{\log mt}\, t^{2/3}\lambda^{-2/3}L^3 \\
&\quad + C_2\sqrt{\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{h}}\, m^{-1/6}\sqrt{\log mt}\, t^{1/6}\lambda^{-1/6}L^{7/2} \\
&\leq \gamma_{t-1}\|\mathbf{g}(\mathbf{x};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}} + C_1 m^{-1/6}\sqrt{\log mt}\, t^{2/3}\lambda^{-2/3}L^3 \\
&\quad + C_2\sqrt{\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{h}}\, m^{-1/6}\sqrt{\log mt}\, t^{1/6}\lambda^{-1/6}L^{7/2}, \qquad (4.10.13)
\end{aligned}
$$

where the first inequality holds due to (4.10.11), (4.10.12) and the fact that $f(\mathbf{x};\boldsymbol{\theta}_0)$ due to the initialization scheme of $\boldsymbol{\theta}_0$, the second one holds due to the triangle inequality, the third one holds Lemmas 4.6.1, 4.10.5, 4.10.6, the fourth one holds due to Lemma 4.6.2 and the fact $\boldsymbol{\theta}^* \in \mathcal{C}_{t-1}$.

Then we have proved (4.6.2) by (4.10.13). To prove (4.6.3), we have

$$
\begin{aligned}
&h(\mathbf{x}_{t,a_t^*}) - h(\mathbf{x}_{t,a_t}) \\
&\leq f(\mathbf{x}_{t,a_t^*};\boldsymbol{\theta}_{t-1}) + \gamma_{t-1}\|\mathbf{g}(\mathbf{x}_{t,a_t^*};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}} - h(\mathbf{x}_{t,a_t}) \\
&\quad + C_1 m^{-1/6}\sqrt{\log mt}\, t^{2/3}\lambda^{-2/3}L^3 + C_2\sqrt{\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{h}}\, m^{-1/6}\sqrt{\log mt}\, t^{1/6}\lambda^{-1/6}L^{7/2} \\
&\leq f(\mathbf{x}_{t,a_t};\boldsymbol{\theta}_{t-1}) + \gamma_{t-1}\|\mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}} - h(\mathbf{x}_{t,a_t}) \\
&\quad + C_1 m^{-1/6}\sqrt{\log mt}\, t^{2/3}\lambda^{-2/3}L^3 + C_2\sqrt{\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{h}}\, m^{-1/6}\sqrt{\log mt}\, t^{1/6}\lambda^{-1/6}L^{7/2} \\
&\leq 2(\gamma_{t-1}\|\mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}} \\
&\quad + C_1 m^{-1/6}\sqrt{\log mt}\, t^{2/3}\lambda^{-2/3}L^3 + C_2\sqrt{\mathbf{h}^\top \mathbf{H}^{-1}\mathbf{h}}\, m^{-1/6}\sqrt{\log mt}\, t^{1/6}\lambda^{-1/6}L^{7/2}), \qquad (4.10.14)
\end{aligned}
$$

where the first inequality holds due to (4.10.13), the second one holds due to the definition of a_t ($a_t = \operatorname{argmax}_{a\in[K]} f(\mathbf{x}_{t,a};\boldsymbol{\theta}_{t-1}) + \gamma_{t-1}\|\mathbf{g}(\mathbf{x}_{t,a};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}}$), the third one holds due to (4.10.13) again. Finally, from (4.10.14) we have

$$h(\mathbf{x}_{t,a_t^*}) - h(\mathbf{x}_{t,a_t})$$

$$
\begin{aligned}
&\leq \min\left\{2\gamma_{t-1}\|\mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}} + 2C_1 m^{-1/6}\sqrt{\log mt}\, t^{2/3}\lambda^{-2/3}L^3\right. \\
&\quad \left. + 2C_2\sqrt{\mathbf{h}^\top\mathbf{H}^{-1}\mathbf{h}}\, m^{-1/6}\sqrt{\log mt}\, t^{1/6}\lambda^{-1/6}L^{7/2}, 1\right\} \\
&\leq \min\left\{2\gamma_{t-1}\|\mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}}, 1\right\} + 2C_1 m^{-1/6}\sqrt{\log mt}\, t^{2/3}\lambda^{-2/3}L^3 \\
&\quad + 2C_2\sqrt{\mathbf{h}^\top\mathbf{H}^{-1}\mathbf{h}}\, m^{-1/6}\sqrt{\log mt}\, t^{1/6}\lambda^{-1/6}L^{7/2} \\
&\leq 2\gamma_{t-1}\min\left\{\|\mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|_{\mathbf{Z}_{t-1}^{-1}}, 1\right\} + 2C_1 m^{-1/6}\sqrt{\log mt}\, t^{2/3}\lambda^{-2/3}L^3 \\
&\quad + 2C_2\sqrt{\mathbf{h}^\top\mathbf{H}^{-1}\mathbf{h}}\, m^{-1/6}\sqrt{\log mt}\, t^{1/6}\lambda^{-1/6}L^{7/2}, \qquad (4.10.15)
\end{aligned}
$$

where the first inequality holds due to the fact that $0 \leq h(\mathbf{x}_{t,a_t^*}) - h(\mathbf{x}_{t,a_t}) \leq 1$, the second inequality holds due to the fact that $\min\{a+b,1\} \leq \min\{a,1\} + b$, the third inequality holds due to the fact $\gamma_{t-1} \geq \sqrt{\lambda}S \geq 1$. Finally, by the fact that $\sqrt{2\mathbf{h}\mathbf{H}^{-1}\mathbf{h}} \leq S$, the proof completes. □

4.10.4 Proof of Lemma 4.6.4

In this section we prove Lemma 4.6.4, we need the following lemma from Abbasi-Yadkori et al. (2011).

Lemma 4.10.7 (Lemma 11, Abbasi-Yadkori et al. (2011))**.** We have the following inequality:

$$
\sum_{t=1}^{T}\min\left\{\|\mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|^2_{\mathbf{Z}_{t-1}^{-1}}, 1\right\} \leq 2\log\frac{\det\mathbf{Z}_T}{\det\lambda\mathbf{I}}.
$$

Proof of Lemma 4.6.4. First by the definition of γ_t, we know that γ_t is a monotonic function w.r.t. $\det\mathbf{Z}_t$. By the definition of $\mathbf{Z}_t$, we know that $\mathbf{Z}_T \succeq \mathbf{Z}_t$, which implies that $\det\mathbf{Z}_t \leq \det\mathbf{Z}_T$. Thus, $\gamma_t \leq \gamma_T$. Second, by Lemma 4.10.7 we know that

$$
\begin{aligned}
&\sum_{t=1}^{T}\min\left\{\|\mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta}_{t-1})/\sqrt{m}\|^2_{\mathbf{Z}_{t-1}^{-1}}, 1\right\} \\
&\leq 2\log\frac{\det\mathbf{Z}_T}{\det\lambda\mathbf{I}}
\end{aligned}
$$

$$\leq 2\log\frac{\det\bar{\mathbf{Z}}_T}{\det\lambda\mathbf{I}} + C_1 m^{-1/6}\sqrt{\log m}L^4T^{5/3}\lambda^{-1/6}, \tag{4.10.16}$$

where the second inequality holds due to Lemma 4.10.3. Next we are going to bound $\log\det\bar{\mathbf{Z}}_T$. Denote $\mathbf{G} = [\mathbf{g}(\mathbf{x}^1;\boldsymbol{\theta}_0)/\sqrt{m},\ldots,\mathbf{g}(\mathbf{x}^{TK};\boldsymbol{\theta}_0)/\sqrt{m}] \in \mathbb{R}^{p\times(TK)}$, then we have

$$\begin{aligned}
\log\frac{\det\bar{\mathbf{Z}}_T}{\det\lambda\mathbf{I}} &= \log\det\bigg(\mathbf{I} + \sum_{t=1}^{T}\mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta}_0)\mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta}_0)^\top/(m\lambda)\bigg)\\
&\leq \log\det\bigg(\mathbf{I} + \sum_{i=1}^{TK}\mathbf{g}(\mathbf{x}^i;\boldsymbol{\theta}_0)\mathbf{g}(\mathbf{x}^i;\boldsymbol{\theta}_0)^\top/(m\lambda)\bigg)\\
&= \log\det\Big(\mathbf{I} + \mathbf{G}\mathbf{G}^\top/\lambda\Big)\\
&= \log\det\Big(\mathbf{I} + \mathbf{G}^\top\mathbf{G}/\lambda\Big),
\end{aligned} \tag{4.10.17}$$

where the inequality holds naively, the third equality holds since for any matrix $\mathbf{A}\in\mathbb{R}^{p\times TK}$, we have $\det(\mathbf{I}+\mathbf{A}\mathbf{A}^\top) = \det(\mathbf{I}+\mathbf{A}^\top\mathbf{A})$. We can further bound (4.10.17) as follows:

$$\begin{aligned}
\log\det\Big(\mathbf{I} + \mathbf{G}^\top\mathbf{G}/\lambda\Big) &= \log\det\Big(\mathbf{I} + \mathbf{H}/\lambda + (\mathbf{G}^\top\mathbf{G} - \mathbf{H})/\lambda\Big)\\
&\leq \log\det\Big(\mathbf{I} + \mathbf{H}/\lambda\Big) + \langle(\mathbf{I}+\mathbf{H}/\lambda)^{-1}, (\mathbf{G}^\top\mathbf{G}-\mathbf{H})/\lambda\rangle\\
&\leq \log\det\Big(\mathbf{I} + \mathbf{H}/\lambda\Big) + \|(\mathbf{I}+\mathbf{H}/\lambda)^{-1}\|_F\|\mathbf{G}^\top\mathbf{G}-\mathbf{H}\|_F/\lambda\\
&\leq \log\det\Big(\mathbf{I} + \mathbf{H}/\lambda\Big) + \sqrt{TK}\|\mathbf{G}^\top\mathbf{G}-\mathbf{H}\|_F\\
&\leq \log\det\Big(\mathbf{I} + \mathbf{H}/\lambda\Big) + 1\\
&= \widetilde{d}\log(1+TK/\lambda) + 1,
\end{aligned} \tag{4.10.18}$$

where the first inequality holds due to the concavity of $\log\det(\cdot)$, the second inequality holds due to the fact that $\langle\mathbf{A},\mathbf{B}\rangle \leq \|\mathbf{A}\|_F\|\mathbf{B}\|_F$, the third inequality holds due to the facts that $\mathbf{I}+\mathbf{H}/\lambda \succeq \mathbf{I}$, $\lambda\geq 1$ and $\|\mathbf{A}\|_F \leq \sqrt{TK}\|\mathbf{A}\|_2$ for any $\mathbf{A}\in\mathbb{R}^{TK\times TK}$, the fourth inequality holds by Lemma 4.10.1 with the choice of m, the fifth inequality holds by the definition of effective dimension in Definition 4.5.3, and the last inequality holds due to the choice of λ.

Substituting (4.10.18) into (4.10.17), we obtain that

$$\log \frac{\det \bar{\mathbf{Z}}_T}{\det \lambda \mathbf{I}} \leq \widetilde{d} \log(1 + TK/\lambda) + 1. \tag{4.10.19}$$

Substituting (4.10.19) into (4.10.16), we have

$$\sum_{t=1}^{T} \min \left\{ \|\mathbf{g}(\mathbf{x}_{t,a_t}; \boldsymbol{\theta}_{t-1})/\sqrt{m}\|^2_{\mathbf{Z}_{t-1}^{-1}}, 1 \right\} \leq 2\widetilde{d} \log(1 + TK/\lambda) + 2 + C_1 m^{-1/6} \sqrt{\log m} L^4 T^{5/3} \lambda^{-1/6}. \tag{4.10.20}$$

We now bound γ_T, which is

$$\begin{aligned}
\gamma_T &= \sqrt{1 + C_1 m^{-1/6} \sqrt{\log m} L^4 T^{7/6} \lambda^{-7/6}} \\
&\quad \cdot \left(\nu \sqrt{\log \frac{\det \mathbf{Z}_T}{\det \lambda \mathbf{I}} + C_2 m^{-1/6} \sqrt{\log m} L^4 T^{5/3} \lambda^{-1/6} - 2 \log \delta} + \sqrt{\lambda} S \right) \\
&\quad + (\lambda + C_3 TL) \Big[(1 - \eta m \lambda)^{J/2} \sqrt{T/(m\lambda)} + m^{-2/3} \sqrt{\log m} L^{7/2} T^{5/3} \lambda^{-5/3} (1 + \sqrt{T/\lambda}) \Big] \\
&\leq \sqrt{1 + C_1 m^{-1/6} \sqrt{\log m} L^4 T^{7/6} \lambda^{-7/6}} \\
&\quad \cdot \left(\nu \sqrt{\log \frac{\det \bar{\mathbf{Z}}_T}{\det \lambda \mathbf{I}} + 2C_2 m^{-1/6} \sqrt{\log m} L^4 T^{5/3} \lambda^{-1/6} - 2 \log \delta} + \sqrt{\lambda} S \right) \\
&\quad + (\lambda + C_3 TL) \Big[(1 - \eta m \lambda)^{J/2} \sqrt{T/(m\lambda)} + m^{-2/3} \sqrt{\log m} L^{7/2} T^{5/3} \lambda^{-5/3} (1 + \sqrt{T/\lambda}) \Big],
\end{aligned} \tag{4.10.21}$$

where the inequality holds due to Lemma 4.10.3. Finally, we have

$$\begin{aligned}
&\sqrt{\sum_{t=1}^{T} \gamma_{t-1}^2 \min \left\{ \|\mathbf{g}(\mathbf{x}_{t,a_t}; \boldsymbol{\theta}_{t-1})/\sqrt{m}\|^2_{\mathbf{Z}_{t-1}^{-1}}, 1 \right\}} \\
&\leq \gamma_T \sqrt{\sum_{t=1}^{T} \min \left\{ \|\mathbf{g}(\mathbf{x}_{t,a_t}; \boldsymbol{\theta}_{t-1})/\sqrt{m}\|^2_{\mathbf{Z}_{t-1}^{-1}}, 1 \right\}} \\
&\leq \sqrt{\log \frac{\det \bar{\mathbf{Z}}_T}{\det \lambda \mathbf{I}} + C_1 m^{-1/6} \sqrt{\log m} L^4 T^{5/3} \lambda^{-1/6}} \Bigg[\sqrt{1 + C_1 m^{-1/6} \sqrt{\log m} L^4 T^{7/6} \lambda^{-7/6}} \\
&\quad \cdot \left(\nu \sqrt{\log \frac{\det \bar{\mathbf{Z}}_T}{\det \lambda \mathbf{I}} + 2C_2 m^{-1/6} \sqrt{\log m} L^4 T^{5/3} \lambda^{-1/6} - 2 \log \delta} + \sqrt{\lambda} S \right) \\
&\quad + (\lambda + C_3 TL) \Big[(1 - \eta m \lambda)^{J/2} \sqrt{T/(m\lambda)} + m^{-3/2} \sqrt{\log m} L^{7/2} T^{5/3} \lambda^{-5/3} (1 + \sqrt{T/\lambda}) \Big] \Bigg]
\end{aligned}$$

$$\leq \sqrt{\widetilde{d}\log(1+TK/\lambda)+1+C_1 m^{-1/6}\sqrt{\log m}L^4T^{5/3}\lambda^{-1/6}}\bigg[\sqrt{1+C_1 m^{-1/6}\sqrt{\log m}L^4T^{7/6}\lambda^{-7/6}}$$
$$\cdot\bigg(\nu\sqrt{\widetilde{d}\log(1+TK/\lambda)+1+2C_2 m^{-1/6}\sqrt{\log m}L^4T^{5/3}\lambda^{-1/6}-2\log\delta}+\sqrt{\lambda}S\bigg)$$
$$+(\lambda+C_3TL)\Big[(1-\eta m\lambda)^{J/2}\sqrt{T/(m\lambda)}+m^{-3/2}\sqrt{\log m}L^{7/2}T^{5/3}\lambda^{-5/3}(1+\sqrt{T/\lambda})\Big]\bigg],$$

where the first inequality holds due to the fact that $\gamma_{t-1}\leq\gamma_T$, the second inequality holds due to (4.10.20) and (4.10.21), the third inequality holds due to (4.10.19). This completes our proof. □

4.11 Proofs of Technical Lemmas in Section 4.10

4.11.1 Proof of Lemma 4.10.1

In this section we prove Lemma 4.10.1, we need the following lemma from Arora et al. (2019):

Lemma 4.11.1 (Theorem 3.1, Arora et al. (2019))**.** Fix $\epsilon>0$ and $\delta\in(0,1)$. Suppose that

$$m=\Omega\bigg(\frac{L^6\log(L/\delta)}{\epsilon^4}\bigg),$$

then for any $i,j\in[TK]$, with probability at least $1-\delta$ over random initialization of $\boldsymbol{\theta}_0$, we have

$$|\langle\mathbf{g}(\mathbf{x}^i;\boldsymbol{\theta}_0),\mathbf{g}(\mathbf{x}^j;\boldsymbol{\theta}_0)\rangle/m-\mathbf{H}_{i,j}|\leq\epsilon. \tag{4.11.1}$$

Proof of Lemma 4.10.1. Taking union bound over $i,j\in[TK]$, we have that if

$$m=\Omega\bigg(\frac{L^6\log(T^2K^2L/\delta)}{\epsilon^4}\bigg),$$

then with probability at least $1-\delta$, (4.11.1) holds for all $(i,j)\in[TK]\times[TK]$. Therefore, we have

$$\|\mathbf{G}^\top\mathbf{G}-\mathbf{H}\|_F=\sqrt{\sum_{i=1}^{TK}\sum_{j=1}^{TK}|\langle\mathbf{g}(\mathbf{x}^i;\boldsymbol{\theta}_0),\mathbf{g}(\mathbf{x}^j;\boldsymbol{\theta}_0)\rangle/m-\mathbf{H}_{i,j}|^2}\leq TK\epsilon.$$

□

4.11.2 Proof of Lemma 4.10.2

In this section we prove Lemma 4.10.2. During the proof, for simplicity, we omit the subscript t by default. We define the following quantities:

$$\begin{aligned}
&\mathbf{J}^{(j)} = \Big(\mathbf{g}(\mathbf{x}_{1,a_1};\boldsymbol{\theta}^{(j)}),\ldots,\mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta}^{(j)})\Big) \in \mathbb{R}^{(md+m^2(L-2)+m)\times t},\\
&\mathbf{H}^{(j)} = [\mathbf{J}^{(j)}]^\top \mathbf{J}^{(j)} \in \mathbb{R}^{t\times t},\\
&\mathbf{f}^{(j)} = (f(\mathbf{x}_{1,a_1};\boldsymbol{\theta}^{(j)}),\ldots,f(\mathbf{x}_{t,a_t};\boldsymbol{\theta}^{(j)}))^\top \in \mathbb{R}^{t\times 1},\\
&\mathbf{y} = (r_{1,a_1},\ldots,r_{t,a_t}) \in \mathbb{R}^{t\times 1}.
\end{aligned}$$

Then the update rule of $\boldsymbol{\theta}^{(j)}$ can be written as follows:

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} - \eta\big[\mathbf{J}^{(j)}(\mathbf{f}^{(j)} - \mathbf{y}) + m\lambda(\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{(0)})\big]. \tag{4.11.2}$$

We also define the following auxiliary sequence $\{\widetilde{\boldsymbol{\theta}}^{(k)}\}$ during the proof:

$$\widetilde{\boldsymbol{\theta}}^{(0)} = \boldsymbol{\theta}^{(0)},\ \widetilde{\boldsymbol{\theta}}^{(j+1)} = \widetilde{\boldsymbol{\theta}}^{(j)} - \eta\big[\mathbf{J}^{(0)}([\mathbf{J}^{(0)}]^\top(\widetilde{\boldsymbol{\theta}}^{(j)} - \widetilde{\boldsymbol{\theta}}^{(0)}) - \mathbf{y}) + m\lambda(\widetilde{\boldsymbol{\theta}}^{(j)} - \widetilde{\boldsymbol{\theta}}^{(0)})\big].$$

Next lemma provides perturbation bounds for $\mathbf{J}^{(j)}, \mathbf{H}^{(j)}$ and $\|\mathbf{f}^{(j+1)} - \mathbf{f}^{(j)} - [\mathbf{J}^{(j)}]^\top(\boldsymbol{\theta}^{(j+1)} - \boldsymbol{\theta}^{(j)})\|_2$.

Lemma 4.11.2. There exist constants $\{\bar{C}_i\}_{i=1}^6 > 0$ such that for any $\delta > 0$, if τ satisfies that

$$\bar{C}_1 m^{-3/2} L^{-3/2}[\log(TKL^2/\delta)]^{3/2} \le \tau \le \bar{C}_2 L^{-6}[\log m]^{-3/2},$$

then with probability at least $1-\delta$, if for any $j \in [J]$, $\|\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{(0)}\|_2 \le \tau$, we have the following inequalities for any $j, s \in [J]$,

$$\big\|\mathbf{J}^{(j)}\big\|_F \le \bar{C}_4\sqrt{tmL}, \tag{4.11.3}$$

$$\|\mathbf{J}^{(j)} - \mathbf{J}^{(0)}\|_F \le \bar{C}_5\sqrt{tm\log m}\tau^{1/3}L^{7/2}, \tag{4.11.4}$$

$$\big\|\mathbf{f}^{(s)} - \mathbf{f}^{(j)} - [\mathbf{J}^{(j)}]^\top(\boldsymbol{\theta}^{(s)} - \boldsymbol{\theta}^{(j)})\big\|_2 \le \bar{C}_6\tau^{4/3}L^3\sqrt{tm\log m}, \tag{4.11.5}$$

$$\|\mathbf{y}\|_2 \le \sqrt{t}. \tag{4.11.6}$$

Next lemma gives an upper bound for $\|\mathbf{f}^{(j)} - \mathbf{y}\|_2$.

Lemma 4.11.3. There exist constants $\{\bar{C}_i\}_{i=1}^4 > 0$ such that for any $\delta > 0$, if τ, η satisfy that

$$\begin{aligned}
&\bar{C}_1 m^{-3/2} L^{-3/2} [\log(TKL^2/\delta)]^{3/2} \leq \tau \leq \bar{C}_2 L^{-6} [\log m]^{-3/2},, \\
&\eta \leq \bar{C}_3 (m\lambda + tmL)^{-1}, \\
&\tau^{8/3} \leq \bar{C}_4 m (\lambda\eta)^2 L^{-6} t^{-1} (\log m)^{-1},
\end{aligned}$$

then with probability at least $1 - \delta$, if for any $j \in [J]$, $\|\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{(0)}\|_2 \leq \tau$, we have that for any $j \in [J]$, $\|\mathbf{f}^{(j)} - \mathbf{y}\|_2 \leq 2\sqrt{t}$.

Next lemma gives an upper bound of the distance between auxiliary sequence $\|\widetilde{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}^{(0)}\|_2$.

Lemma 4.11.4. There exist constants $\{\bar{C}_i\}_{i=1}^3 > 0$ such that for any $\delta \in (0, 1)$, if τ, η satisfy that

$$\begin{aligned}
&\bar{C}_1 m^{-3/2} L^{-3/2} [\log(TKL^2/\delta)]^{3/2} \leq \tau \leq \bar{C}_2 L^{-6} [\log m]^{-3/2},, \\
&\eta \leq \bar{C}_3 (tmL + m\lambda)^{-1},
\end{aligned}$$

then with probability at least $1 - \delta$, we have that for any $j \in [J]$,

$$\begin{aligned}
&\big\|\widetilde{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}^{(0)}\big\|_2 \leq \sqrt{t/(m\lambda)}, \\
&\big\|\widetilde{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}^{(0)} - \bar{\mathbf{Z}}^{-1}\bar{\mathbf{b}}/\sqrt{m}\big\|_2 \leq (1 - \eta m\lambda)^{j/2}\sqrt{t/(m\lambda)}
\end{aligned}$$

With above lemmas, we prove Lemma 4.10.2 as follows.

Proof of Lemma 4.10.2. Set $\tau = 2\sqrt{t/(m\lambda)}$. First we assume that $\|\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{(0)}\|_2 \leq \tau$ for all $0 \leq j \leq J$. Then with this assumption and the choice of m, τ, we have that Lemma 4.11.2, 4.11.3 and 4.11.4 hold. Then we have

$$\big\|\boldsymbol{\theta}^{(j+1)} - \widetilde{\boldsymbol{\theta}}^{(j+1)}\big\|_2 = \big\|\boldsymbol{\theta}^{(j)} - \widetilde{\boldsymbol{\theta}}^{(j)} - \eta(\mathbf{J}^{(j)} - \mathbf{J}^{(0)})(\mathbf{f}^{(j)} - \mathbf{y}) - \eta m\lambda(\boldsymbol{\theta}^{(j)} - \widetilde{\boldsymbol{\theta}}^{(j)})$$

$$
\begin{aligned}
&\quad - \eta \mathbf{J}^{(0)}(\mathbf{f}^{(j)} - [\mathbf{J}^{(0)}]^\top(\widetilde{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}^{(0)}))\big\|_2 \\
&= \Big\|(1-\eta m\lambda)(\boldsymbol{\theta}^{(j)} - \widetilde{\boldsymbol{\theta}}^{(j)}) - \eta(\mathbf{J}^{(j)} - \mathbf{J}^{(0)})(\mathbf{f}^{(j)} - \mathbf{y}) \\
&\quad - \eta \mathbf{J}^{(0)}\Big[\mathbf{f}^{(j)} - [\mathbf{J}^{(0)}]^\top(\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{(0)}) + [\mathbf{J}^{(0)}]^\top(\boldsymbol{\theta}^{(j)} - \widetilde{\boldsymbol{\theta}}^{(j)})\Big]\Big\|_2 \\
&\leq \underbrace{\eta\big\|(\mathbf{J}^{(j)} - \mathbf{J}^{(0)})(\mathbf{f}^{(j)} - \mathbf{y})\big\|_2}_{I_1} + \underbrace{\eta\|\mathbf{J}^{(0)}\|_2\big\|\mathbf{f}^{(j)} - [\mathbf{J}^{(0)}](\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{(0)})\big\|_2}_{I_2} \\
&\quad + \underbrace{\big\|\big[\mathbf{I} - \eta(m\lambda\mathbf{I} + \mathbf{H}^{(0)})\big](\widetilde{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}^{(j)})\big\|_2}_{I_3}, \qquad (4.11.7)
\end{aligned}
$$

where the inequality holds due to triangle inequality. We now bound I_1, I_2 and I_3 separately. For I_1, we have

$$
I_1 \leq \eta\big\|\mathbf{J}^{(j)} - \mathbf{J}^{(0)}\big\|_2\|\mathbf{f}^{(j)} - \mathbf{y}\|_2 \leq \eta C_2 t\sqrt{m\log m}\tau^{1/3}L^{7/2}, \qquad (4.11.8)
$$

where $C_2 > 0$ is a constant, the first inequality holds due to the definition of matrix spectral norm and the second inequality holds due to (4.11.4) in Lemma 4.11.2 and Lemma 4.11.3. For I_2, we have

$$
I_2 \leq \eta\big\|\mathbf{J}^{(0)}\big\|_2\Big\|\mathbf{f}^{(j)} - \mathbf{J}^{(0)}(\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{(0)})\Big\|_2 \leq \eta C_3 tmL^{7/2}\tau^{4/3}\sqrt{\log m}, \qquad (4.11.9)
$$

where $C_3 > 0$, the first inequality holds due to matrix spectral norm, the second inequality holds due to (4.11.3) and (4.11.5) in Lemma 4.11.2 and the fact that $\mathbf{f}^{(0)} = \mathbf{0}$ by random initialization over $\boldsymbol{\theta}^{(0)}$. For I_3, we have

$$
I_3 \leq \big\|\mathbf{I} - \eta(m\lambda\mathbf{I} + \mathbf{H}^{(0)})\big\|_2\big\|\widetilde{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}^{(j)}\big\|_2 \leq (1-\eta m\lambda)\big\|\widetilde{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}^{(j)}\big\|_2, \qquad (4.11.10)
$$

where the first inequality holds due to spectral norm inequality, the second inequality holds since

$$
\eta(m\lambda\mathbf{I} + \mathbf{H}^{(0)}) = \eta(m\lambda\mathbf{I} + [\mathbf{J}^{(0)}]^\top\mathbf{J}^{(0)}) \preceq \eta(m\lambda\mathbf{I} + C_1 tmL\mathbf{I}) \preceq \mathbf{I},
$$

for some $C_1 > 0$, the first inequality holds due to (4.11.3) in Lemma 4.11.2, the second inequality holds due to the choice of η.

Substituting (4.11.8), (4.11.9) and (4.11.10) into (4.11.7), we obtain

$$\left\|\boldsymbol{\theta}^{(j+1)}-\widetilde{\boldsymbol{\theta}}^{(j+1)}\right\|_2 \leq (1-\eta m\lambda)\left\|\boldsymbol{\theta}^{(j)}-\widetilde{\boldsymbol{\theta}}^{(j)}\right\|_2 + C_4\big(\eta t\sqrt{m\log m}\tau^{1/3}L^{7/2}+\eta t m L^{7/2}\tau^{4/3}\sqrt{\log m}\big), \tag{4.11.11}$$

where $C_4>0$ is a constant. By recursively applying (4.11.11) from 0 to j, we have

$$\begin{aligned}\left\|\boldsymbol{\theta}^{(j+1)}-\widetilde{\boldsymbol{\theta}}^{(j+1)}\right\|_2 &\leq C_4\frac{\eta t\sqrt{m\log m}\tau^{1/3}L^{7/2}+\eta t m L^{7/2}\tau^{4/3}\sqrt{\log m}}{\eta m\lambda}\\ &= C_5 m^{-2/3}\sqrt{\log m}L^{7/2}t^{5/3}\lambda^{-5/3}(1+\sqrt{t/\lambda})\\ &\leq \frac{\tau}{2},\end{aligned} \tag{4.11.12}$$

where $C_5>0$ is a constant, the equality holds by the definition of τ, the last inequality holds due to the choice of m, where

$$m^{1/6}\geq C_6\sqrt{\log m}L^{7/2}t^{7/6}\lambda^{-7/6}(1+\sqrt{t/\lambda}),$$

and $C_6>0$ is a constant. Thus, for any $j\in[J]$, we have

$$\|\boldsymbol{\theta}^{(j)}-\boldsymbol{\theta}^{(0)}\|_2 \leq \|\widetilde{\boldsymbol{\theta}}^{(j)}-\boldsymbol{\theta}^{(0)}\|_2 + \|\boldsymbol{\theta}^{(j)}-\widetilde{\boldsymbol{\theta}}^{(j)}\|_2 \leq \sqrt{t/(m\lambda)}+\tau/2=\tau, \tag{4.11.13}$$

where the first inequality holds due to triangle inequality, the second inequality holds due to Lemma 4.11.4. (4.11.13) suggests that our assumption $\|\boldsymbol{\theta}^{(j)}-\boldsymbol{\theta}^{(0)}\|_2\leq\tau$ holds for any j. Note that we have the following inequality by Lemma 4.11.4:

$$\left\|\widetilde{\boldsymbol{\theta}}^{(j)}-\boldsymbol{\theta}^{(0)}-(\bar{\mathbf{Z}})^{-1}\bar{\mathbf{b}}/\sqrt{m}\right\|_2 \leq (1-\eta m\lambda)^j\sqrt{t/(m\lambda)}. \tag{4.11.14}$$

Using (4.11.12) and (4.11.14), we have

$$\left\|\boldsymbol{\theta}^{(j)}-\boldsymbol{\theta}^{(0)}-\bar{\mathbf{Z}}^{-1}\bar{\mathbf{b}}/\sqrt{m}\right\|_2 \leq (1-\eta m\lambda)^{j/2}\sqrt{t/(m\lambda)}+C_5 m^{-2/3}\sqrt{\log m}L^{7/2}t^{5/3}\lambda^{-5/3}(1+\sqrt{t/\lambda}).$$

This completes the proof. □

4.11.3 Proof of Lemma 4.10.3

In this section we prove Lemma 4.10.3.

Proof of Lemma 4.10.3. Set $\tau = 2\sqrt{t/(m\lambda)}$. By Lemma 4.10.2 we have that $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_0\|_2 \leq \tau$ for $i \in [t]$. $\|\mathbf{Z}_t\|_2$ can be bounded as follows.

$$\begin{aligned}
\|\mathbf{Z}_t\|_2 &= \Big\|\lambda\mathbf{I} + \sum_{i=1}^{t} \mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_{i-1})\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_{i-1})^\top/m\Big\|_2 \\
&\leq \lambda + \Big\|\lambda\mathbf{I} + \sum_{i=1}^{t} \mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_{i-1})\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_{i-1})^\top/m\Big\|_2 \\
&\leq \lambda + \sum_{i=1}^{t} \big\|\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_{i-1})\big\|_2^2/m \\
&\leq \lambda + C_0 tL,
\end{aligned}$$

where $C_0 > 0$ is a constant, the first inequality holds due to the fact that $\|\mathbf{a}\mathbf{a}^\top\|_F = \|\mathbf{a}\|_2^2$, the second inequality holds due to Lemma 4.10.6 with the fact that $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_0\|_2 \leq \tau$. We bound $\|\mathbf{Z}_t - \bar{\mathbf{Z}}_t\|_2$ as follows. We have

$$\begin{aligned}
\|\mathbf{Z}_t - \bar{\mathbf{Z}}_t\|_F &= \Big\|\sum_{i=1}^{t} \Big(\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_0)\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_0)^\top - \mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_i)\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_i)^\top\Big)/m\Big\|_F \\
&\leq \sum_{i=1}^{t} \Big\|\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_0)\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_0)^\top - \mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_i)\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_i)^\top\Big\|_F/m \\
&\leq \sum_{i=1}^{t} \Big(\big\|\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_0)\big\|_2 + \big\|\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_i)\big\|_2\Big)\big\|\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_0) - \mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_i)\big\|_2/m,
\end{aligned} \tag{4.11.15}$$

where the first inequality holds due to triangle inequality, the second inequality holds the fact that $\|\mathbf{a}\mathbf{a}^\top - \mathbf{b}\mathbf{b}^\top\|_F \leq (\|\mathbf{a}\|_2 + \|\mathbf{b}\|_2)\|\mathbf{a} - \mathbf{b}\|_2$ for any vectors $\mathbf{a}, \mathbf{b}$. To bound (4.11.15), we have

$$\big\|\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_0)\big\|_2, \big\|\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_i)\big\|_2 \leq C_1\sqrt{mL}, \tag{4.11.16}$$

where $C_1 > 0$ is a constant, the inequality holds due to Lemma 4.10.6 with the fact that $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_0\|_2 \leq \tau$. We also have

$$\big\|\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_0) - \mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_i)\big\|_2 \leq C_2\sqrt{\log m}\tau^{1/3}L^3\|\mathbf{g}(\mathbf{x}_j;\boldsymbol{\theta}_0)\|_2 \leq C_3\sqrt{m\log m}\tau^{1/3}L^{7/2}, \tag{4.11.17}$$

where $C_2, C_3 > 0$ are constants, the first inequality holds due to Lemma 4.10.5 with the fact that $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_0\|_2 \leq \tau$, the second inequality holds due to Lemma 4.10.6. Substituting (4.11.16) and (4.11.17) into (4.11.15), we have

$$\|\mathbf{Z}_t - \bar{\mathbf{Z}}_t\|_F \leq C_4 t\sqrt{\log m}\tau^{1/3}L^4,$$

where $C_4 > 0$ is a constant. We now bound $\log\det\bar{\mathbf{Z}}_t - \log\det\mathbf{Z}_t$. It is easy to verify that $\bar{\mathbf{Z}}_t = \lambda\mathbf{I} + \bar{\mathbf{J}}\bar{\mathbf{J}}^\top$, $\mathbf{Z}_t = \lambda\mathbf{I} + \mathbf{J}\mathbf{J}^\top$, where

$$\bar{\mathbf{J}} = \Big(\mathbf{g}(\mathbf{x}_{1,a_1};\boldsymbol{\theta}_0),\ldots,\mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta}_0)\Big)/\sqrt{m},$$
$$\mathbf{J} = \Big(\mathbf{g}(\mathbf{x}_{1,a_1};\boldsymbol{\theta}_0),\ldots,\mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta}_{t-1})\Big)/\sqrt{m}.$$

We have the following inequalities:

$$\begin{aligned}
\log\frac{\det(\bar{\mathbf{Z}}_t)}{\det(\lambda\mathbf{I})} - \log\frac{\det(\mathbf{Z}_t)}{\det(\lambda\mathbf{I})} &= \log\det(\mathbf{I} + \bar{\mathbf{J}}\bar{\mathbf{J}}^\top/\lambda) - \log\det(\mathbf{I} + \mathbf{J}\mathbf{J}^\top/\lambda)\\
&= \log\det(\mathbf{I} + \bar{\mathbf{J}}^\top\bar{\mathbf{J}}/\lambda) - \log\det(\mathbf{I} + \mathbf{J}^\top\mathbf{J}/\lambda)\\
&\leq \langle(\mathbf{I} + \mathbf{J}^\top\mathbf{J}/\lambda)^{-1}, \bar{\mathbf{J}}^\top\bar{\mathbf{J}} - \mathbf{J}^\top\mathbf{J}\rangle\\
&\leq \|(\mathbf{I} + \mathbf{J}^\top\mathbf{J}/\lambda)^{-1}\|_F\|\bar{\mathbf{J}}^\top\bar{\mathbf{J}} - \mathbf{J}^\top\mathbf{J}\|_F\\
&\leq \sqrt{t}\|(\mathbf{I} + \mathbf{J}^\top\mathbf{J}/\lambda)^{-1}\|_2\|\bar{\mathbf{J}}^\top\bar{\mathbf{J}} - \mathbf{J}^\top\mathbf{J}\|_F\\
&\leq \sqrt{t}\|\bar{\mathbf{J}}^\top\bar{\mathbf{J}} - \mathbf{J}^\top\mathbf{J}\|_F,
\end{aligned} \tag{4.11.18}$$

where the second equality holds due to the fact that $\det(\mathbf{I} + \mathbf{A}\mathbf{A}^\top) = \det(\mathbf{I} + \mathbf{A}^\top\mathbf{A})$, the first inequality holds due to the fact that $\log\det$ function is convex, the second inequality hold due to the fact that $\langle\mathbf{A},\mathbf{B}\rangle \leq \|\mathbf{A}\|_F\|\mathbf{B}\|_F$, the third inequality holds since $\mathbf{I} + \mathbf{J}^\top\mathbf{J}/\lambda$ is a t-dimension matrix, the fourth inequality holds since $\mathbf{I} + \mathbf{J}^\top\mathbf{J}/\lambda \succeq \mathbf{I}$. We have

$$\|\bar{\mathbf{J}}^\top\bar{\mathbf{J}} - \mathbf{J}^\top\mathbf{J}\|_F$$

$$
\begin{aligned}
&\leq t \max_{1\leq i,j\leq t} \left| \mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_0)^\top \mathbf{g}(\mathbf{x}_{j,a_j};\boldsymbol{\theta}_0) - \mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_i)^\top \mathbf{g}(\mathbf{x}_{j,a_j};\boldsymbol{\theta}_j) \right| / m \\
&\leq t \max_{1\leq i,j\leq t} \left\| \mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_0) - \mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_i) \right\|_2 \left\| \mathbf{g}(\mathbf{x}_{j,a_j};\boldsymbol{\theta}_j) \right\|_2 / m \\
&\quad + \left\| \mathbf{g}(\mathbf{x}_{j,a_j};\boldsymbol{\theta}_0) - \mathbf{g}(\mathbf{x}_{j,a_j};\boldsymbol{\theta}_j) \right\|_2 \left\| \mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}_0) \right\|_2 / m \\
&\leq C_5 t \sqrt{\log m} \tau^{1/3} L^4, \qquad (4.11.19)
\end{aligned}
$$

where $C_5 > 0$ is a constant, the first inequality holds due to the fact that $\|\mathbf{A}\|_F \leq t \max |\mathbf{A}_{i,j}|$ for any $\mathbf{A} \in \mathbb{R}^{t\times t}$, the second inequality holds due to the fact $|\mathbf{a}^\top \mathbf{a}' - \mathbf{b}^\top \mathbf{b}'| \leq \|\mathbf{a}-\mathbf{b}\|_2 \|\mathbf{b}'\|_2 + \|\mathbf{a}'-\mathbf{b}'\|_2 \|\mathbf{a}\|_2$, the third inequality holds due to (4.11.16) and (4.11.17). Substituting (4.11.19) into (4.11.18), we obtain

$$
\log \frac{\det(\bar{\mathbf{Z}}_t)}{\det(\lambda \mathbf{I})} - \log \frac{\det(\mathbf{Z}_t)}{\det(\lambda \mathbf{I})} \leq C_5 t^{3/2} \sqrt{\log m} \tau^{1/3} L^4.
$$

Using the same method, we also have

$$
\log \frac{\det(\mathbf{Z}_t)}{\det(\lambda \mathbf{I})} - \log \frac{\det(\bar{\mathbf{Z}}_t)}{\det(\lambda \mathbf{I})} \leq C_5 t^{3/2} \sqrt{\log m} \tau^{1/3} L^4.
$$

This completes our proof.

□

4.12 Proofs of Lemmas in Section 4.11

4.12.1 Proof of Lemma 4.11.2

In this section we give the proof of Lemma 4.11.2.

Proof of Lemma 4.11.2. It can be verified that τ satisfies the conditions of Lemmas 4.10.4, 4.10.5 and 4.10.6. Thus, Lemmas 4.10.4, 4.10.5 and 4.10.6 hold. We will show that for any $j \in [J]$, the following inequalities hold. First, we have

$$
\left\| \mathbf{J}^{(j)} \right\|_F \leq \sqrt{t} \max_{i\in[t]} \left\| \mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}^{(j)}) \right\|_2 \leq C_1 \sqrt{tmL}, \qquad (4.12.1)
$$

where $C_1 > 0$ is a constant, the first inequality holds due to the fact that $\|\mathbf{J}^{(j)}\|_F \leq \sqrt{t}\|\mathbf{J}^{(j)}\|_{2,\infty}$, the second inequality holds due to Lemma 4.10.6.

We also have

$$\|\mathbf{J}^{(j)} - \mathbf{J}^{(0)}\|_F \leq C_2\sqrt{\log m}\tau^{1/3}L^3\|\mathbf{J}^{(0)}\|_F \leq C_3\sqrt{tm\log m}\tau^{1/3}L^{7/2}, \tag{4.12.2}$$

where $C_2, C_3 > 0$ are constants, the first inequality holds due to Lemma 4.10.5 with the assumption that $\|\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{(0)}\|_2 \leq \tau$, the second inequality holds due to (4.12.1).

We also have

$$\begin{aligned}
&\big\|\mathbf{f}^{(s)} - \mathbf{f}^{(j)} - [\mathbf{J}^{(j)}]^\top(\boldsymbol{\theta}^{(s)} - \boldsymbol{\theta}^{(j)})\big\|_2 \\
&\leq \max_{i\in[t]} \sqrt{t}\big|f(\mathbf{x}_{i,a_i};\boldsymbol{\theta}^{(s)}) - f(\mathbf{x}_{i,a_i};\boldsymbol{\theta}^{(j)}) - \langle\mathbf{g}(\mathbf{x}_{i,a_i};\boldsymbol{\theta}^{(j)}), \boldsymbol{\theta}^{(s)} - \boldsymbol{\theta}^{(j)}\rangle\big| \\
&\leq C_4\tau^{4/3}L^3\sqrt{tm\log m},
\end{aligned}$$

where $C_4 > 0$ is a constant, the first inequality holds due to the the fact that $\|\mathbf{x}\|_2 \leq \sqrt{t}\max|x_i|$ for any $\mathbf{x} \in \mathbb{R}^t$, the second inequality holds due to Lemma 4.10.4 with the assumption that $\|\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{(0)}\|_2 \leq \tau, \|\boldsymbol{\theta}^{(s)} - \boldsymbol{\theta}^{(0)}\|_2 \leq \tau$.

For $\|\mathbf{y}\|_2$, we have $\|\mathbf{y}\|_2 \leq \sqrt{t}\max_{1\leq i\leq t}|r(\mathbf{x}_{i,a_i})| \leq \sqrt{t}$. This completes our proof.

□

4.12.2 Proof of Lemma 4.11.3

Proof of Lemma 4.11.3. It can be verified that τ satisfies the conditions of Lemma 4.11.2, thus Lemma 4.11.2 holds. Recall that the loss function L is defined as

$$L(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{f}(\boldsymbol{\theta}) - \mathbf{y}\|_2^2 + \frac{m\lambda}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}\|_2^2.$$

We define $\mathbf{J}(\boldsymbol{\theta})$ and $\mathbf{f}(\boldsymbol{\theta})$ as follows:

$$\mathbf{J}(\boldsymbol{\theta}) = \Big(\mathbf{g}(\mathbf{x}_{1,a_1};\boldsymbol{\theta}), \ldots, \mathbf{g}(\mathbf{x}_{t,a_t};\boldsymbol{\theta})\Big) \in \mathbb{R}^{(md+m^2(L-2)+m)\times t},$$

$$\mathbf{f}(\boldsymbol{\theta}) = (f(\mathbf{x}_{1,a_1};\boldsymbol{\theta}),\ldots,f(\mathbf{x}_{t,a_t};\boldsymbol{\theta}))^\top \in \mathbb{R}^{t\times 1}.$$

Suppose $\|\boldsymbol{\theta}-\boldsymbol{\theta}^{(0)}\|_2 \le \tau$. Then by the fact that $\|\cdot\|_2^2/2$ is 1-strongly convex and 1-smooth, we have the following inequalities:

$$\begin{aligned}
&L(\boldsymbol{\theta}') - L(\boldsymbol{\theta}) \\
&\le \langle \mathbf{f}(\boldsymbol{\theta})-\mathbf{y}, \mathbf{f}(\boldsymbol{\theta}')-\mathbf{f}(\boldsymbol{\theta})\rangle + \frac{1}{2}\big\|\mathbf{f}(\boldsymbol{\theta}')-\mathbf{f}(\boldsymbol{\theta})\big\|_2^2 + m\lambda\langle \boldsymbol{\theta}-\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}'-\boldsymbol{\theta}\rangle + \frac{m\lambda}{2}\big\|\boldsymbol{\theta}'-\boldsymbol{\theta}\big\|_2^2 \\
&= \langle \mathbf{f}(\boldsymbol{\theta})-\mathbf{y}, [\mathbf{J}(\boldsymbol{\theta})]^\top(\boldsymbol{\theta}'-\boldsymbol{\theta})+\mathbf{e}\rangle + \frac{1}{2}\big\|[\mathbf{J}(\boldsymbol{\theta})]^\top(\boldsymbol{\theta}'-\boldsymbol{\theta})+\mathbf{e}\big\|_2^2 \\
&\quad + m\lambda\langle \boldsymbol{\theta}-\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}'-\boldsymbol{\theta}\rangle + \frac{m\lambda}{2}\big\|\boldsymbol{\theta}'-\boldsymbol{\theta}\big\|_2^2 \\
&= \langle \mathbf{J}(\boldsymbol{\theta})(\mathbf{f}(\boldsymbol{\theta})-\mathbf{y}) + m\lambda(\boldsymbol{\theta}-\boldsymbol{\theta}^{(0)}), \boldsymbol{\theta}'-\boldsymbol{\theta}\rangle + \langle \mathbf{f}(\boldsymbol{\theta})-\mathbf{y}, \mathbf{e}\rangle \\
&\quad + \frac{1}{2}\big\|[\mathbf{J}(\boldsymbol{\theta})]^\top(\boldsymbol{\theta}'-\boldsymbol{\theta})+\mathbf{e}\big\|_2^2 + \frac{m\lambda}{2}\big\|\boldsymbol{\theta}'-\boldsymbol{\theta}\big\|_2^2 \\
&= \langle \nabla L(\boldsymbol{\theta}), \boldsymbol{\theta}'-\boldsymbol{\theta}\rangle + \underbrace{\langle \mathbf{f}(\boldsymbol{\theta})-\mathbf{y}, \mathbf{e}\rangle + \frac{1}{2}\big\|[\mathbf{J}(\boldsymbol{\theta})]^\top(\boldsymbol{\theta}'-\boldsymbol{\theta})+\mathbf{e}\big\|_2^2 + \frac{m\lambda}{2}\big\|\boldsymbol{\theta}'-\boldsymbol{\theta}\big\|_2^2}_{I_1}, \qquad (4.12.3)
\end{aligned}$$

where $\mathbf{e} = \mathbf{f}(\boldsymbol{\theta}') - \mathbf{f}(\boldsymbol{\theta}) - \mathbf{J}(\boldsymbol{\theta})^\top(\boldsymbol{\theta}'-\boldsymbol{\theta})$. I_1 can be bounded as follows:

$$\begin{aligned}
I_1 &\le \|\mathbf{f}(\boldsymbol{\theta})-\mathbf{y}\|_2\|\mathbf{e}\|_2 + \|\mathbf{J}(\boldsymbol{\theta})\|_2^2\|\boldsymbol{\theta}'-\boldsymbol{\theta}\|_2^2 + \|\mathbf{e}\|_2^2 + \frac{m\lambda}{2}\big\|\boldsymbol{\theta}'-\boldsymbol{\theta}\big\|_2^2 \\
&\le \frac{C_1}{2}\left((m\lambda+tmL)\big\|\boldsymbol{\theta}'-\boldsymbol{\theta}\big\|_2^2\right) + \|\mathbf{f}(\boldsymbol{\theta})-\mathbf{y}\|_2\|\mathbf{e}\|_2 + \|\mathbf{e}\|_2^2, \qquad (4.12.4)
\end{aligned}$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds due to the fact that $\|\mathbf{J}(\boldsymbol{\theta})\|_2 \le C_2\sqrt{tmL}$ with $\|\boldsymbol{\theta}-\boldsymbol{\theta}^{(0)}\|_2 \le \tau$ by (4.11.3) in Lemma 4.11.2. Substituting (4.12.4) into (4.12.3), we obtain

$$L(\boldsymbol{\theta}') - L(\boldsymbol{\theta}) \le \langle \nabla L(\boldsymbol{\theta}), \boldsymbol{\theta}'-\boldsymbol{\theta}\rangle + \frac{C_1}{2}\left((m\lambda+tmL)\big\|\boldsymbol{\theta}'-\boldsymbol{\theta}\big\|_2^2\right) + \|\mathbf{f}(\boldsymbol{\theta})-\mathbf{y}\|_2\|\mathbf{e}\|_2 + \|\mathbf{e}\|_2^2. \qquad (4.12.5)$$

Taking $\boldsymbol{\theta}' = \boldsymbol{\theta} - \eta\nabla L(\boldsymbol{\theta})$, then by (4.12.5), we have

$$L\big(\boldsymbol{\theta}-\eta\nabla L(\boldsymbol{\theta})\big) - L(\boldsymbol{\theta}) \le -\eta\|\nabla L(\boldsymbol{\theta})\|_2^2\big[1 - C_1(m\lambda+tmL)\eta\big] + \|\mathbf{f}(\boldsymbol{\theta})-\mathbf{y}\|_2\|\mathbf{e}\|_2 + \|\mathbf{e}\|_2^2. \qquad (4.12.6)$$

By the 1-strongly convexity of $\|\cdot\|_2^2$, we further have

$$
\begin{aligned}
&L(\boldsymbol{\theta}') - L(\boldsymbol{\theta}) \\
&\geq \langle \mathbf{f}(\boldsymbol{\theta}) - \mathbf{y}, \mathbf{f}(\boldsymbol{\theta}') - \mathbf{f}(\boldsymbol{\theta}) \rangle + m\lambda \langle \boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle + \frac{m\lambda}{2} \big\| \boldsymbol{\theta}' - \boldsymbol{\theta} \big\|_2^2 \\
&= \langle \mathbf{f}(\boldsymbol{\theta}) - \mathbf{y}, [\mathbf{J}(\boldsymbol{\theta})]^\top (\boldsymbol{\theta}' - \boldsymbol{\theta}) + \mathbf{e} \rangle + m\lambda \langle \boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle + \frac{m\lambda}{2} \big\| \boldsymbol{\theta}' - \boldsymbol{\theta} \big\|_2^2 \\
&= \langle \nabla L(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle + \frac{m\lambda}{2} \big\| \boldsymbol{\theta}' - \boldsymbol{\theta} \big\|_2^2 + \langle \mathbf{f}(\boldsymbol{\theta}) - \mathbf{y}, \mathbf{e} \rangle \\
&\geq \langle \nabla L(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle + \frac{m\lambda}{2} \big\| \boldsymbol{\theta}' - \boldsymbol{\theta} \big\|_2^2 - \|\mathbf{f}(\boldsymbol{\theta}) - \mathbf{y}\|_2 \|\mathbf{e}\|_2 \\
&\geq -\frac{\|\nabla L(\boldsymbol{\theta})\|_2^2}{2m\lambda} - \|\mathbf{f}(\boldsymbol{\theta}) - \mathbf{y}\|_2 \|\mathbf{e}\|_2, \qquad (4.12.7)
\end{aligned}
$$

where the second inequality holds due to Cauchy-Schwarz inequality, the last inequality holds due to the fact that $\langle \mathbf{a}, \mathbf{x} \rangle + c\|\mathbf{x}\|_2^2 \geq -\|\mathbf{a}\|_2^2/(4c)$ for any vectors $\mathbf{a}, \mathbf{x}$ and $c > 0$. Substituting (4.12.7) into (4.12.6), we obtain

$$
\begin{aligned}
&L\big(\boldsymbol{\theta} - \eta \nabla L(\boldsymbol{\theta})\big) - L(\boldsymbol{\theta}) \\
&\leq 2m\lambda\eta(1 - C_1(m\lambda + tmL)\eta)\big[L(\boldsymbol{\theta}') - L(\boldsymbol{\theta}) + \|\mathbf{f}(\boldsymbol{\theta}) - \mathbf{y}\|_2 \|\mathbf{e}\|_2\big] + \|\mathbf{f}(\boldsymbol{\theta}) - \mathbf{y}\|_2 \|\mathbf{e}\|_2 + \|\mathbf{e}\|_2^2 \\
&\leq m\lambda\eta\big[L(\boldsymbol{\theta}') - L(\boldsymbol{\theta}) + \|\mathbf{f}(\boldsymbol{\theta}) - \mathbf{y}\|_2 \|\mathbf{e}\|_2\big] + \|\mathbf{f}(\boldsymbol{\theta}) - \mathbf{y}\|_2 \|\mathbf{e}\|_2 + \|\mathbf{e}\|_2^2 \\
&\leq m\lambda\eta\big[L(\boldsymbol{\theta}') - L(\boldsymbol{\theta}) + \|\mathbf{f}(\boldsymbol{\theta}) - \mathbf{y}\|_2^2/8 + 2\|\mathbf{e}\|_2^2\big] + m\lambda\eta\|\mathbf{f}(\boldsymbol{\theta}) - \mathbf{y}\|_2^2/8 + 2\|\mathbf{e}\|_2^2/(m\lambda\eta) + \|\mathbf{e}\|_2^2 \\
&\leq m\lambda\eta\big(L(\boldsymbol{\theta}') - L(\boldsymbol{\theta})/2\big) + \|\mathbf{e}\|_2^2\big(1 + 2m\lambda\eta + 2/(m\lambda\eta)\big), \qquad (4.12.8)
\end{aligned}
$$

where the second inequality holds due to the choice of η, third inequality holds due to Young's inequality, fourth inequality holds due to the fact that $\|\mathbf{f}(\boldsymbol{\theta}) - \mathbf{y}\|_2^2 \leq 2L(\boldsymbol{\theta})$. Now taking $\boldsymbol{\theta} = \boldsymbol{\theta}^{(j)}$ and $\boldsymbol{\theta}' = \boldsymbol{\theta}^{(0)}$, rearranging (4.12.8), with the fact that $\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} - \eta \nabla L(\boldsymbol{\theta}^{(j)})$, we have

$$
\begin{aligned}
&L(\boldsymbol{\theta}^{(j+1)}) - L(\boldsymbol{\theta}^{(0)}) \\
&\leq (1 - m\lambda\eta/2)\big[L(\boldsymbol{\theta}^{(j)}) - L(\boldsymbol{\theta}^{(0)})\big] + m\lambda\eta/2 L(\boldsymbol{\theta}^{(0)}) + \|\mathbf{e}\|_2^2\big(1 + 2m\lambda\eta + 2/(m\lambda\eta)\big) \\
&\leq (1 - m\lambda\eta/2)\big[L(\boldsymbol{\theta}^{(j)}) - L(\boldsymbol{\theta}^{(0)})\big] + m\lambda\eta/2 \cdot t + m\lambda\eta/2 \cdot t \\
&\leq (1 - m\lambda\eta/2)\big[L(\boldsymbol{\theta}^{(j)}) - L(\boldsymbol{\theta}^{(0)})\big] + m\lambda\eta t, \qquad (4.12.9)
\end{aligned}
$$

where the second inequality holds due to the fact that $L(\boldsymbol{\theta}^{(0)}) = \|\mathbf{f}(\boldsymbol{\theta}^{(0)}) - \mathbf{y}\|_2^2/2 = \|\mathbf{y}\|_2^2/2 \leq t$, and

$$\big(1 + 2m\lambda\eta + 2/(m\lambda\eta)\big)\|\mathbf{e}\|_2^2 \leq 3/(m\lambda\eta) \cdot C_2\tau^{8/3}L^6 tm\log m \leq tm\lambda\eta/2, \tag{4.12.10}$$

where the first inequality holds due to (4.11.5) in Lemma 4.11.2, the second inequality holds due to the choice of τ. Recursively applying (4.12.9) for u times, we have

$$L(\boldsymbol{\theta}^{(j+1)}) - L(\boldsymbol{\theta}^{(0)}) \leq \frac{m\lambda\eta t}{m\lambda\eta/2} = 2t,$$

which implies that $\|\mathbf{f}^{(j+1)} - \mathbf{y}\|_2 \leq 2\sqrt{t}$. This completes our proof. □

4.12.3 Proof of Lemma 4.11.4

In this section we prove Lemma 4.11.4.

Proof of Lemma 4.11.4. It can be verified that τ satisfies the conditions of Lemma 4.11.2, thus Lemma 4.11.2 holds. It is worth noting that $\widetilde{\boldsymbol{\theta}}^{(j)}$ is the sequence generated by applying gradient descent on the following problem:

$$\min_{\boldsymbol{\theta}} \widetilde{\mathcal{L}}(\boldsymbol{\theta}) = \frac{1}{2}\|[\mathbf{J}^{(0)}]^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}) - \mathbf{y}\|_2^2 + \frac{m\lambda}{2}\big\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}\big\|_2^2.$$

Then $\|\boldsymbol{\theta}^{(0)} - \widetilde{\boldsymbol{\theta}}^{(j)}\|_2$ can be bounded as

$$\begin{aligned}
\frac{m\lambda}{2}\|\boldsymbol{\theta}^{(0)} - \widetilde{\boldsymbol{\theta}}^{(j)}\|_2^2 &\leq \frac{1}{2}\|[\mathbf{J}^{(0)}]^\top(\widetilde{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}^{(0)}) - \mathbf{y}\|_2^2 + \frac{m\lambda}{2}\big\|\widetilde{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}^{(0)}\big\|_2^2 \\
&\leq \frac{1}{2}\|[\mathbf{J}^{(0)}]^\top(\widetilde{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^{(0)}) - \mathbf{y}\|_2^2 + \frac{m\lambda}{2}\big\|\widetilde{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^{(0)}\big\|_2^2 \\
&\leq t/2,
\end{aligned}$$

where the first inequality holds trivially, the second inequality holds due to the monotonic decreasing property brought by gradient descent, the third inequality holds due to (4.11.6) in Lemma 4.11.2. It is easy to verify that $\widetilde{\mathcal{L}}$ is a $m\lambda$-strongly convex and function and $C_1(tmL + m\lambda)$-smooth function, since

$$\nabla^2\widetilde{\mathcal{L}} \preceq \big(\big\|\mathbf{J}^{(0)}\big\|_2^2 + m\lambda\big)\mathbf{I} \preceq C_1(tmL + m\lambda),$$

where the first inequality holds due to the definition of $\widetilde{\mathcal{L}}$, the second inequality holds due to (4.11.3) in Lemma 4.11.2. Since we choose $\eta \leq C_2(tmL + m\lambda)^{-1}$ for some small enough $C_2 > 0$, then by standard results of gradient descent on ridge linear regression, $\widetilde{\boldsymbol{\theta}}^{(j)}$ converges to $\boldsymbol{\theta}^{(0)} + (\bar{\mathbf{Z}})^{-1}\bar{\mathbf{b}}/\sqrt{m}$ with the convergence rate

$$
\begin{aligned}
\big\|\widetilde{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}^{(0)} - \bar{\mathbf{Z}}^{-1}\mathbf{b}/\sqrt{m}\big\|_2^2 &\leq (1-\eta m\lambda)^j \cdot \frac{2}{m\lambda}(\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}\big(\boldsymbol{\theta}^{(0)} + \bar{\mathbf{Z}}^{-1}\mathbf{b}/\sqrt{m}\big)) \\
&\leq \frac{2(1-\eta m\lambda)^j}{m\lambda}\mathcal{L}(\boldsymbol{\theta}^{(0)}) \\
&= \frac{2(1-\eta m\lambda)^j}{m\lambda} \cdot \frac{\|\mathbf{y}\|_2^2}{2} \\
&\leq (1-\eta m\lambda)^j t,
\end{aligned}
$$

where the first inequality holds due to the convergence result for gradient descent and the fact that $\boldsymbol{\theta}^{(0)} + (\bar{\mathbf{Z}})^{-1}\bar{\mathbf{b}}/\sqrt{m}$ is the minimal solution to $\mathcal{L}$, the second inequality holds since $\mathcal{L} \geq 0$, the last inequality holds due to Lemma 4.11.2.

□

4.13 A Variant of NeuralUCB

In this section, we present a variant of NeuralUCB called NeuralUCB$_0$. Compared with Algorithm 5, The main differences between NeuralUCB and NeuralUCB$_0$ are as follows: NeuralUCB uses gradient descent to train a deep neural network to learn the reward function $h(\mathbf{x})$ based on observed contexts and rewards. In contrast, NeuralUCB$_0$ uses matrix inversions to obtain parameters in closed forms. At each round, NeuralUCB uses the current DNN parameters ($\boldsymbol{\theta}_t$) to compute an upper confidence bound. In contrast, NeuralUCB$_0$ computes the UCB using the initial parameters ($\boldsymbol{\theta}_0$).

Algorithm 7 NeuralUCB$_0$

Input: number of rounds T, regularization parameter λ, exploration parameter ν, confidence parameter δ, norm parameter S, network width m, network depth L

Initialization: Generate each entry of $\mathbf{W}_l$ independently from $N(0, 2/m)$ for $1 \le l \le L-1$, and each entry of $\mathbf{W}_L$ independently from $N(0, 1/m)$. Define $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{g}(\mathbf{x}; \boldsymbol{\theta}_0)/\sqrt{m}$, where $\boldsymbol{\theta}_0 = [\mathrm{vec}(\mathbf{W}_1)^\top, \ldots, \mathrm{vec}(\mathbf{W}_L)^\top]^\top \in \mathbb{R}^p$

$\mathbf{Z}_0 = \lambda \mathbf{I},\ \mathbf{b}_0 = \mathbf{0}$

for $t = 1, \ldots, T$ **do**

Observe $\{\mathbf{x}_{t,a}\}_{a=1}^K$ and compute

$$(a_t, \widetilde{\boldsymbol{\theta}}_{t,a_t}) = \underset{a \in [K], \boldsymbol{\theta} \in \mathcal{C}_{t-1}}{\mathrm{argmax}} \langle \boldsymbol{\phi}(\mathbf{x}_{t,a}), \boldsymbol{\theta} - \boldsymbol{\theta}_0 \rangle \tag{4.13.1}$$

Play a_t and receive reward r_{t,a_t}

Compute

$$\mathbf{Z}_t = \mathbf{Z}_{t-1} + \boldsymbol{\phi}(\mathbf{x}_{t,a_t}) \boldsymbol{\phi}(\mathbf{x}_{t,a_t})^\top \in \mathbb{R}^{p \times p}, \quad \mathbf{b}_t = \mathbf{b}_{t-1} + r_{t,a_t} \boldsymbol{\phi}(\mathbf{x}_{t,a_t}) \in \mathbb{R}^p$$

Compute $\boldsymbol{\theta}_t = \mathbf{Z}_t^{-1} \mathbf{b}_t + \boldsymbol{\theta}_0 \in \mathbb{R}^p$

Construct $\mathcal{C}_t$ as

$$\mathcal{C}_t = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}_t - \boldsymbol{\theta}\|_{\mathbf{Z}_t} \le \gamma_t\}, \quad \text{where} \quad \gamma_t = \nu \sqrt{\log \frac{\det \mathbf{Z}_t}{\det \lambda \mathbf{I}} - 2 \log \delta} + \sqrt{\lambda} S \tag{4.13.2}$$

end for

CHAPTER 5

Conclusion and Future Directions

This dissertation established a theoretical foundation of the uncertainty-aware RL. Our primary aim is to develop a sample-efficient RL approach for MDPs with large state and action spaces. To achieve this, we propose an RL algorithm that incorporates both epistemic uncertainty and aleatoric uncertainty. By using function approximation, we demonstrate through theoretical analysis that our algorithm achieves a statistical complexity close to the minimax optimal level when learning the optimal policy by establishing matching upper and lower bounds on the regret. In our second objective, we focus on specific scenarios, namely the batch learning setting and the rare policy switch setting. We introduce epistemic uncertainty-aware RL algorithms with limited adaptivity for these settings. Our proposed algorithms exhibit a reduced and nearly optimal number of policy updates compared to the vanilla baseline algorithm. Additionally, we present a gradient-based method that effectively computes epistemic uncertainty. This estimation method is applied to the neural contextual bandit problem, resulting in a novel algorithm with a convergence guarantee.

The dissertation suggests several potential avenues for future research. The primary focus of our existing works has been on providing complexity results that are independent of specific problems, as well as demonstrating the optimality of proposed algorithms through the construction of challenging MDP instances. However, a significant and challenging task that remains open is the development of problem-specific algorithms that approach optimality in terms of complexity. Additionally, it would be intriguing to explore the construction of more efficient uncertainty estimates for applications with distinct problem structures, such as

language processing tasks. Such investigations would not only enhance our understanding of the fundamental properties of RL but also broaden the applicability of RL to a wider range of problem domains.

Bibliography

Abbasi-Yadkori, Y., Pál, D. and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*.

Abe, N., Biermann, A. W. and Long, P. M. (2003). Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica* **37** 263–293.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L. and Schapire, R. (2014). Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*. PMLR.

Agarwal, A., Kakade, S. and Yang, L. F. (2020). Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*.

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*.

Allen-Zhu, Z. and Li, Y. (2019). What can ResNet learn efficiently, going beyond kernels? In *Advances in Neural Information Processing Systems*.

Allen-Zhu, Z., Li, Y. and Song, Z. (2019). A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*.

Allesiardo, R., Féraud, R. and Bouneffouf, D. (2014). A neural networks committee for the contextual bandit problem. In *International Conference on Neural Information Processing*. Springer.

Altschuler, J. and Talwar, K. (2018). Online learning over a finite action set with limited switching. In *Conference On Learning Theory*. PMLR.

Arora, R., Dekel, O. and Tewari, A. (2012). Online bandit learning against an adaptive

adversary: from regret to policy regret. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. and Wang, R. (2019). On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*.

Audibert, J.-Y., Munos, R. and Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science* **410** 1876–1902.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* **3** 397–422.

Auer, P., Cesa-Bianchi, N., Freund, Y. and Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing* **32** 48–77.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M. and Yang, L. (2020). Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*. PMLR.

Azar, M. G., Munos, R. and Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning* **91** 325–349.

Azar, M. G., Osband, I. and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*. PMLR.

Azizzadenesheli, K., Brunskill, E. and Anandkumar, A. (2018). Efficient exploration through Bayesian deep Q-networks. In *2018 Information Theory and Applications Workshop (ITA)*. IEEE.

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series* **19** 357–367.

Bai, Y., Xie, T., Jiang, N. and Wang, Y.-X. (2019). Provably efficient q-learning with low switching cost. In *Advances in Neural Information Processing Systems*, vol. 32.

Beygelzimer, A. and Langford, J. (2009). The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L. and Schapire, R. E. (2011). Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning* **5** 1–122.

Bubeck, S., Munos, R., Stoltz, G. and Szepesvári, C. (2011). X-armed bandits. *Journal of Machine Learning Research* **12** 1655–1695.

Cai, Q., Yang, Z., Jin, C. and Wang, Z. (2020). Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*.

Cao, Y. and Gu, Q. (2019). Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*.

Cao, Y. and Gu, Q. (2020). Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Cesa-Bianchi, N., Dekel, O. and Shamir, O. (2013). Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems*, vol. 26.

Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*.

Chen, L., Yu, Q., Lawrence, H. and Karbasi, A. (2020). Minimax regret of switching-constrained online convex optimization: No phase transition. In *Advances in Neural Information Processing Systems*, vol. 33.

Chen, Z., Cao, Y., Zou, D. and Gu, Q. (2021). How much over-parameterization is sufficient to learn deep relu networks? In *International Conference on Learning Representations*.

Chu, W., Li, L., Reyzin, L. and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*.

Dani, V., Hayes, T. P. and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*.

Daniely, A. (2017). SGD learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*.

Dann, C. and Brunskill, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*.

Dann, C., Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J. and Schapire, R. E. (2018). On oracle-efficient pac rl with rich observations. In *Advances in neural information processing systems*.

Dekel, O., Ding, J., Koren, T. and Peres, Y. (2014). Bandits with switching costs: T 2/3 regret. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*.

Devroye, L., Lugosi, G. and Neu, G. (2015). Random-walk perturbations for online combinatorial optimization. *IEEE Transactions on Information Theory* **61** 4099–4106.

Du, S., Lee, J., Li, H., Wang, L. and Zhai, X. (2019a). Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*.

Du, S. S., Kakade, S. M., Lee, J. D., Lovett, S., Mahajan, G., Sun, W. and Wang, R. (2021). Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*. PMLR.

Du, S. S., Kakade, S. M., Wang, R. and Yang, L. F. (2019b). Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*.

Du, S. S., Zhai, X., Poczos, B. and Singh, A. (2019c). Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*, vol. 38. Siam.

Esfandiari, H., Karbasi, A., Mehrabian, A. and Mirrokni, V. (2021). Regret bounds for batched bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Faury, L., Abeille, M., Calauzènes, C. and Fercoq, O. (2020). Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*.

Féraud, R., Allesiardo, R., Urvoy, T. and Clérot, F. (2016). Random forest for the contextual bandit problem. In *Artificial Intelligence and Statistics*.

Filippi, S., Cappe, O., Garivier, A. and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*.

Foster, D. and Rakhlin, A. (2020). Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*. PMLR.

Foster, D. J., Rakhlin, A., Simchi-Levi, D. and Xu, Y. (2021). Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. In *Conference on Learning Theory*.

Freedman, D. (1975). On tail probabilities for martingales. *The Annals of Probability* **3** 100–118.

Gao, M., Xie, T., Du, S. S. and Yang, L. F. (2021). A provably efficient algorithm for linear markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494* .

Gao, Z., Han, Y., Ren, Z. and Zhou, Z. (2019). Batched multi-armed bandits problem. In *Advances in Neural Information Processing Systems*, vol. 32.

Geulen, S., Vöcking, B. and Winkler, M. (2010). Regret minimization for online buffering problems using the weighted majority algorithm. In *COLT*. Citeseer.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Han, Y., Zhou, Z., Zhou, Z., Blanchet, J., Glynn, P. W. and Ye, Y. (2020). Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321* .

Hanin, B. (2019). Universal function approximation by deep neural nets with bounded width and relu activations. *Mathematics* **7** 992.

Hanin, B. and Sellke, M. (2017). Approximating continuous functions by ReLU nets of minimal width. *arXiv preprint arXiv:1710.11278* .

He, J., Zhou, D. and Gu, Q. (2021a). Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*.

He, J., Zhou, D. and Gu, Q. (2021b). Nearly minimax optimal reinforcement learning for discounted mdps. *Advances in Neural Information Processing Systems* **34** 22288–22300.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 423–447.

Hora, S. C. (1996). Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety* **54** 217–223.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning* **110** 457–506.

Jacot, A., Gabriel, F. and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*.

Jaghargh, M. R. K., Krause, A., Lattanzi, S. and Vassilvtiskii, S. (2019). Consistent online optimization: Convex and submodular. In *The 22nd International Conference on Artificial Intelligence and Statistics*.

Jaksch, T., Ortner, R. and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* **11** 1563–1600.

Jia, Z., Yang, L., Szepesvari, C. and Wang, M. (2020). Model-based reinforcement learning with value-targeted regression. In *L4DC*.

Jiang, N. and Agarwal, A. (2018). Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J. and Schapire, R. E. (2017). Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

Jin, C., Allen-Zhu, Z., Bubeck, S. and Jordan, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*.

Jin, C., Yang, Z., Wang, Z. and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*.

Jun, K.-S., Bhargava, A., Nowak, R. D. and Willett, R. (2017). Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems 30 (NIPS)*.

Kakade, S. M., Shalev-Shwartz, S. and Tewari, A. (2008). Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th international conference on Machine learning*.

Kalai, A. and Vempala, S. (2005). Efficient algorithms for online decision problems. *Journal of Computer and System Sciences* **71** 291–307.

Kirschner, J. and Krause, A. (2018). Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*.

Kleinberg, R., Slivkins, A. and Upfal, E. (2008). Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*. ACM.

Krause, A. and Ong, C. S. (2011). Contextual Gaussian process bandit optimization. In *Advances in neural information processing systems*.

Kveton, B., Zaheer, M., Szepesvári, C., Li, L., Ghavamzadeh, M. and Boutilier, C. (2020). Randomized exploration in generalized linear bandits. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*.

Langford, J. and Zhang, T. (2008). The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems 20 (NIPS)*.

Lattimore, T., Crammer, K. and Szepesvári, C. (2015). Linear multi-resource allocation with semi-bandit feedback. In *Advances in Neural Information Processing Systems*.

Lattimore, T. and Hutter, M. (2012). PAC bounds for discounted MDPs. In *International Conference on Algorithmic Learning Theory*. Springer.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Lattimore, T., Szepesvári, C. and Weisz, G. (2020). Learning with good feature representations in bandits and in RL with a generative model. In *International Conference on Machine Learning*.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86** 2278–2324.

Li, L., Chu, W., Langford, J. and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*.

Li, L., Lu, Y. and Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

Li, Y. and Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*.

Li, Y., Wang, Y., Chen, X. and Zhou, Y. (2021). Tight regret bounds for infinite-armed linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Li, Y., Wang, Y. and Zhou, Y. (2019). Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*.

Liang, S. and Srikant, R. (2017). Why deep neural networks for function approximation? In *International Conference on Learning Representations*.

Lipton, Z., Li, X., Gao, J., Li, L., Ahmed, F. and Deng, L. (2018). BBQ-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Lu, Z., Pu, H., Wang, F., Hu, Z. and Wang, L. (2017). The expressive power of neural networks: A view from the width. In *Advances in neural information processing systems*.

Maurer, A. and Pontil, M. (2009). Empirical Bernstein bounds and sample variance penalization. In *COLT*.

Modi, A., Jiang, N., Tewari, A. and Singh, S. (2020). Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*.

Neu, G. and Pike-Burke, C. (2020). A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems* **33** 1392–1403.

OpenAI, R. (2023). Gpt-4 technical report. *arXiv* .

Perchet, V., Rigollet, P., Chassang, S., Snowberg, E. et al. (2016). Batched bandit problems. *The Annals of Statistics* **44** 660–681.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Riquelme, C., Tucker, G. and Snoek, J. (2018). Deep Bayesian bandits showdown. In *International Conference on Learning Representations*.

Ruan, Y., Yang, J. and Zhou, Y. (2021). Linear bandits with limited adaptivity and learning distributional optimal design. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*.

Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research* **35** 395–411.

Russo, D., Roy, B. V., Kazerouni, A., Osband, I. and Wen, Z. (2018). A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning* **11** 1–96.

Sidford, A., Wang, M., Wu, X., Yang, L. and Ye, Y. (2018). Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems 31*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. et al. (2016). Mastering the game of Go with deep neural networks and tree search. *nature* **529** 484.

Simchowitz, M. and Jamieson, K. G. (2019). Non-asymptotic gap-dependent regret bounds for tabular MDPs. In *Advances in Neural Information Processing Systems*.

Srinivas, N., Krause, A., Kakade, S. and Seeger, M. (2010). Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*. Omnipress.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A. and Langford, J. (2019). Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*.

Sutton, R. S. and Barto, A. G. (1998). *Introduction to reinforcement learning*, vol. 135. MIT Press, Cambridge.

Telgarsky, M. (2015). Representation benefits of deep feedforward networks. *arXiv preprint arXiv:1509.08101* .

Telgarsky, M. (2016). Benefits of depth in neural networks. *arXiv preprint arXiv:1602.04485* .

Tossou, A., Basu, D. and Dimitrakakis, C. (2019). Near-optimal optimistic reinforcement learning using empirical Bernstein inequalities. *arXiv preprint arXiv:1905.12425* .

Valko, M., Korda, N., Munos, R., Flaounas, I. and Cristianini, N. (2013). Finite-time analysis of kernelised contextual bandits. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*. UAI'13, AUAI Press, Arlington, Virginia, USA.

Wang, R., Du, S. S., Yang, L. and Kakade, S. (2020a). Is long horizon rl more difficult than short horizon rl? *Advances in Neural Information Processing Systems* **33**.

Wang, R., Salakhutdinov, R. R. and Yang, L. (2020b). Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems* **33**.

Wang, Y., Wang, R., Du, S. S. and Krishnamurthy, A. (2020c). Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*.

Weisz, G., Amortila, P. and Szepesvári, C. (2021). Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*. PMLR.

Wu, Y., György, A. and Szepesvári, C. (2015). Online learning with gaussian payoffs and side observations. In *Advances in Neural Information Processing Systems*.

Yang, L. and Wang, M. (2019). Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*.

Yang, L. and Wang, M. (2020). Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*. PMLR.

Yang, Z., Jin, C., Wang, Z., Wang, M. and Jordan, M. I. (2020). On function approximation in reinforcement learning: Optimism in the face of large state spaces. In *Advances in Neural Information Processing Systems*, vol. 33.

Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks* **94** 103–114.

Yarotsky, D. (2018). Optimal approximation of continuous functions by very deep relu networks. In *Conference on learning theory*. PMLR.

Zahavy, T. and Mannor, S. (2019). Deep neural linear bandits: Overcoming catastrophic forgetting through likelihood matching. *arXiv preprint arXiv:1901.08612* .

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M. and Lazaric, A. (2020a). Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*.

Zanette, A. and Brunskill, E. (2019). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*.

Zanette, A., Lazaric, A., Kochenderfer, M. J. and Brunskill, E. (2020b). Learning near optimal policies with low inherent Bellman error. In *International Conference on Machine Learning*.

Zhang, L., Yang, T., Jin, R., Xiao, Y. and Zhou, Z.-H. (2016). Online stochastic linear optimization under one-bit feedback. In *International Conference on Machine Learning*.

Zhang, Z. and Ji, X. (2019). Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems*.

Zhang, Z., Ji, X. and Du, S. S. (2021a). Is reinforcement learning more difficult than bandits? A near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*.

Zhang, Z., Zhou, Y. and Ji, X. (2020). Almost optimal model-free reinforcement learning via reference-advantage decomposition. In *Advances in Neural Information Processing Systems 33*.

Zhang, Z., Zhou, Y. and Ji, X. (2021b). Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. In *International Conference on Machine Learning*.

Zhou, D., Gu, Q. and Szepesvari, C. (2021a). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*. PMLR.

Zhou, D., He, J. and Gu, Q. (2021b). Provably efficient reinforcement learning for discounted MDPs with feature mapping. In *International Conference on Machine Learning*.

Zou, D., Cao, Y., Zhou, D. and Gu, Q. (2019). Stochastic gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning* .

Zou, D. and Gu, Q. (2019). An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*.