UNIVERSITY OF CALIFORNIA

Los Angeles

Learning in Safety-critical, Lifelong, and Multi-agent Systems: Bandits and RL Approaches

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Electrical and Computer Engineering

by

Sanae Amani Geshnigani

2023

ABSTRACT OF THE DISSERTATION

Learning in Safety-critical, Lifelong, and Multi-agent Systems: Bandits and RL Approaches

by

Sanae Amani Geshnigani
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2023
Professor Lin Yang, Chair

Sequential decision-making problems arise at every occasion that agents repeatedly interact with an unknown environment in an effort to maximize a certain notion of reward gained from interactions with this environment. Examples are abundant in online advertising, online gaming, robotics, deep learning, dynamic pricing, network routing, etc. In particular, multi-armed bandits (MAB) model the interaction between the agent and the unknown environment as follows. The agent repeatedly acts by pulling arms and after an arm is pulled, she receives a stochastic reward; the goal at the end of this process is to select actions that maximize the expected cumulative reward without knowledge of the arms' distributions. Albeit simple, this model is widely applicable. On the other hand, many sequential decision making occasions deal with more complicated environments modeled through Markov Decision Processes (MDPs) where the environment's status constantly changes as a result of taking actions and makes learning even more challenging. The field of reinforcement learning (RL) defines a principled foundation for this methodology, based on classical dynamic programming algorithms for solving MDPs.

Our research goal is to expand the applicability of bandit and RL algorithms to new application domains: specifically, safety-critical, lifelong and distributed physical systems, such as robotics, wireless networks, the power grid and medical trials.

One distinguishing feature of many of such "new" potential applications of bandits and RL is their *safety-critical* nature. Specifically, the algorithm's chosen policies must satisfy certain system constraints that if violated can lead to catastrophic results for the system. Importantly, the specifics of these constraints often change based on the interactions with the unknown environment; thus, they are often unknown themselves. This leads to the new challenge of balancing the goal of reward maximization with the restriction of playing policies that are safe. We modeled this problem through bandits and RL frameworks with linear reward and constraint structures. It turns out that even this seemingly simple safe linear bandit and RL formulations are more intricate than the original setting without safety constraints. In particular, simple variations of existing algorithms can be shown to be highly suboptimal. Using appropriate tools from high-dimensional probability and exploration-exploitation dilemma, we were able to design novel algorithms and to guarantee that they not only respect the safety constraints, but also have performance comparable to the setting without safety constraints.

Recently, there has been a surging interest in designing lifelong learning agents that can continuously learn to solve multiple sequential decision making problems in their lifetimes. This scenario is in particular motivated by building multi-purpose embodied intelligence, such as robots working in a weakly structured environment. Typically, curating all tasks beforehand for such problems is nearly infeasible, and the problems the agent is tasked with may be adaptively selected based on the agent's past behaviors. Consider a household robot as an example. Since each household is unique, it is difficult to anticipate upfront all scenarios the robot would encounter. In this direction, we theoretically study lifelong RL in a regret minimization setting, where the agent needs to solve a sequence of tasks using rewards in an unknown environment while balancing exploration and exploitation. Motivated by the embodied intelligence scenario, we suppose that tasks differ in rewards, but share the same state and action spaces and transition dynamics.

Another distinguishing feature of the envisioned applications of bandit algorithms is that interactions involve multiple *distributed* agents/learners (e.g., wireless/sensor networks). This calls for extensions of the traditional bandit setting to networked systems. In many

such systems, it is critical to maintain an efficient communication among the network while achieving a good performance in terms of accumulated reward, usually measured as network's regret. In view of this, for the problem of distributed contextual linear bandits, we prove a minimax lower bound on the communication cost of any distributed contextual linear bandit algorithm with stochastic contexts that is optimal in terms of regret. We further propose an algorithm whose regret is optimal and communication rate matches this lower bound, and therefore it is provably optimal in terms of *both regret and communication rate*.

The dissertation of Sanae Amani Geshnigani is approved.

Lieven Vandenberghe

Jonathan Chau-Yan Kao

Christina Panagio Fragouli

Lin Yang, Committee Chair

University of California, Los Angeles

2023

To my big-hearted parents, kindest brother, and caring partner
for their unconditional love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I want to express my profound gratitude to Professor Lin Yang, my advisor, whose exceptional kindness, generosity, and patience have been the pillars of my academic journey. Beyond his role as a mentor, I am truly thankful for the expansive academic freedom he graciously granted me, allowing me to explore and grow as a researcher. Professor Yang has played a crucial role in expanding my academic network, introducing me to other outstanding researchers and fostering connections that have enriched my research experience. Through his mentorship, I have had the privilege of engaging with a broader community of scholars, gaining insights and perspectives that have undoubtedly contributed to the depth of my work.

Words cannot express my gratitude and appreciation to Professor Christos Thrampoulidis, who guided me right from the start of my journey as a researcher. His patience and mentorship shaped the way I approach research. During the tough times, he was my rock of support. I've always looked up to his professionalism and unique, dedicated research style, trying my best to follow in his footsteps.

My sincerest gratitude is extended to all my other wonderful collaborators, Tor Lattimore, Andras Gyorgy, and Ching-An Cheng, for their patience and the enriching insights I've acquired through our collaborative efforts. Our work was partially supported by the following: NSF and DARPA grants, and Amazon science hub doctoral fellowship.

I'm also grateful to my committee members, Professor Christina Fragouli for her continuous support, insightful comments, and invaluable suggestions; Professor Lieven Vandenberghe, who provided me with a solid understanding of optimization fundamentals; and Professor Jonathan Kao, who significantly contributed to my knowledge in deep learning.

I want to express my love to all my dear friends including my long-distant friend Mahnaz, Bahareh, Matina, Atefeh, Golara, Sina, Daniel, and so many others - you were the ones who got me through this.

Finally, I want to express profound gratitude to my family, whose unwavering support

has made this journey possible. A special acknowledgment goes to my partner and closest ally, Borna, for being a continuous wellspring of love, kindness, and unwavering support. My heartfelt appreciation extends to my parents, Saman Dokht and Bahman, for their boundless, unconditional love and for being my sanctuary in challenging times, despite the geographical distance. I am immensely thankful to my brother, Ali, whose positive influence and presence have warmed my heart throughout all these years.

VITA

2021–2023	Ph.D. in Electrical and Computer Engineering, University of California, Los Angeles
2018–2020	M.Sc. in Electrical and Computer Engineering, University of California, Santa Barbara
2013–2018	B.Sc. in Electrical Engineering, Sharif University of Technology
2023	AI Research PhD Intern, PayPal Inc., San Jose, CA
2023	Amazon Science Hub Doctoral Fellowship, UCLA.
2022	Research Intern, Google, Mountain View, CA
2022	AI Research PhD Intern, PayPal Inc., San Jose, CA

PUBLICATIONS

S. Amani, T. Lattimore, A. Gyorgy, L. F. Yang, "Distributed Contextual Linear Bandits with Minimax Optimal Communication Cost", *International Conference on Machine Learning (ICML)*, 2023.

S. Amani, L. F. Yang, C.-A. Cheng, "Provably Efficient Lifelong Reinforcement Learning with Linear Representation", *International Conference on Learning Representations (ICLR)*, 2023.

S. Amani, L. F. Yang, "Doubly Pessimistic Algorithms for Strictly Safe Off-Policy Optimization", *Annual Conference on Information Sciences and Systems (CISS)*, 2022.

S. Amani, C. Thrampoulidis, "UCB-based Algorithms for Multinomial Logistic Regression

Bandits", *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

S. Amani, C. Thrampoulidis, L. F. Yang, "Safe Reinforcement Learning with Linear Function Approximation", *International Conference on Machine Learning (ICML)*, 2021.

S. Amani, M. Alizadeh, C. Thrampoulidis, "Regret Bound for Safe Gaussian Process Bandit Optimization", *IEEE International Symposium on Information Theory (ISIT)*, 2021.

S. Amani, C. Thrampoulidis, "Decentralized Multi-Agent Linear Bandits with Safety Constraints", *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

A. Moradipari, **S. Amani**, M. Alizadeh, C. Thrampoulidis, "Safe Linear Thompson Sampling with Side Information", *IEEE Transactions on Signal Processing*, 2021.

S. Amani, M. Alizadeh, C. Thrampoulidis, "Generalized Linear Bandits with Safety Constraints", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

S. Amani, M. Alizadeh, C. Thrampoulidis, "Linear Stochastic Bandits Under Safety Constraints", *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

CHAPTER 1

Introduction

Sequential decision making under uncertainty has received enormous attention in recent years. It specifically refers to occasions where an agent/learner repeatedly interacts with an unknown environment in an effort to maximize a certain notion of reward obtained throughout these interactions. The multiarmed bandit (MAB) problem and reinforcement learning (RL), in which a decision maker allocates a single resource by repeatedly choosing one among a set of competing alternative options, exemplify the explore vs. exploit trade-off, i.e., choosing between the most informative and the most rewarding actions trade-off. Sequential hypothesis testing concerns the speed-accuracy trade-off: deciding quickly versus reliably on a set of alternatives. One way of determining a bandit/RL algorithm is efficient in terms of this tradeoff, is to keep track of its *regret*, which is defined as the difference between accumulated reward by the algorithm and that of the best algorithm in the hindsight in expectation.

Our goal is to expand the applicability of bandit and RL algorithms to new application domains: specifically, safety-critical and distributed physical systems, such as robotics, wireless networks, the power grid and medical trials. In particular, it is desirable to theoretically study and design algorithms that are provably efficient and have descent regret performances. In the following sections of this chapter, we specify what "descent regret" mathematically refers to.

In the rest of this chapter we establish notation and a few necessary concepts and definitions used in chapters 2, 3, 4, 5, and 6. The results presented in chapters 2, 3, 4, 5 and 6 have been published as [11], [14], [17], [16], and [12], respectively.

1.1 Notation

Throughout this dissertation, we use lower-case letters for scalars, lower-case bold letters for vectors, and upper-case bold letters for matrices. The Euclidean norm of $\mathbf{x}$ is denoted by $\|\mathbf{x}\|_2$ and the spectral norm of a matrix $\mathbf{M}$ is denoted by $\|\mathbf{M}\|$. We denote the transpose of any column vector $\mathbf{x}$ by $\mathbf{x}^\top$. For any vectors $\mathbf{x}$ and $\mathbf{y}$, we use $\langle \mathbf{x}, \mathbf{y} \rangle$ to denote their inner product. We denote the Kronecker product by $\mathbf{A} \otimes \mathbf{B}$. Let $\mathbf{A}$ be a positive semi-definite $d \times d$ matrix and $\boldsymbol{\nu} \in \mathbb{R}^d$. The weighted 2-norm of $\boldsymbol{\nu}$ with respect to $\mathbf{A}$ is defined by $\|\boldsymbol{\nu}\|_{\mathbf{A}} = \sqrt{\boldsymbol{\nu}^\top \mathbf{A} \boldsymbol{\nu}}$. For square matrices $\mathbf{A}$ and $\mathbf{B}$, we use $\mathbf{A} \preceq \mathbf{B}$ to denote $\mathbf{B} - \mathbf{A}$ is positive semi-definite. We denote the minimum and maximum eigenvalue of $\mathbf{A}$ by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$. The maximum of two numbers α, β is denoted $\alpha \vee \beta$. For a real number α, we denote $\{\alpha\}^+ = \max\{\alpha, 0\}$. For a positive integer n, $[n]$ denotes the set $\{1, 2, \ldots, n\}$, while for positive integers $m \leq n$, $[m : n]$ denotes the set $\{m, m+1, \ldots, n\}$. We use $\mathbf{e}_i$ to denote the i-th standard basis vector. $I(X; Y)$ denotes the mutual information between two random variables X and Y. Finally, we use standard $\tilde{\mathcal{O}}$ notation for big-O notation that ignores logarithmic factors.

1.2 Stochastic Linear Bandit

In a stochastic linear bandit setting, at each round t, the agent is given a decision set $\mathcal{D}_t \subset \mathbb{R}^d$ [1]. At each round t, the agent chooses an action $\mathbf{x}_t \in \mathcal{D}_t$ and observes reward $y_t = \langle \boldsymbol{\theta}_*, \mathbf{x}_t \rangle + \eta_t$, where $\boldsymbol{\theta}_* \in \mathbb{R}^d$ is an unknown vector and η_t is random additive noise.

1.3 Cumulative Regret in Stochastic Linear Bandit

Let T be the total number of rounds. We define the cumulative regret of the entire network as:

$$R_T := \sum_{t=1}^{T} \langle \boldsymbol{\theta}_*, \mathbf{x}_* \rangle - \langle \boldsymbol{\theta}_*, \mathbf{x}_t \rangle. \tag{1.1}$$

[1] $\mathcal{D}_t$ may be fixed or changing throughout the learning horizon.

The optimal action $\mathbf{x}_*$ is defined with respect to $\mathcal{D}_t$ as $\arg\max_{\mathbf{x}\in\mathcal{D}_t}\langle\boldsymbol{\theta}_*, \mathbf{x}\rangle$. The goal is to minimize the cumulative regret and achieve regret that is sublinear in T.

1.4 Finite-horizon Markov decision process

A finite-horizon Markov decision process (MDP) is denoted by $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{S}$ is the state set, $\mathcal{A}$ is the action set, H is the length of each episode (horizon), $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ are the transition probabilities, and $r = \{r_h\}_{h=1}^H$ are the reward functions. For each time-step $h \in [H]$, $\mathbb{P}_h(s'|s, a)$ denotes the probability of transitioning to state s' upon playing action a at state s, and $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function. We consider the learning problem where $\mathcal{S}$ and $\mathcal{A}$ are known, while the transition probabilities $\mathbb{P}_h$ and rewards r_h are *unknown* to the agent and must be learned online. The agent interacts with its unknown environment described by M in episodes. In particular, at each episode k and time-step $h \in [H]$, the agent observes the state s_h^k, plays an action $a_h^k \in \mathcal{A}$, and observes a reward $r_h^k := r_h(s_h^k, a_h^k)$.

A deterministic policy is a function $\pi : \mathcal{S} \times [H] \to \mathcal{A}$, such that $\pi(s, h)$ is the action the policy π suggests the agent to play at time-step $h \in [H]$ and state $s \in \mathcal{S}$. A randomized policy $\pi : \mathcal{S} \times [H] \to \Delta_{\mathcal{A}}$ maps states and time-steps to distributions over actions such that $a \sim \pi(s, h)$ is the action the policy π suggests the agent to play at time-step $h \in [H]$ when being at state $s \in \mathcal{S}$.

For each $h \in [H]$, the cumulative expected reward obtained under a π during and after time-step h, known as the value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$, is defined by

$$V_h^\pi(s) := \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}\left(s_{h'}, \pi(s_{h'}, h')\right) \middle| s_h = s\right], \tag{1.2}$$

where the expectation is over the environment. We also define the state-action value action $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for a policy π at time-step $h \in [H]$ by

$$Q_h^\pi(s, a) := \mathbb{E}\left[\sum_{h'=h+1}^{H} r_{h'}\left(s_{h'}, \pi(s_{h'}, h')\right) \middle| s_h = s, a_h = a\right]. \tag{1.3}$$

To simplify the notation, for any function f, we denote $[\mathbb{P}_h f](s,a) := \mathbb{E}_{s' \sim \mathbb{P}_h(.|s,a)} f(s')$. Let π_* be the optimal policy such that $V_h^{\pi_*}(s) := V_h^*(s) = \sup_\pi V_h^\pi(s)$ for all $(s,h) \in \mathcal{S} \times [H]$. Thus, for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and $a \in \mathcal{A}$, the Bellman equations for a deterministic policy π and the optimal deterministic policy are:

$$Q_h^\pi(s,a) = r_h(s,a) + [\mathbb{P}_h V_{h+1}^\pi](s,a), \quad V_h^\pi(s) = Q_h^\pi(s, \pi(s,h)), \tag{1.4}$$

$$Q_h^*(s,a) = r_h(s,a) + [\mathbb{P}_h V_{h+1}^*](s,a), \quad V_h^*(s) = \max_{a \in \mathcal{A}} Q_h^*(s,a), \tag{1.5}$$

where $V_{H+1}^\pi(s) = V_{H+1}^*(s) = 0$.

The Bellman equations for a randomized policy π and the optimal randomized policy are:

$$\tilde{Q}_h^\pi(s,a) = r_h(s,a) + [\mathbb{P}_h \tilde{V}_{h+1}^\pi](s,a), \quad \tilde{V}_h^\pi(s) = \mathbb{E}_{a \sim \pi(s,h)} \left[\tilde{Q}_h^\pi(s,a) \right], \tag{1.6}$$

$$\tilde{Q}_h^*(s,a) = r_h(s,a) + [\mathbb{P}_h \tilde{V}_{h+1}^*](s,a), \quad \tilde{V}_h^*(s) = \max_\theta \mathbb{E}_{a \sim \theta} \left[\tilde{Q}_h^*(s,a) \right], \tag{1.7}$$

CHAPTER 2

Safety in Linear Stochastic Bandits

2.1 Introduction

The stochastic multi-armed bandit (MAB) problem is a sequential decision-making problem where, at each step of a T-period run, a learner plays one of k arms and observes a corresponding loss that is sampled independently from an underlying distribution with unknown parameters. The learner's goal is to minimize the pseudo-regret, i.e., the difference between the expected T-period loss incurred by the decision making algorithm and the optimal loss if the unknown parameters were given. The linear stochastic bandit problem generalizes MAB to the setting where each arm is associated with a feature vector x and the expected loss of each arm is equal to the inner product of its feature vector x and an unknown parameter vector $\boldsymbol{\mu}$. There are several variants of linear stochastic bandits that consider finite or infinite number of arms, as well as the case where the set of feature vectors changes over time. A detailed account of previous work in this area will be provided in Section 2.3.

Bandit algorithms have found many applications in systems that repeatedly deal with unknown stochastic environments (such as humans) and seek to optimize a long-term reward by simultaneously learning and exploiting the unknown environment (e.g., ad display optimization algorithms with unknown user preferences, path routing, ranking in search engines). They are also naturally relevant for many cyber-physical systems with humans in the loop (e.g., pricing end-use demand in societal-scale infrastructure systems such as power grids or transportation networks to minimize system costs given the limited number of user interactions possible). However, existing bandit heuristics might not be directly applicable in these latter cases. One critical reason is the existence of safety guarantees that have to be met at every single round.

For example, when managing demand to minimize costs in a power system, it is required that the operational constraints of the power grid are not violated in response to our actions (these can be formulated as linear constraints that depend on the demand). Thus, for such systems, it becomes important to develop new bandit algorithms that account for critical safety requirements.

Given the high level of uncertainty about the system parameters in the initial rounds, any such bandit algorithm will be initially highly constrained in terms of safe actions that can be chosen. However, as further samples are obtained and the algorithm becomes more confident about the value of the unknown parameters, it is intuitive that safe actions become easier to distinguish and it seems plausible that the effect of the system safety requirements on the growth of regret can be diminished.

In this chapter, we formulate a variant of linear stochastic bandits where at each round t, the learner's choice of arm should also satisfy a *safety constraint* that is dependent on the unknown parameter vector $\boldsymbol{\mu}$. While the formulation presented is certainly an abstraction of the complications that might arise in the systems discussed above, we believe that it is a natural first step towards understanding and evaluating the effect of safety constraints on the performance of bandit heuristics.

Specifically, we assume that the learner's goal is twofold: 1) Minimize the T-period cumulative pseudo-regret; 2) Ensure that a linear side constraint of the form $\boldsymbol{\mu}^\top \mathbf{B}\mathbf{x} \leq c$ is respected at every round during the T-period run of the algorithm, where B and c are known. See Section 2.2 for details. Given the learner's uncertainty about $\boldsymbol{\mu}$, the existence of this safety constraint effectively restricts the learner's choice of actions to what we will refer to as the *safe decision set* at each round t. To tackle this constraint, in Section 2.4, we present Safe-LUCB as a safe version of the standard linear UCB (LUCB) algorithm [40, 2, 104]. In Section 2.5, we provide general regret bounds that characterize the effect of safety constraints on regret. We show that the regret of the modified algorithm is dependent on the parameter $\Delta = c - \boldsymbol{\mu}^\top \mathbf{B}\mathbf{x}^*$, where $\mathbf{x}^*$ denotes the optimal safe action given $\boldsymbol{\mu}$. When $\Delta > 0$ and is known to the learner, we show that the regret of Safe-LUCB is $\tilde{\mathcal{O}}(\sqrt{T})$; thus, the effect of the system safety requirements on the growth of regret can be diminished (for large enough

T). In Section 2.6, we also present a heuristic modification of Safe-LUCB that empirically approaches the same regret without a-priori knowledge of the value of Δ. On the other hand, when $\Delta = 0$, the regret of Safe-LUCB is $\tilde{\mathcal{O}}(T^{2/3})$. Technical proofs and some further discussions are deferred to the appendix provided in the supplementary material.

2.1.1 Key Contributions

Bandit algorithms have various application in safety-critical systems, where it is important to respect the system constraints that rely on the bandit's unknown parameters at every round. In this chapter, we formulate a linear stochastic multi-armed bandit problem with safety constraints that depend (linearly) on an unknown parameter vector. As such, the learner is unable to identify all safe actions and must act conservatively in ensuring that her actions satisfy the safety constraint at all rounds (at least with high probability). For these bandits, we propose a new UCB-based algorithm called Safe-LUCB, which includes necessary modifications to respect safety constraints. The algorithm has two phases. During the pure exploration phase the learner chooses her actions at random from a restricted set of safe actions with the goal of learning a good approximation of the entire unknown safe set. Once this goal is achieved, the algorithm begins a safe exploration-exploitation phase where the learner gradually expands their estimate of the set of safe actions while controlling the growth of regret. We provide a general regret bound for the algorithm, as well as a problem dependent bound that is connected to the location of the optimal action within the safe set. We then propose a modified heuristic that exploits our problem dependent analysis to improve the regret.

2.2 Safe Linear Stochastic Bandit Problem

Cost model. The learner is given a convex compact decision set $\mathcal{D}_0 \subset \mathbb{R}^d$. At each round t, the learner chooses an action $\mathbf{x}_t \in \mathcal{D}_0$ which results in an observed loss ℓ_t that is linear on the unknown parameter $\boldsymbol{\mu}$ with additive random noise η_t, i.e., $\ell_t := c_t(\mathbf{x}_t) := \boldsymbol{\mu}^\top \mathbf{x}_t + \eta_t$.

Safety Constraint. The learning environment is subject to a side constraint that restricts the choice of actions by dividing $\mathcal{D}_0$ into a safe and an unsafe set. The learner is restricted to actions $\mathbf{x}_t$ from the *safe set* $\mathcal{D}^{\mathrm{s}}(\boldsymbol{\mu})$. As notation suggests, the safe set depends on the unknown parameter. Since $\boldsymbol{\mu}$ is unknown, the learner is unable to identify the safe set and must act conservatively in ensuring that actions $\mathbf{x}_t$ are feasible for all t. In this chapter, we assume that $\mathcal{D}^{\mathrm{s}}(\boldsymbol{\mu})$ is defined via a linear constraint

$$\boldsymbol{\mu}^\top \mathbf{B}\mathbf{x}_t \leq c, \tag{2.1}$$

which needs to be satisfied by $\mathbf{x}_t$ at all rounds t with high probability. Thus, $\mathcal{D}^{\mathrm{s}}(\boldsymbol{\mu})$ is defined as,

$$\mathcal{D}^{\mathrm{s}}(\boldsymbol{\mu}) := \{\mathbf{x} \in \mathcal{D}_0 \,:\, \boldsymbol{\mu}^\top \mathbf{B}\mathbf{x} \leq c\}. \tag{2.2}$$

The matrix $\mathbf{B} \in \mathbb{R}^{d\times d}$ and the positive constant $c > 0$ are known to the learner. However, after playing any action $\mathbf{x}_t$, the value $\boldsymbol{\mu}^\top \mathbf{B}\mathbf{x}_t$ is *not* observed by the learner. When clear from context, we drop the argument $\boldsymbol{\mu}$ in the definition of the safe set and simply refer to it as $\mathcal{D}^{\mathrm{s}}$.

Regret. Let T be the total number of rounds. If $\mathbf{x}_t,\ t \in [T]$ are the actions chosen, then the *cumulative pseudo-regret* ([22]) of the learner's algorithm for choosing the actions $\mathbf{x}_t$ is defined by $R_T = \sum_{t=1}^{T} \boldsymbol{\mu}^\top \mathbf{x}_t - \boldsymbol{\mu}^\top \mathbf{x}^*$, where $\mathbf{x}^*$ is the optimal *safe* action that minimizes the loss ℓ_t in expectation, i.e., $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathcal{D}^{\mathrm{s}}(\boldsymbol{\mu})} \boldsymbol{\mu}^\top \mathbf{x}$.

Goal. The goal of the learner is to keep R_T as small as possible. At the bare minimum, we require that the algorithm leads to $R_T/T \to 0$ (as T grows large). In contrast to existing linear stochastic bandit formulations, we require that the chosen actions $\mathbf{x}_t, t \in [T]$ are safe (i.e., belong in $\mathcal{D}^{\mathrm{s}}$ (2.2)) with high probability. For the rest of this chapter, we simply use regret to refer to the pseudo-regret R_T.

In Section 2.4.1 we place some further technical assumptions on $\mathcal{D}_0$ (bounded), on $\mathcal{D}^{\mathrm{s}}$ (non-empty), on $\boldsymbol{\mu}$ (bounded) and on the distribution of η_t (subgaussian).

2.3 Prior Work

Our algorithm relies on a modified version of the famous UCB algorithm known as UCB1, which was first developed by [23]. For linear stochastic bandits, the regret of the LUCB algorithm was analyzed by, e.g., [40, 2, 104, 105, 38] and it was shown that the regret grows at the rate of $\sqrt{T}\log(T)$. Extensions to generalized linear bandit models have also been considered by, e.g., [50, 83]. There are two different contexts where constraints have been applied to the stochastic MAB problem. The first line of work considers the MAB problem with global budget (a.k.a. knapsack) constraints where each arm is associated with a random resource consumption and the objective is to maximize the total reward before the learner runs out of resources, see, e.g., [25, 9, 136, 26]. The second line of work considers stage-wise safety for bandit problems in the context of ensuring that the algorithm's regret performance stays above a fixed percentage of the performance of a baseline strategy at every round during its run [68, 138]. In [68], which is most closely related to our setting, the authors study a variant of LUCB in which the chosen actions are constrained such that the *cumulative* reward remains *strictly* greater than $(1-\alpha)$ times a given baseline reward for all t. In both of the above mentioned lines of work, the constraint applies to the cumulative resource consumption (or reward) across the entire run of the algorithm. As such, the set of permitted actions at each round vary depending on the round and on the history of the algorithm. This is unlike our constraint, which is applied at each individual round, is deterministic, and does *not* depend on the history of past actions.

In a more general context, the concept of safe learning has received significant attention in recent years from different communities. Most existing work that consider mechanisms for *safe exploration* in unknown and stochastic environments are in reinforcement learning or control. However, the notion of safety has many diverse definitions in this literature. For example, [93] proposes an algorithm that allows safe exploration in Markov Decision Processes (MDP) in order to avoid fatal absorbing states that must never be visited during the exploration process. By considering constrained MDPs that are augmented with a set of auxiliary cost functions and replacing them with surrogates that are easy to estimate, [7] purposes a policy

search algorithm for constrained reinforcement learning with guarantees for near constraint satisfaction at each iteration. In the framework of global optimization or active data selection, [106, 28] assume that the underlying system is safety-critical and present active learning frameworks that use Gaussian Processes (GP) as non-parametric models to learn the safe decision set. More closely related to our setting, [114, 113] extend the application of UCB to *nonlinear* bandits with nonlinear constraints modeled through Gaussian processes (GPs). The algorithms in [114, 113] come with convergence guarantees, but *no* regret bounds as provided in our work. Regret guarantees imply convergence guarantees from an optimization perspective (see [110]), *but not the other way around*. Such approaches for safety-constrained optimization using GPs have shown great promise in robotics applications with safety constraints [95, 10].With a control theoretic point of view, [56] combines reachability analysis and machine learning for autonomously learning the dynamics of a target vehicle and [21] designs a learning-based MPC scheme that provides deterministic guarantees on robustness when the underlying system model is linear and has a known level of uncertainty. In a very recent related work [128], the authors propose and analyze a (safe) variant of the Frank-Wolfe algorithm to solve a smooth optimization problem with unknown linear constraints that are accessed by the learner via stochastic zeroth-order feedback. The main goal in [128] is to provide a convergence rate for more general convex objective, whereas we aim to provide *regret bounds* for a linear but otherwise unknown objective.

2.4 A Safe-LUCB Algorithm

Our proposed algorithm is a safe version of LUCB. As such, it relies on the well-known heuristic principle of *optimism in the face of uncertainty* (OFU). The algorithm constructs a confidence set $\mathcal{C}_t$ at each round t, within which the unknown parameter $\boldsymbol{\mu}$ lies with high probability. In the absence of any constraints, the learner chooses the most "favorable" environment $\boldsymbol{\mu}$ from the set $\mathcal{C}_t$ and plays the action $\mathbf{x}_t$ that minimizes the expected loss in that environment. However, the presence of the constraint (2.1) complicates the choice of the learner. To address this, we propose an algorithm called *safe linear upper confidence*

bound (Safe-LUCB), which attempts to minimize regret while making sure that the safety constraints (2.1) are satisfied. Safe-LUCB is summarized in Algorithm 1 and a detailed presentation follows in Sections 2.4.2 and 2.4.3, where we discuss the *pure-exploration* and *safe exploration-exploitation* phases of the algorithm, respectively. Before these, in Section 2.4.1 we introduce the necessary conditions under which our proposed algorithm operates and achieves good regret bounds as will be shown in Section 2.5.

2.4.1 Model Assumptions

Let $\mathcal{F}_t = \sigma(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t+1}, \eta_1, \eta_2, \ldots, \eta_t)$ be the σ-algebra (or, history) at round t. We make the following standard assumptions on the noise distribution, on the parameter $\boldsymbol{\mu}$ and on the actions.

Assumption 1 (Subgaussian noise)**.** *For all t, η_t is conditionally zero-mean R-sub-Gaussian for fixed constant $R \geq 0$, i.e., $\mathbb{E}[\eta_t \,|\, x_{1:t}, \eta_{1:t-1}] = 0$ and $\mathbb{E}[e^{\lambda \eta_t} \,|\, \mathcal{F}_{t-1}] \leq \exp(\lambda^2 R^2/2), \quad \forall \lambda \in \mathbb{R}$.*

Assumption 2 (Boundedness)**.** *There exist positive constants S, L such that $\|\boldsymbol{\mu}\|_2 \leq S$ and $\|\mathbf{x}\|_2 \leq L, \forall \mathbf{x} \in \mathcal{D}_0$. Also, $\boldsymbol{\mu}^\top \mathbf{x} \in [-1, 1], \forall \mathbf{x} \in \mathcal{D}_0$.*

In order to avoid trivialities, we also make the following assumption. This, together with the assumption that $C > 0$ in (2.1), guarantee that the safe set $\mathcal{D}^{\mathrm{s}}(\boldsymbol{\mu})$ is non-empty (for every $\boldsymbol{\mu}$).

Assumption 3 (Non-empty safe set)**.** *The decision set $\mathcal{D}_0$ is a convex body in $\mathbb{R}^d$ that contains the origin in its interior.*

2.4.2 Pure Exploration Phase

The pure exploration phase of the algorithm runs for rounds $t \in [T']$, where T' is passed as input to the algorithm. In Section 2.5, we will show how to appropriately choose its value to guarantee that the cumulative regret is controlled. During this phase, the algorithm selects

Algorithm 1 Safe-LUCB

Pure exploration phase:

for $t = 1, 2, \ldots, T'$ **do**

  Randomly choose $\mathbf{x}_t \in \mathcal{D}^w$ (defined in (2.3)) and observe loss $\ell_t = c_t(\mathbf{x}_t)$.

end for

Safe exploration-exploitation phase:

for $t = T' + 1, 2, \ldots, T$ **do**

  Set $\mathbf{A}_t = \lambda I + \sum_{\tau=1}^{t-1} \mathbf{x}_\tau \mathbf{x}_\tau^\top$ and compute $\hat{\boldsymbol{\mu}}_t = \mathbf{A}_t^{-1} \sum_{\tau=1}^{t-1} \ell_\tau \mathbf{x}_\tau$

  $\mathcal{C}_t = \{\boldsymbol{\nu} \in \mathbb{R}^d : \|\boldsymbol{\nu} - \hat{\boldsymbol{\mu}}_t\|_{\mathbf{A}_t} \leq \beta_t\}$ and β_t chosen as in (2.7)

  $\mathcal{D}_t^{\mathrm{s}} = \{\mathbf{x} \in \mathcal{D}_0 : \boldsymbol{\nu}^\top \mathbf{B}\mathbf{x} \leq c, \forall \boldsymbol{\nu} \in \mathcal{C}_t\}$

  $\mathbf{x}_t = \arg\min_{\mathbf{x} \in \mathcal{D}_t^{\mathrm{s}}} \min_{v \in \mathcal{C}_t} \boldsymbol{\nu}^\top \mathbf{x}$

  Choose $\mathbf{x}_t$ and observe loss $\ell_t = c_t(\mathbf{x}_t)$.

end for

random actions from a safe subset $\mathcal{D}^w \subset \mathcal{D}_0$ that we define next. For every chosen action $\mathbf{x}_t$, we observe a loss ℓ_t. The collected action-loss pairs $(\mathbf{x}_t, \ell_t)$ over the T' rounds are used in the second phase to obtain a good estimate of $\boldsymbol{\mu}$. We will see in Section 2.4.3 that this is important since the quality of the estimate of $\boldsymbol{\mu}$ determines our belief of which actions are safe. Now, let us define the safe subset $\mathcal{D}^w$.

The safe set $\mathcal{D}^{\mathrm{s}}$ is unknown to the learner (since $\boldsymbol{\mu}$ is unknown). However, it can be deduced from the constraint (2.1) and the boundedness Assumption 2 on $\boldsymbol{\mu}$, that the following subset $\mathcal{D}^w \subset \mathcal{D}_0$ is safe:

$$\mathcal{D}^w := \{\mathbf{x} \in \mathcal{D}_0 : \max_{\|v\|_2 \leq S} \boldsymbol{\nu}^\top \mathbf{B}\mathbf{x} \leq c\} = \{\mathbf{x} \in \mathcal{D}_0 : \|\mathbf{B}\mathbf{x}\|_2 \leq c/S\}. \tag{2.3}$$

Note that the set $\mathcal{D}^w$ is only a conservative (inner) approximation of $\mathcal{D}^{\mathrm{s}}$, but this is inevitable, since the learner has not yet collected enough information on the unknown parameter $\boldsymbol{\mu}$.

In order to make the choice of random actions $\mathbf{x}_t, t \in [T']$ concrete, let $X \sim \mathrm{Unif}(\mathcal{D}^w)$ be a d-dimensional random vector uniformly distributed in $\mathcal{D}^w$ according to the probability measure given by the normalized volume in $\mathcal{D}^w$ (recall that $\mathcal{D}^w$ is a convex body by Assumption 3). During rounds $t \in [T']$, Safe-LUCB chooses safe IID actions $\mathbf{x}_t \overset{\mathrm{iid}}{\sim} X$. For future reference, we

denote the covariance matrix of X by $\Sigma = \mathbb{E}[XX^\top]$ and its minimum eigenvalue by

$$\lambda_- := \lambda_{\min}(\Sigma) > 0. \tag{2.4}$$

Remark 1. *Since $\mathcal{D}_0$ is compact with zero in its interior, we can always find $0 < \epsilon \leq C/S$ such that*

$$\widetilde{\mathcal{D}^w} := \{\mathbf{x} \in \mathbb{R}^d \,|\, \|\mathbf{B}\mathbf{x}\|_2 = \epsilon\} \subset \mathcal{D}^w. \tag{2.5}$$

Thus, an effective way to choose (random) actions $\mathbf{x}_t$ during the safe-exploration phase for which an explicit expression for λ_- is easily derived, is as follows. For simplicity, we assume B is invertible. Let ϵ be the largest value $0 < \epsilon \leq c/S$ such that (2.5) holds. Then, generate samples $\mathbf{x}_t \sim$ Unif$(\widetilde{\mathcal{D}^w}), t = 1, \ldots, T'$, by choosing $\mathbf{x}_t = \epsilon \mathbf{B}^{-1}\mathbf{z}_t$, where $\mathbf{z}_t$ are i.i.d samples on the unit sphere $\mathcal{S}^{d-1}$. Clearly, $\mathbb{E}[z_t z_t^\top] = \frac{1}{d}\mathbf{I}$. Thus, $\Sigma := \mathbb{E}[\mathbf{x}_t \mathbf{x}_t^\top] = \frac{\epsilon^2}{d}\left(\mathbf{B}^\top \mathbf{B}\right)^{-1}$, from which it follows that $\lambda_- := \lambda_{min}(\Sigma) = \frac{\epsilon}{d\,\lambda_{max}\left(\mathbf{B}^\top \mathbf{B}\right)} = \frac{\epsilon^2}{d\|\mathbf{B}\|^2}$.

2.4.3 Safe Exploration-Exploitation Phase

We implement the OFU principle *while respecting the safety constraints.* First, at each $t = T' + 1, T' + 2 \ldots, T$, the algorithm uses the previous action-observation pairs and obtains a λ-regularized least-squares estimate $\hat{\boldsymbol{\mu}}_t$ of $\boldsymbol{\mu}$ with regularization parameter $\lambda > 0$ as follows:

$$\hat{\boldsymbol{\mu}}_t = \mathbf{A}_t^{-1} \sum_{\tau=1}^{t-1} \ell_\tau \mathbf{x}_\tau, \text{ where } \mathbf{A}_t = \lambda \mathbf{I} + \sum_{\tau=1}^{t-1} \mathbf{x}_\tau \mathbf{x}_\tau^\top.$$

Then, based on $\hat{\boldsymbol{\mu}}_t$ the algorithm builds a *confidence set*

$$\mathcal{C}_t := \{\boldsymbol{\nu} \in \mathbb{R}^d : \|\boldsymbol{\nu} - \hat{\boldsymbol{\mu}}_t\|_{\mathbf{A}_t} \leq \beta_t\}, \tag{2.6}$$

where, β_t is chosen according to Theorem 1 below ([2]) to guarantee that $\boldsymbol{\mu} \in \mathcal{C}_t$ with high probability.

Theorem 1 (Confidence region, [2])**.** *Let Assumptions 1 and 2 hold. Fix any $\delta \in (0, 1)$ and let β_t in (2.6) be chosen as follows,*

$$\beta_t = R\sqrt{d \log\left(\frac{1 + (t-1)L^2/\lambda}{\delta}\right)} + \lambda^{1/2} S, \qquad \textit{for all} \quad t > 0. \tag{2.7}$$

Then, with probability at least $1-\delta$*, for all* $t>0$*, it holds that* $\boldsymbol{\mu}\in\mathcal{C}_t$*.*

The remaining steps of the algorithm also build on existing principles of UCB algorithms. However, here we introduce necessary modifications to account for the safety constraint (2.1). Specifically, we choose the actions with the following two principles.

Caution in the face of constraint violation. At each round t, the algorithm performs conservatively, to ensure that the constraint (2.1) is satisfied for the chosen action $\mathbf{x}_t$. As such, at the beginning of each round $t=T'+1,\dots,T$, Safe-LUCB forms the so-called *safe decision set* denoted as $\mathcal{D}_t^{\mathrm{s}}$:

$$\mathcal{D}_t^{\mathrm{s}}=\{\mathbf{x}\in\mathcal{D}_0:\boldsymbol{\nu}^\top\mathbf{B}\mathbf{x}\le c,\forall v\in\mathcal{C}_t\}. \tag{2.8}$$

Recall from Theorem 1 that $\boldsymbol{\mu}\in\mathcal{C}_t$ with high probability. Thus, $\mathcal{D}_t^{\mathrm{s}}$ is guaranteed to be a set of safe actions that satisfy (2.1) with the same probability. On the other hand, note that $\mathcal{D}_t^{\mathrm{s}}$ is still a conservative inner approximation of $\mathcal{D}^{\mathrm{s}}(\boldsymbol{\mu})$ (actions in it are safe for *all* parameter vectors in $\mathcal{C}_t$, not only for the true $\boldsymbol{\mu}$). This (unavoidable) conservative definition of safe decision sets could contribute to the growth of the regret. This is further studied in Section 2.5.

Optimism in the face of uncertainty in cost. After choosing safe actions randomly at rounds $1,\dots,T'$, the algorithm creates the safe decision set $\mathcal{D}_t^{\mathrm{s}}$ at all rounds $t\ge T'+1$, and chooses an action $\mathbf{x}_t$ based on the OFU principle. Specifically, a pair $(\mathbf{x}_t,\tilde{\boldsymbol{\mu}}_t)$ is chosen such that

$$\tilde{\boldsymbol{\mu}}_t^\top\mathbf{x}_t=\min_{\mathbf{x}\in\mathcal{D}_t^{\mathrm{s}},v\in\mathcal{C}_t}\boldsymbol{\nu}^\top\mathbf{x}. \tag{2.9}$$

2.5 Regret Analysis of Safe-LUCB

2.5.1 The Regret of Safety

In the safe linear bandit problem, the safe set $\mathcal{D}^{\mathrm{s}}$ is not known, since $\boldsymbol{\mu}$ is unknown. Therefore, at each round, the learner chooses actions from a conservative inner approximation of $\mathcal{D}^{\mathrm{s}}$.

Intuitively, the better this approximation, the more likely that the optimistic actions of Safe-LUCB lead to good cumulant regret, ideally of the same order as that of LUCB in the original linear bandit setting.

A key difference in the analysis of Safe-LUCB compared to the classical LUCB is that $\mathbf{x}^*$ may not lie within the estimated safe set $\mathcal{D}_t^{\mathrm{s}}$ at each round. To see what changes, consider the standard decomposition of the instantaneous regret $r_t,\ t = T' + 1, \ldots, T$ in two terms as follows (e.g., [40, 2]):

$$r_t := \boldsymbol{\mu}^\top \mathbf{x}_t - \boldsymbol{\mu}^\top \mathbf{x}^* = \underbrace{\boldsymbol{\mu}^\top \mathbf{x}_t - \tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}_t}_{\text{Term I}} + \underbrace{\tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}_t - \boldsymbol{\mu}^\top \mathbf{x}^*}_{\text{Term II}}, \tag{2.10}$$

where, $(\tilde{\boldsymbol{\mu}}_t, \mathbf{x}_t)$ is the optimistic pair, i.e. the solution to the minimization in Step 10 of Algorithm 1. On the one hand, controlling Term I, is more or less standard and closely follows previous such bounds on UCB-type algorithms (e.g., [2]); see Appendix A.2.2 for details. On the other hand, controlling Term II, which we call *the regret of safety* is more delicate. This complication lies at the heart of the new formulation with additional safety constraints. When safety constraints are absent, classical LUCB guarantees that Term II is non-positive. Unfortunately, this is *not* the case here: $\mathbf{x}^*$ does *not* necessarily belong to $\mathcal{D}_t^{\mathrm{s}}$ in (2.8), thus Term II can be positive. This extra regret of safety is the price paid by Safe-LUCB for choosing safe actions at each round. Our main contribution towards establishing regret guarantees is upper bounding Term II. We show in Section 2.5.2 that the pure-exploration phase is critical in this direction.

2.5.2 Learning the Safe Set

The challenge in controlling the regret of safety is that, in general, $\mathcal{D}_t^{\mathrm{s}} \neq \mathcal{D}^{\mathrm{s}}$. At a high level, we proceed as follows (see Appendix A.2.3 for details). First, we relate Term II with a certain notion of "distance" in the direction of $\mathbf{x}^*$ between the estimated set $\mathcal{D}_t^{\mathrm{s}}$ at rounds $t = T' + 1, \ldots, T$ and the true safe set $\mathcal{D}^{\mathrm{s}}$. Next, we show that this "distance" term can be controlled by appropriately lower bounding the minimum eigenvalue $\lambda_{\min}(\mathbf{A}_t)$ of the Gram matrix $\mathbf{A}_t$. Due to the interdependency of the actions $\mathbf{x}_t$, it is difficult to directly establish such a lower bound for each round t. Instead, we use that $\lambda_{\min}(\mathbf{A}_t) \geq \lambda_{\min}(\mathbf{A}_{T'+1}),\ t \geq T' + 1$

and we are able to bound $\lambda_{\min}(\mathbf{A}_{T'+1})$ thanks to the pure exploration phase of Safe-LUCB . Hence, the pure exploration phase guarantees that $\mathcal{D}_t^{\mathrm{s}}$ is a sufficiently good approximation to the true $\mathcal{D}^{\mathrm{s}}$ once the exploration-exploitation phase begins.

Lemma 1. *Let $\mathbf{A}_{T'+1} = \lambda \mathbf{I} + \sum_{t=1}^{T'} \mathbf{x}_t \mathbf{x}_t^\top$ be the Gram matrix corresponding to the first T' actions of Safe-LUCB (pure-exploration phase). Recall the definition of λ_- in (2.4). Then, for any $\delta \in (0,1)$, it holds with probability at least $1-\delta$,*

$$\lambda_{min}(\mathbf{A}_{T'+1}) \geq \lambda + \frac{\lambda_- T'}{2}, \tag{2.11}$$

provided that $T' \geq t_\delta := \frac{8L^2}{\lambda_-} \log(\frac{d}{\delta})$.

The proof of the lemma and technical details relating the result to a desired bound on Term II are deferred to Appendixes A.1 and A.2.3, respectively.

2.5.3 Problem Dependent Upper Bound

In this section, we present a problem-dependent upper bound on the regret of Safe-LUCB in terms of the following critical parameter, which we call the *safety gap*:

$$\Delta := c - \boldsymbol{\mu}^\top \mathbf{B} \mathbf{x}^*. \tag{2.12}$$

Note that $\Delta \geq 0$. In this section, we assume that Δ is known to the learner. The next lemma shows that if $\Delta > 0$ [1] , then choosing $T' = \mathcal{O}(\log T)$ guarantees that $\mathbf{x}^* \in \mathcal{D}_t^{\mathrm{s}}$ for all $t = T'+1, \ldots, T$.

Lemma 2 ($\mathbf{x}^* \in \mathcal{D}_t^{\mathrm{s}}$)**.** *Let Assumptions 1, 2 and 3 hold. Fix any $\delta \in (0,1)$ and assume a positive safety gap $\Delta > 0$. Initialize Safe-LUCB with (recall the definition of t_δ in Lemma 1)*

$$T' \geq T_\Delta := \left(\frac{8L^2 \|B\|^2 \beta_T^2}{\lambda_- \, \Delta^2} - \frac{2\lambda}{\lambda_-} \right) \vee t_\delta. \tag{2.13}$$

Then, with probability at least $1-\delta$, for all $t = T'+1, \ldots, T$ it holds that $\mathbf{x}^* \in \mathcal{D}_t^s$.

[1] We remark that the case $\Delta > 0$ studied here is somewhat reminiscent of the assumption $\alpha r_\ell > 0$ in [68].

In light of our discussion in Sections 2.5.1 and 2.5.2, once we have established that $\mathbf{x}^* \in \mathcal{D}_t^{\mathrm{s}}$ for $t = T' + 1, \ldots, T$, the regret of safety becomes nonpositive and we can show that the algorithm performs just like classical LUCB during the exploration-exploitation phase [2]. This is formalized in Theorem 2 showing that when $\Delta > 0$ (and is known), then the regret of Safe-LUCB is $\tilde{\mathcal{O}}(\sqrt{T})$.

Theorem 2 (Problem-dependent bound; $\Delta > 0$)**.** *Let the same assumptions as in Lemma 2 hold. Initialize Safe-LUCB with $T' \geq T_\Delta$ specified in* (2.13)*. Then, for $T \geq T'$, with probability at least $1 - 2\delta$, the cumulative regret of Safe-LUCB satisfies*

$$R_T \leq 2T' + 2\beta_T \sqrt{2d\,(T - T')\,\log\left(\frac{2TL^2}{d(\lambda\_T' + 2\lambda)}\right)}\,. \tag{2.14}$$

Specifically, choosing $T' = T_\Delta$ guarantees cumulant regret $\mathcal{O}(T^{1/2}\log T)$.

The bound in (2.14) is a contribution of two terms. The first one is a trivial bound on the regret of the exploration-only phase of Safe-LUCB and is proportional to its duration T'. Thanks to Lemma 2 the duration of the exploration phase is limited to T_Δ rounds and T_Δ is (at most) logarithmic in the total number of rounds T. Thus, the first summand in (2.14) contributes only $\mathcal{O}(\log T)$ in the total regret. Note, however, that T_Δ grows larger as the normalized safety gap $\Delta/\|\mathbf{B}\|$ becomes smaller. The second summand in (2.14) contributes $\mathcal{O}(T^{1/2}\log T)$ and bounds the cumulant regret of the exploration-exploitation phase, which takes the bulk of the algorithm. More specifically, it bounds the contribution of Term I in (2.10) since the Term II is zeroed out once $\mathbf{x}^* \in \mathcal{D}_t^{\mathrm{s}}$ thanks to Lemma 2. Finally, note that Theorem 2 requires the total number of rounds T to be large enough for the desired regret performance. This is the price paid for the extra safety constraints compared to the performance of the classical LUCB in the original linear bandit setting. We remark that existing lower bounds for the simpler problem without safety constraints (e.g. [104, 40]), show that the regret $\tilde{\mathcal{O}}(\sqrt{Td})$ of Theorem 2 cannot be improved modulo logarithmic factors. The proofs of Lemma 2 and Theorem 2 are in Appendix A.2.

[2] Our simulation results in Appendix A.6 emphasize the critical role of a sufficiently long pure exploration phase by Safe-LUCB as suggested by Lemma 2. Specifically, Figure 2.1b depicts an instance where *no* exploration leads to significantly worse order of regret.

2.5.4 General Upper Bound

We now extend the results of Section 2.5.3 to instances where the safety gap is zero, i.e. $\Delta = 0$. In this case, we cannot guarantee an exploration phase that results in $\mathbf{x}^* \in \mathcal{D}_t^s, t > T'$ in a reasonable time length T'. Thus, the regret of safety is not necessarily non-positive and it is unclear whether a sub-linear cumulant regret is possible.

Theorem 3 shows that Safe-LUCB achieves regret $\tilde{\mathcal{O}}(T^{2/3})$ when $\Delta = 0$. Note that this (worst-case) bound is also applicable when the safety gap is unknown to the learner. While it is significantly worse than the performance guaranteed by Theorem 2, it proves that Safe-LUCB always leads to $R_T/T \to 0$ as T grows large. The proof is deferred to Appendix A.2.

Theorem 3 (General bound: worst-case)**.** *Suppose Assumptions 1, 2 and 3 hold. Fix any $\delta \in (0, 0.5)$. Initialize Safe-LUCB with $T' \geq t_\delta$ specified in Lemma 1. Then, with probability at least $1 - 2\delta$ the cumulative regret R_T of Safe-LUCB for $T \geq T'$ satisfies*

$$R_T \leq 2T' + 2\beta_T \sqrt{2d(T - T') \log\left(\frac{2TL^2}{d(\lambda_- T' + 2\lambda)}\right)} + \frac{2\sqrt{2}\|B\| L \beta_T (T - T')}{c\sqrt{\lambda_- T' + 2\lambda}}. \tag{2.15}$$

Specifically, choosing $T' = T_0 := \left(\frac{\|\mathbf{B}\| L \beta_T T}{c\sqrt{2\lambda_-}}\right)^{\frac{2}{3}} \vee t_\delta$, guarantees regret $\mathcal{O}(T^{2/3} \log T)$.

Compared to Theorem 2, the bound in (2.15) is now comprised of three terms. The first one captures again the exploration-only phase and is linear in its duration T'. However, note that T' is now $\mathcal{O}(T^{2/3} \log T)$, i.e., of the same order as the total bound. The second term bounds the total contribution of Term I of the exploration-exploitation phase. As usual, its order is $\tilde{\mathcal{O}}(T^{1/2})$. Finally, the additional third term bounds the regret of safety and is of the same order as that of the first term.

2.6 Unknown Safety Gap

In Section 2.5.3 we showed that when the safety gap $\Delta > 0$, then Safe-LUCB achieves good regret performance $\tilde{\mathcal{O}}(\sqrt{T})$. However, this requires that the value of Δ, or at least a

(non-trivial) lower bound on it, be known to the learner so that T' is initialized appropriately according to Lemma 2. This requirement might be restrictive in certain applications. When that is the case, one option is to run Safe-LUCB with a choice of T' as suggested by Theorem 3, but this could result in an unnecessarily long pure exploration period (during which regret grows linearly). Here, we present an alternative. Specifically, we propose a variation of Safe-LUCB refered to as *generalized safe linear upper confidence bound* (GSLUCB). The key idea behind GSLUCB is to build a lower confidence bound Δ_t for the safety gap Δ and calculate the length of the pure exploration phase associated with Δ_t, denoted as T'_t. This allows the learner to stop the pure exploration phase at round t such that condition $t \leq T'_{t-1}$ has been met. While we do not provide a separate regret analysis for GSLUCB, it is clear that the worst case regret performance would match that of Safe-LUCB with $\Delta = 0$. However, our numerical experiment highlights the improvements that GSLUCB can provide for the cases where $\Delta \neq 0$. We give a full explanation of GSLUCB, including how we calculate the lower confidence bound Δ_t, in Appendix A.5.

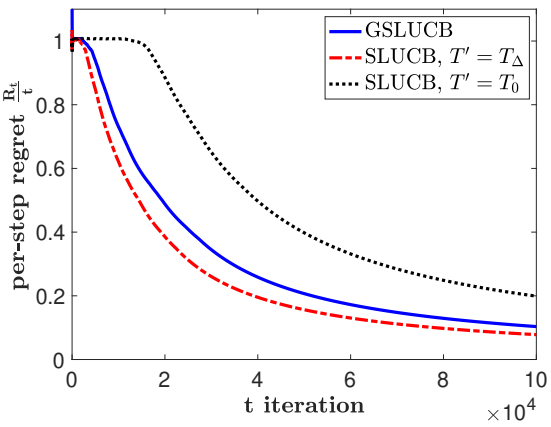
Figure 2.1a compares the average per-step regret of 1) Safe-LUCB with knowledge of Δ; 2) Safe-LUCB without knowledge of Δ (hence, assuming $\Delta = 0$); 3) GSLUCB without knowledge of Δ, in a simplified setting of K-armed linear bandits with strictly positive safety gap (see Appendix A.3). The details on the parameters of the simulations are deferred to Appendix A.6.

2.7 Future Directions and Summary

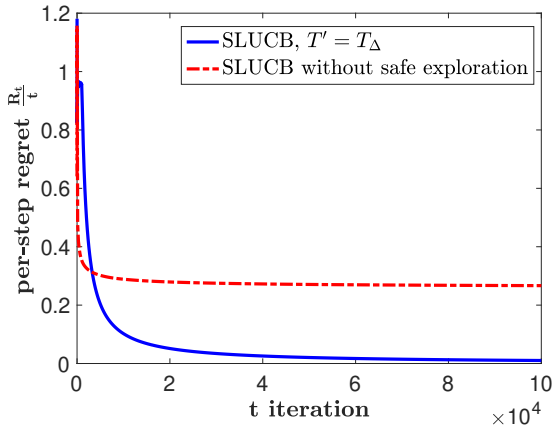
We have formulated a linear stochastic bandit problem with safety constraints that depend linearly on the unknown problem parameter $\boldsymbol{\mu}$. While simplified, the model captures the additional complexity introduced in the problem by the requirement that chosen actions belong to an unknown safe set. As such, it allows us to quantify tradeoffs between learning the safe set and minimizing the regret. Specifically, we propose Safe-LUCB which is comprised of two phases: (i) a pure-exploration phase that speeds up learning the safe set; (ii) a safe exploration-exploitation phase that optimizes minimizing the regret. Our analysis suggests

Algorithm 2 GSLUCB

1: **Pure exploration phase:**
2: $t \leftarrow 1$, $T'_0 = T_0$
3: **while** $\left(t \leq \min\left(T'_{t-1}, T_0\right)\right)$ **do**
4: Randomly choose $x_t \in \mathcal{D}^w$ and observe loss $\ell_t = c_t(x_t)$.
5: Δ_t = Lower confidence bound on Δ at round t
6: **if** $\Delta_t > 0$ **then**
7: $T'_t = T_{\Delta_t}$
8: **else**
9: $T'_t = T_0$
10: **end if**
11: $t \leftarrow t + 1$
12: **end while**
13: **Safe exploration exploitation phase:**
14: Lines 6 - 12 of Safe-LUCB for all remaining rounds.



(a) Average per-step regret of Safe-LUCB and GSLUCB with a decision set of K arms.



(b) Per-step regret of Safe-LUCB with and without pure exploration phase.

Figure 2.1: Simulation of per-step regret.

that the safety gap Δ plays a critical role. When $\Delta > 0$ we show how to achieve regret $\tilde{\mathcal{O}}(\sqrt{T})$ as in the classical linear bandit setting. However, when $\Delta = 0$, the regret of Safe-LUCB is $\tilde{\mathcal{O}}(T^{2/3})$. It is an interesting open problem to establish lower bounds for an arbitrary policy that accounts for the safety constraints. Our analysis of Safe-LUCB suggests that $\Delta = 0$ is a worst-case scenario, but it remains open whether the $\tilde{\mathcal{O}}(T^{2/3})$ regret bound can be improved in that case. Natural extensions of the problem setting to multiple constraints and generalized linear bandits (possibly with generalized linear constraints) might also be of interest.

CHAPTER 3

Safe Reinforcement Learning with Linear Function Approximation

3.1 Introduction

Reinforcement Learning (RL) is the study of an agent trying to maximize its expected cumulative reward by interacting with an unknown environment over time [117]. In most classical RL algorithms, agents aim to maximize a long term gain by exploring all possible actions. However, freely exploring all actions may be harmful in many real-world systems where playing even one unsafe action may lead to catastrophic results. Thus, safety in RL has become a serious issue that restricts the applicability of RL algorithms to many real-world systems. For example, in a self-driving car, it is critical to explore those policies that avoid crash and damage to the car, people and property. Switching cost limitations in medical applications [27] and legal restrictions in financial managements [3] are other examples of safety-critical applications. All the aforementioned safety-critical environments introduce the new challenge of balancing the goal of reward maximization with the restriction of playing safe actions.

To address this major concern, the learning algorithm needs to guarantee that it does not violate certain safety constraints. From a bandit optimization point of view, [11, 96, 13, 94] study a linear bandit problem, in which, at each round, a linear cost constraint needs to be satisfied with high probability. For this problem, they propose no-regret algorithms that with high probability never violate the constraints. There has been a surge of research activity to address the issue of safe exploration in RL when the environment is modeled via the more challenging and complex setting of an unknown MDP. Many of existing algorithms

model the safety in RL via Constrained Markov Decision Process (CMDP), that extends the classical MDP to settings with extra constraints on the total expected cost over a horizon. To address the safety requirements in CMDPs, different approaches such as Primal-Dual Policy Optimization [99, 98, 112], Constrained Policy Optimization [7, 145], and Reward Constrained Policy Optimization [120] have been proposed. These algorithms come with either no theoretical guarantees or asymptotic convergence guarantee in the batch offline setting. In another line of work studying CMDP in online settings, [48, 127, 54, 154, 42, 100, 43, 141, 67] propose algorithms coming with sub-linear bounds on the number of constraint violation. Additionally, the safety constraint considered in the aforementioned papers is defined by the cumulative expected cost over a horizon falling below a certain threshold.

In this chapter, we propose an upper confidence bound (UCB)- based algorithm – termed Safe Linear UCB Q/V Iteration (SLUCB-QVI) – with the focus on deterministic policy selection respecting a more restrictive notion of safety requirements that must be satisfied at each time-step an action is played with high probability. We also present Randomized SLUCB-QVI (RSLUCB-QVI), a safe algorithm focusing on randomized policy selection without any constraint violation. For both algorithms, we assume the underlying MDP has linear structure and prove a regret bound that is order-wise comparable to those of its unsafe counter-parts.

Our main technical contributions allowing us to guarantee sub-linear regret bound while the safety constraints are never violated, include: 1) conservatively selecting actions from properly defined subsets of the unknown safe sets; and 2) exploiting careful algorithmic designs to ensure *optimism in the face of safety constraints*, i.e., the value function of our proposed algorithms are greater than the optimal value functions. See Sections 3.4,3.5, and 3.6 for details.

3.1.1 Key Contributions

Safety in reinforcement learning has become increasingly important in recent years. Yet, existing solutions either fail to strictly avoid choosing unsafe actions, which may lead to

catastrophic results in safety-critical systems, or fail to provide regret guarantees for settings where safety constraints need to be learned. In this chapter, we address both problems by first modeling safety as an unknown linear cost function of states and actions, which must always fall below a certain threshold. We then present algorithms, termed SLUCB-QVI and RSLUCB-QVI, for finite-horizon Markov decision processes (MDPs) with linear function approximation. We show that SLUCB-QVI and RSLUCB-QVI, while with *no safety violation*, achieve a $\tilde{\mathcal{O}}\left(\kappa\sqrt{d^3H^3T}\right)$ regret, nearly matching that of state-of-the-art unsafe algorithms, where H is the duration of each episode, d is the dimension of the feature mapping, κ is a constant characterizing the safety constraints, and T is the total number of action played. We further present numerical simulations that corroborate our theoretical findings.

3.2 Problem formulation

Finite-horizon Markov decision process. We consider a finite-horizon Markov decision process (MDP) denoted by $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, c)$, where $\mathcal{S}$ is the state set, $\mathcal{A}$ is the action set, H is the length of each episode (horizon), $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ are the transition probabilities, $r = \{r_h\}_{h=1}^H$ are the reward functions, and $c = \{c_h\}_{h=1}^H$ are the safety measures. For each time-step $h \in [H]$, $\mathbb{P}_h(s'|s,a)$ denotes the probability of transitioning to state s' upon playing action a at state s, and $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ and $c_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ are reward and constraint functions. We consider the learning problem where $\mathcal{S}$ and $\mathcal{A}$ are known, while the transition probabilities $\mathbb{P}_h$, rewards r_h and safety measures c_h are *unknown* to the agent and must be learned online. The agent interacts with its unknown environment described by M in episodes. In particular, at each episode k and time-step $h \in [H]$, the agent observes the state s_h^k, plays an action $a_h^k \in \mathcal{A}$, and observes a reward $r_h^k := r_h(s_h^k, a_h^k)$ and a noise-perturbed safety measure $z_h^k := c_h(s_h^k, a_h^k) + \epsilon_h^k$, where ϵ_h^k is a random additive noise.

Safety Constraint. We assume that the underlying system is safety-critical and the learning environment is subject to a side constraint that restricts the choice of actions. At each episode k and time-step $h \in [H]$, when being in state s_h^k, the agent must select a *safe*

action a_h^k such that

$$c_h(s_h^k, a_h^k) \leq \tau \tag{3.1}$$

with high probability, where τ is a known constant. We accordingly define the *unknown* safe action sets as

$$\mathcal{A}_h^{\mathbf{safe}}(s) := \{a \in \mathcal{A} \,:\, c_h(s,a) \leq \tau\}, \quad \forall (s,h) \in \mathcal{S} \times [H].$$

Thus, after observing state s_h^k at episode k and time-step $h \in [H]$, the agent's choice of action must belong to $\mathcal{A}_h^{\text{safe}}(s_h^k)$ with high probability. As a motivating example, consider a self-driving car. On the one hand, the agent (car) is rewarded for getting from point one to point two as fast as possible. On the other hand, the driving behavior must be constrained to respect traffic safety standards.

Goal. A *safe* deterministic policy is a function $\pi : \mathcal{S} \times [H] \to \mathcal{A}$, such that $\pi(s,h) \in \mathcal{A}_h^{\mathbf{safe}}(s)$ is the *safe* action the policy π suggests the agent to play at time-step $h \in [H]$ and state $s \in \mathcal{S}$. Thus, we define the set of safe policies by

$$\Pi^{\text{safe}} := \left\{\pi : \pi(s,h) \in \mathcal{A}_h^{\mathbf{safe}}(s),\ \forall (s,h) \in \mathcal{S} \times [H]\right\}.$$

For each $h \in [H]$, the cumulative expected reward obtained under a safe policy $\pi \in \Pi^{\text{safe}}$ during and after time-step h, known as the value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$, is defined by

$$V_h^\pi(s) := \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}\left(s_{h'}, \pi(s_{h'}, h')\right) \middle| s_h = s\right], \tag{3.2}$$

where the expectation is over the environment. We also define the state-action value action $Q_h^\pi : \mathcal{S} \times \mathcal{A}_h^{\mathbf{safe}}(.) \to \mathbb{R}$ for a safe policy $\pi \in \Pi^{\text{safe}}$ at time-step $h \in [H]$ by

$$Q_h^\pi(s,a) := \mathbb{E}\left[\sum_{h'=h+1}^{H} r_{h'}\left(s_{h'}, \pi(s_{h'}, h')\right) \middle| s_h = s, a_h = a\right]. \tag{3.3}$$

To simplify the notation, for any function f, we denote $[\mathbb{P}_h f](s,a) := \mathbb{E}_{s' \sim \mathbb{P}_h(.|s,a)} f(s')$. Let π_* be the optimal *safe* policy such that $V_h^{\pi_*}(s) := V_h^*(s) = \sup_{\pi \in \Pi^{\text{safe}}} V_h^\pi(s)$ for all $(s,h) \in \mathcal{S} \times [H]$.

Thus, for all $(s,h) \in \mathcal{S} \times [H]$ and $a \in \mathcal{A}_h^{\mathbf{safe}}(s)$, the Bellman equations for an arbitrary safe policy $\pi \in \Pi^{\text{safe}}$ and the optimal safe policy are:

$$Q_h^\pi(s,a) = r_h(s,a) + [\mathbb{P}_h V_{h+1}^\pi](s,a), \quad V_h^\pi(s) = Q_h^\pi(s,\pi(s,h)), \tag{3.4}$$

$$Q_h^*(s,a) = r_h(s,a) + [\mathbb{P}_h V_{h+1}^*](s,a), \quad V_h^*(s) = \max_{a \in \mathcal{A}_h^{\mathbf{safe}}(s)} Q_h^*(s,a), \tag{3.5}$$

where $V_{H+1}^\pi(s) = V_{H+1}^*(s) = 0$. Note that in classical RL without safety constraints, the Bellman optimality equation implies that there exists at least one optimal policy that is deterministic (see [30, 118, 117]). When considering solving the Bellman equation for the optimal policy, the presence of safety constraints is equivalent to solving it for an MDP without constraints but with different action sets for each $(s,h) \in \mathcal{S} \times [H]$, i.e., $\mathcal{A}_h^{\text{safe}}(s)$.

Let K be the total number of episodes, s_1^k be the initial state at the beginning of episode $k \in [K]$ and π_k be the high probability *safe* policy chosen by the agent during episode $k \in [K]$. Then the *cumulative pseudo-regret* is defined by

$$R_K := \sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k). \tag{3.6}$$

The agent's goal is to keep R_K as small as possible ($R_K/K \to 0$ as K grows large) *without violating the safety constraint in the process*, i.e., $\pi_k \in \Pi^{\text{safe}}$ for all $k \in [K]$ with high probability.

Linear Function Approximation. We focus on MDPs with linear transition kernels, reward, and cost functions that are encapsulated in the following assumption.

Assumption 4 (Linear MDP [31, 142, 64])**.** *$M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, c)$ is a linear MDP with feature map $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, if for any $h \in [H]$, there exist d unknown measures $\boldsymbol{\mu}_h^* := [\mu_h^{*(1)}, \ldots, \mu_h^{*(d)}]^\top$ over $\mathcal{S}$, and unknown vectors $\boldsymbol{\theta}_h^*, \boldsymbol{\gamma}_h^* \in \mathbb{R}^d$ such that $\mathbb{P}_h(.|s,a) = \langle \boldsymbol{\mu}_h^*(.), \boldsymbol{\phi}(s,a) \rangle$, $r_h(s,a) = \langle \boldsymbol{\theta}_h^*, \boldsymbol{\phi}(s,a) \rangle$, and $c_h(s,a) = \langle \boldsymbol{\gamma}_h^*, \boldsymbol{\phi}(s,a) \rangle$.*

This assumption highlights the definition of linear MDP, in which the Markov transition model, the reward functions, and the cost functions are linear in a feature mapping $\boldsymbol{\phi}$.

3.3 Prior Work

Safe RL with randomized policies: The problem of Safe RL formulated with Constrained Markov Decision Process (CMDP) with a focus on unknown dynamics and *randomized* policies is studied in [48, 127, 54, 154, 42, 100, 43, 141, 67]. In the above-mentioned papers, the goal is to find the optimal randomized policy that maximizes the reward value function $V_r^\pi(s)$ (expected total reward) while ensuring the cost value function $V_c^\pi(s)$ (expected total cost) does not exceed a certain threshold. This safety requirement is defined over a *horizon*, in expectation with respect to the environment and the randomization of the policy, and consequently is less strict than the safety requirement considered in this chapter, which must be satisfied at each time-step an action is played. In addition to their different problem formulations, the theoretical guarantees of these works fundamentally differ from the ones provided in our work. The recent closely-related work of [42] studies constrained finite-horizon MDPs with a linear structure as considered in our work via a primal-dual-type policy optimization algorithm that achieves a $\mathcal{O}(dH^{2.5}\sqrt{T})$ regret and constraint violation and can only be applied to settings with finite action set $\mathcal{A}$. The algorithm of [48] obtains a $\mathcal{O}(|\mathcal{S}|H^2\sqrt{|\mathcal{S}||\mathcal{A}|T})$ regret and constraint violation in the episodic finite-horizon tabular setting via linear program and primal-dual policy optimization. In [100], the authors study an adversarial stochastic shortest path problem under constraints with $\mathcal{O}(|\mathcal{S}|H\sqrt{|\mathcal{A}|T})$ regret and constraint violation. [43] proposes a primal-dual algorithm for solving discounted infinite horizon CMDPs that achieves a global convergence with rate $\mathcal{O}(1/\sqrt{T})$ regarding both the optimality gap and the constraint violation. In contrast to the aforementioned works which can only guarantee bounds on the number of constraint violation, our algorithms *never* violate the safety constraint during the learning process.

Besides primal-dual methods, in [37] Lyapunov functions are leveraged to handle the constraints. [149] proposes a constrained policy gradient algorithm with convergence guarantee. Both above-stated works focus on solving CMDPs with known transition model and constraint function without providing regret guarantees.

Safe RL with GPs and deterministic transition model and policies: In another line of work, [126, 29, 130, 129] use Gaussian processes to model the dynamics with deterministic transitions and/or the value function in order to be able to estimate the constraints and guarantee safe learning. Despite the fact that some of these algorithms are approximately safe, analysing the convergence is challenging and the regret analysis is lacking.

3.4 Safe Linear UCB Q/V Iteration

In this section, we present *Safe Linear Upper Confidence Bound Q/V Iteration* (SLUCB-QVI) summarized in Algorithm 3, which is followed by a high-level description of its performance in Section 3.4. First, we introduce the following necessary assumption and set of notations used in describing Algorithm 3 and its analysis in the next sections.

Assumption 5 (Non-empty safe sets)**.** *For all $s \in \mathcal{S}$, there exists a known safe action $a_0(s)$ such that $a_0(s) \in \mathcal{A}_h^{\boldsymbol{safe}}(s)$ with known safety measure $\tau_h(s) := \left\langle \boldsymbol{\phi}\left(s, a_0\left(s\right)\right), \boldsymbol{\gamma}_h^* \right\rangle < \tau$ for all $h \in [H]$.*

Knowing safe actions $a_0(s)$ is necessary for solving the safe linear MDP setting studied in this chapter, which requires the constraint (3.1) to be satisfied from the very first round. This assumption is also realistic in many practical examples, where the known safe action could be the one suggested by the current strategy of the company or a very cost-neutral action that does not necessarily have high reward but its cost is far from the threshold. It is possible to relax the assumption of knowing the cost of the safe actions $\tau_h(s)$. In this case, the agent starts by playing $a_0(s)$ for $T_h(s)$ rounds at time-steps h in order to construct a conservative estimator for the gap $\tau - \tau_h(s)$. $T_h(s)$ is selected in an adaptive way and in Appendix B.1.4, we show that $\frac{16\log(K)}{(\tau-\tau_h(s))^2} \leq T_h(s) \leq \frac{64\log(K)}{(\tau-\tau_h(s))^2}$. After $T_h(s)$ rounds, the agent relies on these estimates of $\tau_h(s)$ in the computation of estimated safe set of policies (discussed shortly).

Notations. For any vector $\mathbf{x} \in \mathbb{R}^d$, define the normalized vector $\tilde{\mathbf{x}} := \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$. We define the span of the safe feature $\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)$ as $\mathcal{V}_s = \text{ span}\left(\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)\right) := \left\{\alpha\boldsymbol{\phi}\left(s, a_0\left(s\right)\right) : \alpha \in \mathbb{R}\right\}$ and the orthogonal complement of $\mathcal{V}_s$ as $\mathcal{V}_s^{\perp} := \{\mathbf{y} \in \mathbb{R}^d : \langle \mathbf{y}, \mathbf{x}\rangle = 0,\ \forall \mathbf{x} \in \mathcal{V}_s\}$. For any

Algorithm 3 SLUCB-QVI

1: **Input:** $\mathcal{A}$, λ, δ, H, K, τ, $\kappa_h(s)$

2: $\mathbf{A}_h^1 = \lambda I,\ \mathbf{A}_{h,s}^1 = \lambda\left(I - \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)\tilde{\boldsymbol{\phi}}^\top\left(s, a_0\left(s\right)\right)\right)\ \mathbf{b}_h^1 = \mathbf{r}_{h,s}^1 = \mathbf{0},\ \forall (s,h) \in \mathcal{S} \times [H], Q_{H+1}^k(.,.) = 0,\ \forall k \in [K]$

3: **for** episodes $k = 1, \ldots, K$ **do**

4: Observe the initial state s_1^k.

5: **for** time-steps $h = H, \ldots, 1$ **do**

6: Compute $\mathcal{A}_h^k(s)$ as in (3.9) $\forall s \in \mathcal{S}$.

7: Compute $Q_h^k(s,a)$ as in (3.10) $\forall (s,a) \in \mathcal{S} \times \mathcal{A}_h^k(.)$.

8: **end for**

9: **for** time-steps $h = 1, \ldots, H$ **do**

10: Play $a_h^k = \arg\max_{a \in \mathcal{A}_h^k(s_h^k)} Q_h^k(s_h^k, a)$ and observe s_{h+1}^k, r_h^k and z_h^k.

11: **end for**

12: **end for**

$\mathbf{x} \in \mathbb{R}^d$, denote by $\Phi_0(s, \mathbf{x}) := \left\langle \mathbf{x}, \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)\right\rangle \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)$ its projection on $\mathcal{V}_s$, and, by $\Phi_0^\perp(s, \mathbf{x}) := \mathbf{x} - \Phi_0(s, \mathbf{x})$ its projection onto the orthogonal subspace $\mathcal{V}_s^\perp$. Moreover, for ease of notation, let $\boldsymbol{\phi}_h^k := \boldsymbol{\phi}(s_h^k, a_h^k)$.

3.4.1 Overview

From a high-level point of view, our algorithm is the safe version of LSVI-UCB proposed by [64]. In particular, each episode consists of two loops over all time-steps. The first loop (Lines 5-8) updates the quantities $\mathcal{A}_h^k$, estimated safe sets, and Q_h^k, action-value function, that are used to execute the *upper confidence bound* policy $a_h^k = \arg\max_{a \in \mathcal{A}_h^k(s_h^k)} Q_h^k(s_h^k, a)$ in the second loop (Lines 9-11). The key difference between SLUCB-QVI and LSVI-UCB is the requirement that chosen actions a_h^k must always belong to unknown safe sets $\mathcal{A}_h^{\mathbf{safe}}(s_h^k)$. To this end, at each episode $k \in [K]$, in an extra step in the first loop (Line 6), the agent computes a set $\mathcal{A}_h^k(s)$ for all $s \in \mathcal{S}$, which we will show is guaranteed to be a subset of the unknown safe set $\mathcal{A}_h^{\mathbf{safe}}(s)$, and therefore, is a good candidate to select action a_h^k from in

the second loop (Line 10). Construction of $\mathcal{A}_h^k(s)$ depends on an appropriate confidence set around the unknown parameter $\boldsymbol{\gamma}_h^*$ used in the definition of safety constraints (see Assumption 4). Since the agent has knowledge of $\tau_h(s) := \left\langle \boldsymbol{\phi}\left(s, a_0\left(s\right)\right), \boldsymbol{\gamma}_h^* \right\rangle$ (see Assumption 5), it can compute $z_{h,s}^k := \left\langle \Phi_0^\perp\left(s, \boldsymbol{\phi}_h^k\right), \Phi_0^\perp\left(s, \boldsymbol{\gamma}_h^*\right) \right\rangle + \epsilon_h^k = z_h^k - \frac{\left\langle \boldsymbol{\phi}_h^k, \tilde{\boldsymbol{\phi}}\left(s, a_0(s)\right) \right\rangle}{\left\| \boldsymbol{\phi}\left(s, a_0(s)\right) \right\|_2} \tau_h(s)$, i.e., the cost incurred by a_h^k along the subspace $\mathcal{V}_s^\perp$, which is orthogonal to $\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)$. Thus, the agent does not need to build confidence sets around $\boldsymbol{\gamma}_h^*$ along the normalized safe feature vector, $\tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)$. Instead, it only builds the following confidence sets around $\Phi_0^\perp\left(s, \boldsymbol{\gamma}_h^*\right)$ which is along the orthogonal direction of $\tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)$:

$$\mathcal{C}_h^k(s) := \left\{ \boldsymbol{\nu} \in \mathbb{R}^d : \left\| \boldsymbol{\nu} - \boldsymbol{\gamma}_{h,s}^k \right\|_{\mathbf{A}_{h,s}^k} \leq \beta \right\}, \tag{3.7}$$

where $\boldsymbol{\gamma}_{h,s}^k := \left(\mathbf{A}_{h,s}^k\right)^{-1} \mathbf{r}_{h,s}^k$ is the regularized least-squares estimator of $\Phi_0^\perp\left(s, \boldsymbol{\gamma}_h^*\right)$ computed by the inverse of Gram matrix $\mathbf{A}_{h,s}^k := \lambda\left(I - \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right) \tilde{\boldsymbol{\phi}}^\top\left(s, a_0\left(s\right)\right)\right) + \sum_{j=1}^{k-1} \Phi_0^\perp\left(s, \boldsymbol{\phi}_h^j\right) \Phi_0^{\perp,\top}\left(s, \boldsymbol{\phi}_h^j\right)$ and $\mathbf{r}_{h,s}^k := \sum_{j=1}^{k-1} z_{h,s}^j \Phi_0^\perp\left(s, \boldsymbol{\phi}_h^j\right)$. The exploration factor β will be defined shortly in Theorem 4 such that it guarantees that the event

$$\mathcal{E}_1 := \left\{ \Phi_0^\perp\left(s, \boldsymbol{\gamma}_h^*\right) \in \mathcal{C}_h^k(s),\ \forall (s,h,k) \in \mathcal{S} \times [H] \times [K] \right\} \tag{3.8}$$

i.e., $\Phi_0^\perp\left(s, \boldsymbol{\gamma}_h^*\right)$ belongs to the confidence sets $\mathcal{C}_h^k(s)$, holds with high probability. In the implementations, we treat β as a tuning parameter. Conditioned on event $\mathcal{E}_1$, the agent is ready to compute the following inner approximations of the true unknown safe sets $\mathcal{A}_h^{\mathbf{safe}}$ for all $s \in \mathcal{S}$:

$$\begin{aligned} \mathcal{A}_h^k(s) = \Bigg\{ a \in \mathcal{A} : & \frac{\left\langle \Phi_0\left(s, \boldsymbol{\phi}(s,a)\right), \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right) \right\rangle}{\left\| \boldsymbol{\phi}\left(s, a_0\left(s\right)\right) \right\|_2} \tau_h(s) + \left\langle \boldsymbol{\gamma}_{h,s}^k, \Phi_0^\perp\left(s, \boldsymbol{\phi}(s,a)\right) \right\rangle \\ & + \beta \left\| \Phi_0^\perp\left(s, \boldsymbol{\phi}(s,a)\right) \right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}} \leq \tau \Bigg\}. \end{aligned} \tag{3.9}$$

Note that $\frac{\left\langle \Phi_0\left(s, \boldsymbol{\phi}(s,a)\right), \tilde{\boldsymbol{\phi}}\left(s, a_0(s)\right) \right\rangle}{\left\| \boldsymbol{\phi}\left(s, a_0(s)\right) \right\|_2} \tau_h(s)$ is the known cost of action a at state s along direction $\tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)$ and $\max_{\boldsymbol{\nu} \in \mathcal{C}_h^k(s)} \left\langle \Phi_0^\perp\left(s, \boldsymbol{\phi}(s,a)\right), \boldsymbol{\nu} \right\rangle = \left\langle \boldsymbol{\gamma}_{h,s}^k, \Phi_0^\perp\left(s, \boldsymbol{\phi}(s,a)\right) \right\rangle +$

$\beta\left\|\Phi_0^{\perp}\left(s,\boldsymbol{\phi}(s,a)\right)\right\|_{\left(\mathbf{A}_{h,s}^{k}\right)^{-1}}$ is its maximum possible cost in the orthogonal space $\mathcal{V}_s^{\perp}$. Thus, $\frac{\left\langle \Phi_0\left(s,\boldsymbol{\phi}(s,a)\right),\tilde{\boldsymbol{\phi}}\left(s,a_0(s)\right)\right\rangle}{\left\|\tilde{\boldsymbol{\phi}}\left(s,a_0(s)\right)\right\|_2}\tau_h(s) + \left\langle \boldsymbol{\gamma}_{h,s}^{k}, \Phi_0^{\perp}\left(s,\boldsymbol{\phi}(s,a)\right)\right\rangle + \beta\left\|\Phi_0^{\perp}\left(s,\boldsymbol{\phi}(s,a)\right)\right\|_{\left(\mathbf{A}_{h,s}^{k}\right)^{-1}}$ is a high probability upper bound on the true unknown cost $\langle\boldsymbol{\phi}(s,a),\boldsymbol{\gamma}_h^*\rangle$, which implies that $\mathcal{A}_h^k(s)\subset\mathcal{A}_h^{\mathbf{safe}}(s)$.

Proposition 1. *Conditioned on $\mathcal{E}_1$ in (3.8), for all $(s,h,k)\in\mathcal{S}\times[H]\times[K]$, it holds that $\left\langle\boldsymbol{\phi}(s,a),\boldsymbol{\gamma}_h^*\right\rangle\leq\tau,\ \forall a\in\mathcal{A}_h^k(s)$.*

Thus, conditioned on $\mathcal{E}_1$, the decision rule $a_h^k:=\arg\max_{a\in\mathcal{A}_h^k(s_h^k)}Q_h^k(s_h^k,a)$ in Line 10 of Algorithm 3 suggests that a_h^k does not violate the safety constraint. Note that $\mathcal{A}_h^k(s)$ is always non-empty, since as a consequence of Assumption 5, the safe action $a_0(s)$ is always in $\mathcal{A}_h^k(s)$.

Now that the estimated safe sets $\mathcal{A}_h^k(s)$ are constructed, we describe how the action-value functions Q_h^k are computed to be used in the UCB decision rule, selecting the action a_h^k in the second loop of the algorithm. The linear structure of the MDP allows us to parametrize $Q_h^*(s,a)$ by a linear form $\langle\mathbf{w}_h^*,\boldsymbol{\phi}(s,a)\rangle$, where $\mathbf{w}_h^*:=\boldsymbol{\theta}_h^*+\int_{\mathcal{S}}V_{h+1}^*(s')d\boldsymbol{\mu}(s')$. Thus, a natural idea to estimate $Q_h^*(s,a)$ is to solve least-squares problem for $\mathbf{w}_h^*$. In fact, for all $(s,a)\in\mathcal{S}\times\mathcal{A}_h^k(.)$, the agent computes $Q_h^k(s,a)$ defined as

$$Q_h^k(s,a)=\min\left\{\left\langle\mathbf{w}_h^k,\boldsymbol{\phi}(s,a)\right\rangle+\kappa_h(s)\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}},H\right\},\tag{3.10}$$

where $\mathbf{w}_h^k:=\left(\mathbf{A}_h^k\right)^{-1}\mathbf{b}_h^k$ is the regularized least-squares estimator of $\mathbf{w}_h^*$ computed by the inverse of Gram matrix $\mathbf{A}_h^k:=\lambda I+\sum_{j=1}^{k-1}\boldsymbol{\phi}_h^j{\boldsymbol{\phi}_h^j}^{\top}$ and $\mathbf{b}_h^k:=\sum_{j=1}^{k-1}\boldsymbol{\phi}_h^j\left[r_h^j+\max_{a\in\mathcal{A}_{h+1}^k(s_{h+1}^j)}Q_{h+1}^k(s_{h+1}^j,a)\right]$. Here, $\kappa_h(s)\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}$ is an exploration bonus that is characterized by: 1) β that encourages enough exploration regarding the uncertainty about r and $\mathbb{P}$; and 2) $\kappa_h(s)>1$ that encourages enough exploration regarding the uncertainty about c. While we make use of standard analysis of unsafe bandits and MDPs [2] and [64] to define β, appropriately quantifying $\kappa_h(s)$ is the main challenge the presence of safety constraints brings to the analysis of SLUCB-QVI compared to the unsafe LSVI-UCB and it is stated in Lemma 3.

3.5 Theoretical Guarantees of SLUCB-QVI

In this section, we discuss the technical challenges the presence of safety constraints brings to our analysis and provide a regret bound for SLUCB-QVI. Before these, we make the remaining necessary assumptions under which our proposed algorithm operates and achieves good regret bound.

Assumption 6 (Subgaussian noise)**.** *For all* $(h,k) \in [H] \times [K]$*,* ϵ_h^k *is a zero-mean* σ*-subGaussian random variable.*

Assumption 7 (Boundedness)**.** *Without loss of generality,* $\left\|\boldsymbol{\phi}(s,a)\right\|_2 \leq 1$ *for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$*, and* $\max\left(\left\|\boldsymbol{\mu}_h^*(\mathcal{S})\right\|_2, \left\|\boldsymbol{\theta}_h^*\right\|_2, \left\|\boldsymbol{\gamma}_h^*\right\|_2\right) \leq \sqrt{d}$ *for all* $h \in [H]$*.*

Assumption 8 (Star convex sets)**.** *For all* $s \in \mathcal{S}$*, the set* $\mathcal{D}(s) := \left\{\boldsymbol{\phi}(s,a) : a \in \mathcal{A}\right\}$ *is a star convex set around the safe feature* $\boldsymbol{\phi}\left(s, a_0(s)\right)$*, i.e., for all* $\mathbf{x} \in \mathcal{D}(s)$ *and* $\alpha \in [0,1]$*,* $\alpha\mathbf{x} + (1-\alpha)\boldsymbol{\phi}\left(s, a_0(s)\right) \in \mathcal{D}(s)$*.*

Assumptions 6 and 7 are standard in linear MDP and bandit literature [64, 96, 11]. Assumption 8 is necessary to ensure that the agent has the opportunity to explore the feature space around the given safe feature vector $\boldsymbol{\phi}\left(s, a_0(s)\right)$. For example, consider a simple setting where $\mathcal{S} = \{s_1\}, \mathcal{A} = \{a_1, a_2\}, H = 1, \boldsymbol{\mu}^*(s_1) = (1,1), \boldsymbol{\theta}^* = (0,1), \boldsymbol{\gamma}^* = (0,1), \tau = 2$, $a_0(s_1) = a_2$, and $\mathcal{D}(s_1) = \{\boldsymbol{\phi}(s_1,a_1), \boldsymbol{\phi}(s_1,a_2)\} = \{(0,1),(1,0)\}$, which is not a star convex set. Here, both actions a_1 and a_2 are safe. The optimal safe policy always plays a_1, which gives the highest reward. However, if $\mathcal{D}(s_1)$ does not contain the whole line connecting $(1,0)$ and $(0,1)$, the agent keeps playing a_2 and will not be able to explore other safe action and identify that the optimal policy would always select a_1. Also, it is worth mentioning that the star convexity of the sets $\mathcal{D}(s)$ is a milder assumption than convexity assumption considered in existing safe algorithms of [11, 94].

Given these assumptions, we are now ready to present the formal guarantees of SLUCB-QVI in the following theorem.

Theorem 4 (Regret of SLUCB-QVI)**.** *Under Assumptions 4, 5, 6, 7, and 8, there exists an absolute constant* $c_\beta > 0$ *such that for any fixed* $\delta \in (0, 0.5)$*, if we set* $\beta :=$

$\max\left(\sigma\sqrt{d\log\left(\frac{2+\frac{2T}{\lambda}}{\delta}\right)}+\sqrt{\lambda d}, c_\beta dH\sqrt{\log(\frac{dT}{\delta})}\right)$, *and* $\kappa_h(s) := \frac{2H}{\tau-\tau_h(s)}+1$, *then with probability at least* $1-2\delta$, *it holds that* $R_K \leq 2H\sqrt{T\log(\frac{dT}{\delta})}+(1+\kappa)\beta\sqrt{2dHT\log\left(1+\frac{K}{d\lambda}\right)}$, *where* $\kappa := \max_{(s,h)\in\mathcal{S}\times[H]}\kappa_h(s)$

Here, $T = KH$ is the total number of action plays. We observe that the regret bound is of the same order as that of state-of-the-art unsafe algorithms, such as LSVI-UCB [64], with only an additional factor κ in its second term. The complete proof is reported in the Appendix B.1.3. In the following section, we give a sketch of the proof.

3.5.1 Proof Sketch of Theorem 4

First, we state the following theorem borrowed from [2, 64].

Theorem 5 (Thm. 2 in [2] and Lemma B.4 in [64])**.** *For any fixed policy* π, *define* $V_h^k(s) := \max_{a\in\mathcal{A}_h^k(s,a)} Q_h^k(s,a)$, *and the event*

$$\mathcal{E}_2 := \Big\{\Big|\langle \mathbf{w}_h^k, \boldsymbol{\phi}(s,a)\rangle - Q_h^\pi(s,a) + [\mathbb{P}_h(V_{h+1}^\pi - V_{h+1}^k)](s,a)\Big| \leq \beta\big\|\boldsymbol{\phi}(s,a)\big\|_{\left(\mathbf{A}_h^k\right)^{-1}}$$
$$, \forall (a,s,h,k) \in \mathcal{A}\times\mathcal{S}\times[H]\times[K]\Big\},$$

and recall the definition of $\mathcal{E}_1$ *in* (3.8). *Then, under Assumptions 4, 5, 6, 7, and the definition of* β *in Theorem 4, there exists an absolute constant* $c_\beta > 0$, *such that for any fixed* $\delta \in (0, 0.5)$, *with probability at least* $1-\delta$, *the event* $\mathcal{E} := \mathcal{E}_2 \cap \mathcal{E}_1$ *holds.*

As our main technical contribution, in Lemma 3, we prove that when $\kappa_h(s) := \frac{2H}{\tau-\tau_h(s)} + 1$, then *optimism in the face of safety constraint*, i.e., $Q_h^*(s,a) \leq Q_h^k(s,a)$ is guaranteed. Intuitively, this is required because the maximization in Line 10 of Algorithm 3 is not over the entire $\mathcal{A}_h^{\mathbf{safe}}(s_h^k)$, but only a subset of it. Thus, larger values of $\kappa_h(s)$ (compared to $\kappa_h(s) = 1$ in unsafe algorithm LSVI-UCB) are needed to provide enough exploration to the algorithm so that the selected actions in $\mathcal{A}_h^k(s_h^k)$ are -often enough- *optimistic*, i.e., $Q_h^*(s,a) \leq Q_h^k(s,a)$.

Lemma 3 (Optimism in the face of safety constraint in SLUCB-QVI)**.** *Let* $\kappa_h(s) :=$

$\frac{2H}{\tau-\tau_h(s)} + 1$ *and Assumptions 4,5,6,7,8 hold. Then, conditioned on* $\mathcal{E}$*, it holds that* $V_h^*(s) \leq V_h^k(s), \forall (s,h,k) \in \mathcal{S} \times [H] \times [K]$.

We report the proof in Appendix B.1.2. As a direct conclusion of Lemma 3 and on event $\mathcal{E}_2$ defined in Theorem 5, we have

$$\begin{aligned} Q_h^*(s,a) &\leq \left\langle \mathbf{w}_h^k, \boldsymbol{\phi}(s,a) \right\rangle + \beta \left\| \boldsymbol{\phi}(s,a) \right\|_{\left(\mathbf{A}_h^k\right)^{-1}} + [\mathbb{P}_h V_{h+1}^* - V_{h+1}^k](s,a) && \text{(Event } \mathcal{E}_2\text{)} \\ &\leq Q_h^k(s,a). && \text{(Lemma 3)} \end{aligned}$$

This is encapsulated in the following corollary.

Corollary 1 (UCB)**.** *Let* $\kappa_h(s) := \frac{2H}{\tau-\tau_h(s)} + 1$ *and Let Assumptions 4,5,6,7,8 hold. Then, conditioned on* $\mathcal{E}$*, it holds that* $Q_h^*(s,a) \leq Q_h^k(s,a), \forall (a,s,h,k) \in \mathcal{A} \times \mathcal{S} \times [H] \times [K]$.

After proving UCB nature of SLUCB-QVI using Lemma 3, we are ready to exploit the standard analysis of classical unsafe LSVI-UCB [64] to complete the analysis and establish the final regret bound of SLUCB-QVI.

3.6 Extension to Randomized Policy Selection

SLUCB-QVI presented in Section 3.4 can only output a deterministic policy. In this section, we show that our results can be extended to the setting of randomized policy selection, which might be desirable in practice. A randomized policy $\pi : \mathcal{S} \times [H] \to \Delta_{\mathcal{A}}$ maps states and time-steps to distributions over actions such that $a \sim \pi(s,h)$ is the action the policy π suggests the agent to play at time-step $h \in [H]$ when being at state $s \in \mathcal{S}$. At each episode k and time-step $h \in [H]$, when being in state s_h^k, the agent must draw its action a_h^k from a *safe* policy $\pi_k(s_h^k, h)$ such that

$$\mathbb{E}_{a_h^k \sim \pi_k(s_h^k,h)} c_h(s_h^k, a_h^k) \leq \tau \tag{3.11}$$

with high probability. We accordingly define the *unknown* set of safe policies by

$$\tilde{\Pi}^{\mathbf{safe}} := \left\{ \pi : \pi(s,h) \in \Gamma_h^{\mathbf{safe}}(s),\ \forall (s,h) \in \mathcal{S} \times [H] \right\},$$

where $\Gamma_h^{\mathbf{safe}}(s) := \left\{\theta \in \Delta_{\mathcal{A}} : \mathbb{E}_{a\sim\theta} c_h(s,a) \leq \tau\right\}$. Thus, after observing state s_h^k at time-step $h \in [H]$ in episode k, the agent's choice of policy must belong to $\Gamma_h^{\mathbf{safe}}(s_h^k)$ with high probability. In this formulation, the expectation in the definition of (action-) value functions for a policy π is over both the environment and the randomness of policy π. We denote them by $\tilde{V}_h^\pi$ and $\tilde{Q}_h^\pi$ to distinguish them from V_h^π and Q_h^π defined in (3.2) and (3.3) for a deterministic policy π. Let π_* be the optimal safe policy such that $\tilde{V}_h^{\pi_*}(s) := \tilde{V}_h^*(s) = \sup_{\pi\in\tilde{\Pi}^{\mathbf{safe}}} \tilde{V}_h^\pi(s)$ for all $(s,h) \in \mathcal{S}\times[H]$. Thus, for all $(a,s,h) \in \mathcal{A}\times\mathcal{S}\times[H]$, the Bellman equations for a safe policy $\pi \in \tilde{\Pi}^{\mathbf{safe}}$ and the optimal safe policy are

$$\tilde{Q}_h^\pi(s,a) = r_h(s,a) + [\mathbb{P}_h \tilde{V}_{h+1}^\pi](s,a), \quad \tilde{V}_h^\pi(s) = \mathbb{E}_{a\sim\pi(s,h)}\left[\tilde{Q}_h^\pi(s,a)\right], \tag{3.12}$$

$$\tilde{Q}_h^*(s,a) = r_h(s,a) + [\mathbb{P}_h \tilde{V}_{h+1}^*](s,a), \quad \tilde{V}_h^*(s) = \max_{\theta\in\Gamma_h^{\mathbf{safe}}(s)} \mathbb{E}_{a\in\theta}\left[\tilde{Q}_h^*(s,a)\right], \tag{3.13}$$

where $\tilde{V}_{H+1}^\pi(s) = \tilde{V}_{H+1}^*(s) = 0$, and the cumulative regret is defined as $R_K := \sum_{k=1}^K \tilde{V}_1^*(s_1^k) - \tilde{V}_1^{\pi_k}(s_1^k)$. This definition of safety constraint in (3.11) frees us from star-convexity assumption on the sets $\mathcal{D}(s) := \left\{\boldsymbol{\phi}(s,a) : a \in \mathcal{A}\right\}$ (Assumption 8), which is necessary for the deterministic policy selection approach. We propose a modification of SLUCB-QVI which is tailored to this new formulation and termed Randomized SLUCB-QVI (RSLUCB-QVI). This new algorithm also achieves a sub-linear regret with the same order as that of SLUCB-QVI, i.e., $\tilde{\mathcal{O}}\left(\kappa\sqrt{d^3H^3T}\right)$.

While RSLUCB-QVI respects a milder definition of the safety constraint (cf. (3.11)) compared to that considered in SLUCB-QVI (cf. (3.1)), it still possesses significant superiorities over other existing algorithms solving CMDP with randomized policy selection [48, 127, 54, 154, 42, 100, 43, 141, 67]. First, the safety constraint considered in these algorithms is defined by the *cumulative* expected cost over a horizon falling below a certain threshold, while RSLUCB-QVI guarantees that the expected cost incurred at each time-step an action is played (not over a horizon) is less than a threshold. Second, even for this looser definition of safety constraint, the best these algorithms can guarantee in terms of constraint satisfaction is a sub-linear bound on the number of constraint violation, whereas RSLUCB-QVI ensures *no constraint violation*.

3.6.1 Randomized SLUCB-QVI

We now describe RSLUCB-QVI summarized in Algorithm 4. Let $\boldsymbol{\phi}^{\theta}(s) := \mathbb{E}_{a\sim\theta}\boldsymbol{\phi}(s,a)$. At each episode $k \in [K]$, in the first loop, the agent computes the estimated set of true unknown set $\Gamma_h^{\mathbf{safe}}(s)$ for all $s \in \mathcal{S}$ as follows:

$$\begin{aligned}\Gamma_h^k(s) &:= \left\{\theta \in \Delta_{\mathcal{A}} : \mathbb{E}_{a\sim\theta}\left[\frac{\left\langle \Phi_0\left(s,\boldsymbol{\phi}(s,a)\right), \tilde{\boldsymbol{\phi}}\left(s,a_0\left(s\right)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s,a_0\left(s\right)\right)\right\|_2}\tau_h(s)\right] + \max_{\boldsymbol{\nu}\in\mathcal{C}_h^k(s)}\left\langle \Phi_0^{\perp}\left(s,\mathbb{E}_{a\sim\theta}\left[\boldsymbol{\phi}(s,a)\right]\right),\boldsymbol{\nu}\right\rangle \le \tau\right\} \\ &= \left\{\theta \in \Delta_{\mathcal{A}} : \frac{\left\langle \Phi_0\left(s,\boldsymbol{\phi}^{\theta}(s)\right), \tilde{\boldsymbol{\phi}}\left(s,a_0\left(s\right)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s,a_0\left(s\right)\right)\right\|_2}\tau_h(s) + \left\langle \boldsymbol{\gamma}_{h,s}^k, \Phi_0^{\perp}\left(s,\boldsymbol{\phi}^{\theta}(s)\right)\right\rangle \right. \\ &\left. +\beta\left\|\Phi_0^{\perp}\left(s,\boldsymbol{\phi}^{\theta}(s)\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}} \le \tau\right\}. \end{aligned} \tag{3.14}$$

Note that due to the linear structure of the MDP, we can again parametrize $\tilde{Q}_h^*(s,a)$ by a linear form $\langle\tilde{\mathbf{w}}_h^*, \boldsymbol{\phi}(s,a)\rangle$, where $\tilde{\mathbf{w}}_h^* := \boldsymbol{\theta}_h^* + \int_{\mathcal{S}}\tilde{V}_{h+1}^*(s')d\boldsymbol{\mu}(s')$. In the next step, for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, the agent computes

$$\tilde{Q}_h^k(s,a) = \left\langle \tilde{\mathbf{w}}_h^k, \boldsymbol{\phi}(s,a)\right\rangle + \kappa_h(s)\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}, \tag{3.15}$$

where $\tilde{\mathbf{w}}_h^k := \left(\mathbf{A}_h^k\right)^{-1}\tilde{\mathbf{b}}_h^k$ is the regularized least-squares estimator of $\tilde{\mathbf{w}}_h^*$ computed by the Gram matrix $\mathbf{A}_h^k$ and $\tilde{\mathbf{b}}_h^k := \sum_{j=1}^{k-1}\boldsymbol{\phi}_h^j\left[r_h^j + \min\left\{\max_{\theta\in\Gamma_{h+1}^k(s_{h+1}^j)}\mathbb{E}_{a\sim\theta}\left[\tilde{Q}_{h+1}^k(s_{h+1}^j,a)\right], H\right\}\right]$. After these computations in the first loop, the agent draws actions a_h^k from distribution $\Gamma_h^k(s_h^k)$ in the second loop. Define $\tilde{V}_h^k(s) := \min\left\{\max_{\theta\in\Gamma_h^k(s)}\mathbb{E}_{a\sim\theta}\left[\tilde{Q}_h^k(s,a)\right], H\right\}$, and

$$\begin{aligned}\mathcal{E}_3 := \Bigg\{&\left|\langle\tilde{\mathbf{w}}_h^k,\boldsymbol{\phi}(s,a)\rangle - \tilde{Q}_h^{\pi}(s,a) + [\mathbb{P}_h\tilde{V}_{h+1}^{\pi} - \tilde{V}_{h+1}^k](s,a)\right| \le \beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}} \\ &, \forall(a,s,h,k) \in \mathcal{A}\times\mathcal{S}\times[H]\times[K]\Bigg\}.\end{aligned}$$

It can be easily shown that the results stated in Theorem 5 hold for the settings focusing on randomized policies, i.e., under Assumptions 4, 5, 6, and 7, and by the definition of β in Theorem 4, with probability at least $1-2\delta$, the event $\tilde{\mathcal{E}} := \mathcal{E}_1 \cap \mathcal{E}_3$ holds. Therefore, as a direct conclusion of Proposition 1, it is guaranteed that conditioned on $\mathcal{E}_1$, all the policies inside $\Gamma_h^k(s)$ are safe, i.e., $\Gamma_h^k(s) \subset \Gamma_h^{\mathbf{safe}}(s)$. Now, in the following lemma, we quantify $\kappa_h(s)$.

Algorithm 4 RSLUCB-QVI

Input: $\mathcal{A}$, λ, δ, H, K, τ, $\kappa_h(s)$

$\mathbf{A}_h^1 = \lambda I,\ \mathbf{A}_{h,s}^1 = \lambda\left(I - \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)\tilde{\boldsymbol{\phi}}^\top\left(s, a_0\left(s\right)\right)\right)\ \tilde{\mathbf{b}}_h^1 = \mathbf{r}_{h,s}^1 = \mathbf{0},\ \forall (s,h) \in \mathcal{S} \times [H], \tilde{Q}_{H+1}^k(.,.) = 0,\ \forall k \in [K]$

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

  Observe the initial state s_1^k.

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

    Compute $\Gamma_h^k(s)$ as in (3.14) $\forall s \in \mathcal{S}$.

    Compute $\tilde{Q}_h^k(s,a)$ as in (3.15) $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$.

  end for

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

    Play $a_h^k \sim \arg\max_{\theta \in \Gamma_h^k(s_h^k)} \mathbb{E}_{a\sim\theta}\left[\tilde{Q}_h^k(s_h^k, a)\right]$ and observe s_{h+1}^k, r_h^k and z_h^k.

  end for

end for

Lemma 4 (Optimism in the face of safety constraint in RSLUCB-QVI)**.** *Let* $\kappa_h(s) := \frac{2H}{\tau - \tau_h(s)} + 1$ *and Assumptions 4,5,6,7 hold. Then, conditioned on event* $\tilde{\mathcal{E}}$*, it holds that* $\tilde{V}_h^*(s) \leq \tilde{V}_h^k(s), \forall (s,h,k) \in \mathcal{S} \times [H] \times [K]$.

The proof is included in Appendix B.2.1. Using Lemma 4, we show that $\tilde{Q}_h^*(s,a) \leq \tilde{Q}_h^k(s,a), \forall (a,s,h,k) \in \mathcal{A} \times \mathcal{S} \times [H] \times [K]$. This highlights the UCB nature of RSLUCB-QVI, allowing us to exploit the standard analysis of unsafe LSVI-UCB [64] to establish the regret bound.

Theorem 6 (Regret of RSLUCB-QVI)**.** *Under Assumptions 4, 5, 6, and 7, there exists an absolute constant* $c_\beta > 0$ *such that for any fixed* $\delta \in (0, 1/3)$*, and the definition of* β *in Theorem 4, if we set* $\kappa_h(s) := \frac{2H}{\tau - \tau_h(s)} + 1$*, then with probability at least* $1 - 3\delta$*, it holds that* $R_K \leq 2H\sqrt{T\log(\frac{dT}{\delta})} + 2(1+\kappa)\beta\sqrt{2dHT\log\left(1 + \frac{K}{d\lambda}\right)}$*, where* $\kappa := \max_{(s,h)\in\mathcal{S}\times[H]} \kappa_h(s)$.

See Appendix B.2.2 for the proof.

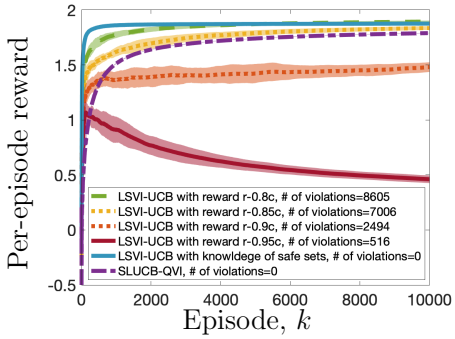
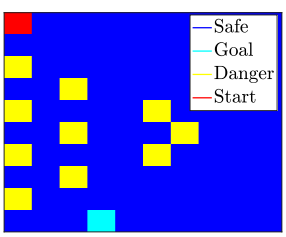
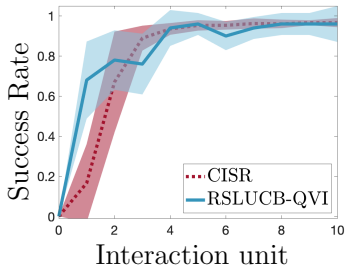


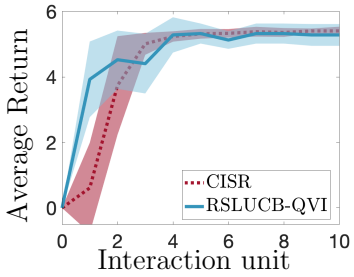
Figure 3.1: Comparison of SLUCB-QVI to the unsafe state-of-the-art verifying that: 1) when LSVI-UCB [64] has knowledge of $\boldsymbol{\gamma}_h^*$, it outperforms SLUCB-QVI (without knowledge of $\boldsymbol{\gamma}_h^*$) as expected; 2) when LSVI-UCB does not know $\boldsymbol{\gamma}_h^*$ (as is the case for SLUCB-QVI) and its goal is to maximize $r - \lambda' c$ instead of r, larger λ' leads to smaller per-episode reward and number of constraint violations while the number of constraint violations for SLUCB-QVI is zero.



(a) Map



(b) Success rate



(c) Average return

Figure 3.2: Comparison of RSLUCB-QVI and CISR [127] in Frozen Lake environment.

3.7 Experiments

In this section, we present numerical simulations [1] to complement and confirm our theoretical findings. We evaluate the performance of SLUCB-QVI on synthetic environments and implement RSLUCB-QVI on the *Frozen Lake* environment from OpenAI Gym [32].

3.7.1 SLUCB-QVI on Synthetic Environments

The results shown in Figure 3.1 depict averages over 20 realizations, for which we have chosen $\delta = 0.01$, $\sigma = 0.01$, $\lambda = 1$, $d = 5$, $\tau = 0.5$, $H = 3$ and $K = 10000$. The parameters $\{\boldsymbol{\theta}_h^*\}_{h\in[H]}$ and $\{\boldsymbol{\gamma}_h^*\}_{h\in[H]}$ are drawn from $\mathcal{N}(0, I_d)$. In order to tune parameters $\{\boldsymbol{\mu}_h^*(.)\}_{h\in[H]}$ and the feature map $\boldsymbol{\phi}$ such that they are compatible with Assumption 4, we consider that the feature space $\{\boldsymbol{\phi}(s,a) : (s,a) \in \mathcal{S}\times\mathcal{A}\}$ is a subset of the d-dimensional simplex and $\mathbf{e}_i^\top \boldsymbol{\mu}_h^*(.)$ is an arbitrary probability measure over $\mathcal{S}$ for all $i \in [d]$. This guarantees that Assumption 4 holds.

Computing safe sets $\mathcal{A}_h^k(s)$ in the first loop of SLUCB-QVI (Line 6), is followed by selecting an action that maximizes a linear function (in feature map $\boldsymbol{\phi}$) over the feature space $\mathcal{D}_h^k(s_h^k) := \left\{\boldsymbol{\phi}(s_h^k, a) : a \in \mathcal{A}_h^k(s_h^k)\right\}$ in its second loop (Line 10). Unfortunately, even if the feature space $\{\boldsymbol{\phi}(s,a) : (s,a) \in \mathcal{S}\times\mathcal{A}\}$ is convex, the set $\mathcal{D}_h^k(s_h^k)$ can have a form over which maximizing the linear function is intractable. In our experiments, we define map $\boldsymbol{\phi}$ such that the sets $\mathcal{D}(s)$ are star convex and *finite* around $\boldsymbol{\phi}\left(s, a_0(s)\right)$ with $N = 100$ (see Definition 1) and therefore, we can show that the optimization problem in Line 10 of SLUCB-QVI can be solved efficiently (see Appendix B.3 for a proof).

Definition 1 (Finite star convex set)**.** *A star convex set $\mathcal{D}$ around $\mathbf{x}_0 \in \mathbb{R}^d$ is finite, if there exist finitely many vectors $\{\mathbf{x}_i\}_{i=1}^N$ such that $\mathcal{D} = \cup_{i=1}^N [\mathbf{x}_0, \mathbf{x}_i]$, where $[\mathbf{x}_0, \mathbf{x}_i]$ is the line connecting $\mathbf{x}_0$ and $\mathbf{x}_i$.*

Figure 3.1 depicts the average per-episode reward of SLUCB-QVI and compares it to that of baseline and emphasizes the value of SLUCB-QVI in terms of respecting the safety constraints at all time-steps. Specifically, we compare SLUCB-QVI with 1) LSVI-UCB [64]

[1] All the experiments are implemented in Matlab on a 2020 MacBook Pro with 32GB of RAM.

when it has knowledge of safety constraints, i.e., $\boldsymbol{\gamma}_h^*$; and 2) LSVI-UCB, when it does not know $\boldsymbol{\gamma}_h^*$ (as is the case for SLUCB-QVI) and its goal is to maximize the function $r - \lambda' c$, with the constraint being pushed into the objective function, for different values of $\lambda' = 0.8, 0.85, 0.9$ and 0.95. Thus, playing costly actions is discouraged via low rewards. The plot verifies that LSVI-UCB with knowledge of $\boldsymbol{\gamma}_h^*$ outperforms SLUCB-QVI without knowledge of $\boldsymbol{\gamma}_h^*$ as expected. Also, larger λ' leads to smaller per-episode reward and number of constraint violations when LSVI-UCB seeks to maximize $r - \lambda' c$ (without knowledge of $\boldsymbol{\gamma}_h^*$) while the number of constraint violations for SLUCB-QVI is zero.

3.7.2 RSLUCB-QVI on Frozen Lake Environment

We evaluate the performance of RSLUCB-QVI in the Frozen Lake environment. The agent seeks to reach a goal in a 10×10 2D map (Figure 3.2a) while avoiding dangers. At each time step, the agent can move in four directions, i.e., $\mathcal{A} = \{a_1 : \text{left}, a_2 : \text{right}, a_3 : \text{down}, a_4 : \text{up}\}$. With probability 0.9 it moves in the desired direction and with probability 0.05 it moves in either of the orthogonal directions. We set $H = 1000$, $K = 10$, $d = |\mathcal{S}| = 100$, and $\boldsymbol{\mu}^*(s) \sim \mathcal{N}(0, I_d)$ for all $s \in \mathcal{S} = \{s_1, \ldots, s_{100}\}$. We then properly specified the feature map $\boldsymbol{\phi}(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ by solving a set of linear equations such that the transition specifics of the environment explained above are respected. In order to interpret the requirement of avoiding dangers as a constraint of form (3.11), we tuned $\boldsymbol{\gamma}^*$ and τ as follows: the cost of playing action $a \in \mathcal{A}$ at state $s \in \mathcal{S}$ is the probability of the agent moving to one of the danger states. Therefore a safe policy ensures that the expected value of probability of moving to a danger state is a small value. To this end, we set $\boldsymbol{\gamma}^* = \sum_{s \in \text{ Danger states}} \boldsymbol{\mu}^*(s)$ and $\tau = 0.1$. Also, for each state $s \in \mathcal{S}$ a safe action, playing which leads to one of the danger states with small probability ($\tau = 0.1$) is given to the agent. We solve a set of linear equations to tune $\boldsymbol{\theta}^*$ such that at each state $s \in \mathcal{S}$, the direction which leads to a state that is closest to the goal state gives the agent a reward 1, while playing other three directions gives it a reward 0.01. This model persuades the agent to move towards to the goal.

After specifying the feature map $\boldsymbol{\phi}$ and tuning all parameters, we implemented RSLUCB-

QVI for 10 interaction units (episodes) i.e, $K = 10$) each consisting of 1000 time-steps (horizon), i.e., $H = 1000$). During each interaction unit (episode) and after each move, the agent can end up in one of three kinds of states: 1) goal, resulting in a successful termination of the interaction unit; 2) danger, resulting in a failure and the consequent termination of the interaction unit; 3) safe. The agent receives a return of 6 for reaching the goal and 0.01 otherwise.

In Figure 3.2, we report the average of success rate and return over 20 agents for each of which we implemented RSLUCB-QVI 10 times and compare our results with that of CISR proposed by [127] in which a teacher helps the agent in selecting safe actions by making interventions. While the performances of both approaches, RSLUCB-QVI and CISR, are fairly comparable, an important point to consider is that each interaction unit (episode) in CISR consists of 10000 time-steps whereas this number is 1000 in RSLUCB-QVI. Notably, the learning rate of RSLUCB-QVI is faster than that of CISR. Also it is noteworthy that we comparedRSLUCB-QVI with CISR when it uses the *optimized* intervention, which gives the best results compared to other types of intervention.

3.8 Summary

In this chapter, we developed SLUCB-QVI and RSLUCB-QVI, two safe RL algorithms in the setting of finite-horizon linear MDP. For these algorithms, we provided sub-linear regret bounds $\tilde{\mathcal{O}}\left(\kappa\sqrt{d^3H^3T}\right)$, where H is the duration of each episode, d is the dimension of the feature mapping, κ is a constant characterizing the safety constraints, and $T = KH$ is the total number of action plays. We proved that with high probability, they never violate the unknown safety constraints. Finally, we implemented SLUCB-QVI and RSLUCB-QVI on synthetic and Frozen Lake environments, respectively, which confirms that our algorithms have performances comparable to that of state-of-the-art that either have knowledge of the safety constraint or take advantage of a teacher's advice helping the agent avoid unsafe actions.

CHAPTER 4

Doubly Pessimistic Algorithms for Strictly Safe Off-Policy Optimization

4.1 Introduction

Offline/Batch reinforcement learning (RL) as a method that uses previously collected datasets in many real-world decision-making applications where obtaining new experiences is costly has received significant attention [76]. For example, the outcome of a treatment in clinical trials can be evaluated only after several years and thus, a bad decision can cause long-term damages. The main focus of offline RL has been on two directions: 1) offline policy evaluation, which aims at estimating value functions of a target policy, and 2) offline policy optimization, which aims to find an optimal policy that maximizes the expected cumulative reward. A key challenge in offline RL is to address the issue of insufficient coverage in the dataset [131] due to the lack of exploration in data collecting process. There has been a surge of research activities investigating appropriate conditions on the data collecting process to guarantee an efficient and successful learning either in policy evaluation or policy optimization regions. For example, see [45, 143, 153, 147, 146].

Most of the existing offline RL methods in the more challenging category of offline policy optimization find a policy that under certain coverage assumptions performs well or at least as well as the behavior policy based on which the available dataset has been collected [76, 51, 71, 150, 101, 65, 72]. However, the learned policy in the above-mentioned works explores all possible actions, even though freely exploring all actions may be harmful in many real-world systems where playing even one unsafe action may lead to catastrophic results. Safety in RL has become increasingly important in recent years. Yet, many of existing

solutions fail to strictly avoid choosing unsafe policies, which may lead to catastrophic results in safety-critical systems. Thus, safety in offline RL has become a serious issue that has restricted the applicability of offline RL algorithms to many risk-sensitive real-world systems. For example, in a self-driving car, it is critical to only explore those policies that avoid crash and damage to the car, people and property. Switching cost limitations in medical applications [27] and legal restrictions in financial managements [3] are other examples of safety-critical applications. All the aforementioned safety-critical environments introduce the new challenge of balancing the goal of reward maximization with the restriction of playing safe actions and studying the influence of safety constraints in the sample complexity of finding an optimal safe policy.

4.1.1 Key Contributions

We study offline reinforcement learning (RL) in the presence of safety requirements: from a dataset collected a priori and without direct access to the true environment, learn an optimal policy that is guaranteed to respect the safety constraints. We focus on a strong notion of safety requirement which is modeled as an unknown cost function of states and actions, whose expected value with respect to the learned policy must fall below a certain threshold at each time-step an action is played with high probability. We present an algorithm in the context of finite-horizon Markov decision processes (MDPs), termed Safe-DPVI that performs in a doubly pessimistic manner when 1) it constructs a conservative set of safe policies; and 2) when it selects a good policy from that conservative set. Without assuming the sufficient coverage of the dataset or any structure for the underlying MDPs, we establish a data-dependent upper bound on the suboptimality gap of the *safe* policy Safe-DPVI returns. We then specialize our results to linear MDPs with appropriate assumptions on dataset being well-explored. Both data-dependent and specialized bounds nearly match that of state-of-the-art unsafe offline RL algorithms, with an additional multiplicative factor $\frac{\sum_{h=1}^{H} \alpha_h}{H}$, where α_h characterizes the safety constraint at time-step h. We further present numerical simulations that corroborate our theoretical findings.

4.2 Problem Statement

In this section, we first introduce the standard episodic Markov decision process (MDP) which is augmented by an extra safety/cost function and describe the data collecting process based on a *behavior* policy in the underlying MDP. Then, we introduce safety constraint which must be satisfied at all time-steps that actions are played with high probability. Finally, we introduce the performance metric.

Episodic Markov decision process. We consider an episodic Markov decision process (MDP) denoted by $M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, R, C)$, where $\mathcal{S}$ is the state set, $\mathcal{A}$ is the action set, H is the length of each episode (horizon), $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ are the transition probabilities, $R = \{R_h\}_{h=1}^H$ are the reward functions, and $C = \{C_h\}_{h=1}^H$ are the safety/cost functions, where $R_h : \mathcal{S} \times \mathcal{A} \to [0,1]$ and $C_h : \mathcal{S} \times \mathcal{A} \to [0,1]$. For each time-step $h \in [H]$, $\mathbb{P}_h(s'|s,a)$ denotes the probability of transitioning to state s' upon playing action a at state s. At each time-step $h \in [H]$, the agent observes the state s_h, plays an action $a_h \in \mathcal{A}$, and observes the next state $s_{h+1} \sim \mathbb{P}_h(.|s_h, a_h)$, a reward $r_h := R_h(s_h, a_h) + \eta_h$, and a cost $c_h := C_h(s_h, a_h) + \epsilon_h$, where η_h and ϵ_h are random additive noise. We consider a learning problem, where $\mathcal{S}$ and $\mathcal{A}$ are known, while the transition probabilities $\mathbb{P}_h$, rewards R_h and costs C_h are *unknown* to the agent and must be learned from a given dataset $\mathcal{D}$. The dataset $\mathcal{D} := \left\{s_h^k, a_h^k, r_h^k, c_h^k\right\}_{h,k=1}^{H,K}$ is collected from K i.i.d. trajectories under a *behavior* policy denoted as $\bar{\pi}$.

Safety Constraint. We assume that the underlying system is safety-critical and the environment is subject to a side constraint that restricts the choice of policies. A policy $\pi := \{\pi_h\}_{h=1}^H$, where $\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}}$ maps $\mathcal{S}$ to distributions over $\mathcal{A}$, is called *safe* if

$$\mathbb{E}_{a \sim \pi_h(.|s)} \left[C_h(s,a)\right] \leq \tau, \ \forall (s,h) \in \mathcal{S} \times [H] \tag{4.1}$$

with high probability. We accordingly define the *unknown* set of safe policies by $\Pi^{\text{safe}} := \left\{\pi : \pi_h(.|s) \in \Gamma_h^{\text{safe}}(s), \ \forall (s,h) \in \mathcal{S} \times [H]\right\}$, where

$$\Gamma_h^{\text{safe}}(s) := \left\{\theta(.|s) \in \Delta_{\mathcal{A}} : \mathbb{E}_{a \sim \theta(.|s)} \left[C_h(s,a)\right] \leq \tau\right\}. \tag{4.2}$$

Thus, after observing state s_h at time-step $h \in [H]$, the agent's choice of policy must belong to $\Gamma_h^{\text{safe}}(s_h)$ with high probability. As a motivating example, consider a self-driving car. On the one hand, the agent (car) is rewarded for getting from point one to point two as fast as possible. On the other hand, the driving behavior must be constrained to respect traffic safety standards.

Performance Metric. We define the state-action and state value function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $V_h^\pi : \mathcal{S} \to \mathbb{R}$ for a policy π at time-step $h \in [H]$ by

$$Q_h^\pi(s,a) := \mathbb{E}\left[\sum_{h'=h+1}^{H} r_{h'}\left(s_{h'}, a_{h'}\right) \middle| s_h = s, a_h = a, \pi\right], V_h^\pi(s) := \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}\left(s_{h'}, a_{h'}\right) \middle| s_h = s, \pi\right],$$

where the expectation is over the environment and the randomness of policy π. To simplify the notation, for any function f, we denote $[\mathbb{P}_h f](s,a) := \mathbb{E}_{s' \sim \mathbb{P}_h(.|s,a)} f(s')$ and $[\mathbb{B}_h f](s,a) := R_h(s,a) + [\mathbb{P}_h f](s,a)$. Let π_* be the optimal *safe* policy such that $V_h^{\pi_*}(s) := V_h^*(s) = \sup_{\pi \in \Pi^{\text{safe}}} V_h^\pi(s)$ for all $(s,h) \in \mathcal{S} \times [H]$. Thus, for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, the Bellman equations for the optimal safe policy and an arbitrary policy $\pi \in \Pi^{\text{safe}}$ are:

$$Q_h^*(s,a) = [\mathbb{B}_h V_{h+1}^*](s,a), \quad V_h^*(s) = \max_{\theta(.|s) \in \Gamma_h^{\text{safe}}(s)} \mathbb{E}_{a \sim \theta(.|s)}\left[Q_h^*(s,a)\right] \tag{4.3}$$

$$Q_h^\pi(s,a) = [\mathbb{B}_h V_{h+1}^\pi](s,a), \quad V_h^\pi(s) = \mathbb{E}_{a \sim \pi_h(.|s)}\left[Q_h^\pi(s,a)\right], \tag{4.4}$$

where $V_{H+1}^\pi(s) = V_{H+1}^*(s) = 0$. Our goal is to learn a *safe* policy that maximizes the cumulative expected reward given the collected dataset $\mathcal{D}$. To this end, we define the following suboptimality gap of a safe policy π given by the initial state $s_1 = s$ as

$$\Delta(\pi; s) := V_1^*(s) - V_1^\pi(s). \tag{4.5}$$

4.3 Prior Work

In online setting, the problem of Safe RL formulated with Constrained Markov Decision Process (CMDP) is studied in [48, 127, 54, 154, 42, 100, 43, 141, 67, 88]. In the above-mentioned papers, the goal is to find the optimal policy in an online manner that maximizes

the reward value function $V_r^\pi(s)$ (expected total reward) over the safe policies that satisfy $V_c^\pi(s) \leq b$, where $V_c^\pi(s)$ is the cumulative expected cost over an entire episode with duration H and b is a threshold. This safety requirement is defined over an *entire* episode, and consequently is less strict than the safety requirement considered in this work, which must be satisfied at each time-step an action is played. The notion of safety has been used in several existing offline RL works. However, they fundamentally differ from the definition of safety considered in our work. For example, safety in [77, 123, 55] means the algorithm returns a policy with performance at least as good as that of behavior/baseline policy, based on which the dataset has been collected. In another line of work, [111, 122] empirically study safety-constrained RL problem and propose algorithms that consist of two distinct offline and online phases and aim to find an optimal policy for which the expected value of the number of unsafe states visits is less than some threshold $\epsilon \in (0, 1)$ in the context of discounted MDPs with discount factor γ. In the offline phase, they estimate the safe set of policies form an available dataset, and then in the online phase, they seek to find the best policy based on the estimated safe set of policies. The definition of safety constraints studied in all the above-stated papers is a special case of the notion of safety considered in our work. For example, if $C(s, a)$ and τ in (4.1) are set to be the probability of transitioning to an unsafe state by playing action a at state s and $\epsilon(1 - \gamma)$ would recover the safety constraint in [122] for infinite-horizon discounted MDPs. Furthermore, in all the online safe papers, having $\tau = b/H$ in the definition of safety constraint considered in our work in (4.1) would recover $V_c^\pi(s) \leq b$. Therefore, the safety requirement considered in our work is much stricter than those in the existing literature, and naturally covers a wider range of applications.

4.4 Safe-DPVI: A General Framework for Safe Offline Policy Optimization

In this section, we formally present Safe Doubly Pessimistic Value Iteration (Safe-DPVI), summarized in Algorithm 5, that employs the dataset and returns a *safe* policy $\hat{\pi}$. We then introduce two uncertainty quantifiers based on which, we are able to control $\Delta(\hat{\pi}; s)$ and

state our main results on the suboptimality gap's bound in Theorem 7.

First, we introduce the following assumption, which is necessary to ensure that the safety constraint in (4.1) is satisfied from the very first time-step.

Assumption 9 (Non-empty safe sets)**.** *There exists a known safe policy π^0 with known costs $\tau_h(s) := \mathbb{E}_{a\sim\pi_h^0(.|s)}\left[C_h(s,a)\right] < \tau$. Thus, the sets $\Gamma_h^{safe}(s)$ are non-empty, as $\pi_h^0(s) \in \Gamma_h^{safe}(s)$.*

This assumption is rather standard and has been widely used in the literature of safe online RL [15, 88] and safe bandits [11, 97]. This assumption is also realistic in many practical examples, where the known safe policy could be the one suggested by the current strategy of the company or a very cost-neutral policy that does not necessarily have high reward but its cost is far from the threshold. Note that the known safe policy π^0 is not necessarily the same as the behavior policy $\bar{\pi}$. If the behavior policy $\bar{\pi}$ is also safe, we can simply treat it as the known safe policy, i.e., $\bar{\pi} = \pi^0$. In Appendix C.3, we show that it is possible to relax the assumption of knowing the costs of the safe policy $\tau_h(s)$ and when $\pi^0 = \bar{\pi}$, this relaxation naturally goes through.

Algorithm 5 Safe Doubly Pessimistic Value Iteration

Input: $\mathcal{D} = \left\{s_h^k, a_h^k, r_h^k, c_h^k\right\}_{h,k=1}^{H,K}$

Initialization: $\hat{V}_{H+1}(s) = 0,\ \forall s \in \mathcal{S}$

for time-steps $h = H, \dots, 1$ **do**

  Compute $[\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a)$, $B'_h(s,a)$ and $\hat{\Gamma}_h(s)$, $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$, as defined in Section 4.5.1 for underlying linear MDP.

  Set $\hat{Q}_h(s,a) = \left\{[\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a) - B'_h(s,a)\right\}^+,\ \forall (s,a) \in \mathcal{S}\times\mathcal{A}$.

  Set $\hat{\pi}_h(.|s) = \arg\max_{\theta(.|s)\in\hat{\Gamma}_h(s)} \mathbb{E}_{a\sim\theta(.|s)}\left[\hat{Q}_h(s,a)\right],\ \forall s \in \mathcal{S}$.

  Set $\bar{V}_h(s) = \mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[\hat{Q}_h(s,a)\right],\ \hat{V}_h(s) = \min\left\{\bar{V}_h(s), H\right\},\ \forall s \in \mathcal{S}$.

end for

Output: $\hat{\pi} = \{\hat{\pi}_h\}_{h=1}^H$

4.4.1 Overview

From a high-level point of view, based on dataset $\mathcal{D}$, Safe-DPVI constructs estimated cost functions $\hat{C}_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, Q-functions $\hat{Q}_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, value functions $\hat{V}_h : \mathcal{S} \to \mathbb{R}$, and Bellman operator $\hat{\mathbb{B}}_h$ such that $[\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a)$ approximates $[\mathbb{B}_h \hat{V}_{h+1}](s,a)$. Note that the algorithm only relies on construction of $[\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a)$ not $\hat{\mathbb{B}}_h$ itself. The algorithm constructs an estimated set of safe policies $\hat{\Pi}$ based on estimated cost functions $\hat{C}_h$. To see how this happens, we first define the following δ-safety uncertainty quantifier and δ-Bellman uncertainty quantifier for $\delta \in (0,1)$ that quantify the uncertainty arising from approximating the cost function C and $[\mathbb{B}_h \hat{V}_{h+1}](s,a)$, respectively.

Definition 2 (Uncertainty quantifiers)**.** *For a fixed $\delta \in (0,1)$, we call $B = \{B_h\}_{h=1}^H$ with $B_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, a δ-safety uncertainty quantifier if $\mathbb{P}\left(\left|C_h(s,a) - \hat{C}_h(s,a)\right| \leq B_h(s,a), \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]\right) \geq 1 - \delta$. We also call $B' = \{B'_h\}_{h=1}^H$ with $B'_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, a δ-Bellman uncertainty quantifier if $\mathbb{P}\left(\left|[\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a) - [\mathbb{B}_h \hat{V}_{h+1}](s,a)\right| \leq B'_h(s,a), \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]\right) \geq 1 - \delta$.*

Thus, if the agent can compute a δ-safety uncertainty quantifier B based on the dataset $\mathcal{D}$ and $u_h^c(s,a) = \hat{C}_h(s,a) + B_h(s,a)$, then, a natural approximation for Π^{safe} is $\hat{\Pi} := \left\{\pi : \pi_h(.|s) \in \hat{\Gamma}_h(s),\ \forall (s,h) \in \mathcal{S} \times [H]\right\}$, where

$$\hat{\Gamma}_h(s) := \left\{\theta(.|s) \in \Delta_{\mathcal{A}} : \mathbb{E}_{a \sim \theta(.|s)}\left[u_h^c(s,a)\right] \leq \tau\right\}. \tag{4.6}$$

Thus, Safe-DPVI constructs $\hat{\Gamma}_h(s)$ pessimistically as it relies on $u_h^c(s,a)$, which is an upper confidence bound on $C_h(s,a)$.

Next time Safe-DPVI applies pessimism is when it computes $\hat{Q}_h$ by incorporating δ-Bellman uncertainty quantifier B' into the value iteration step as follows

$$\hat{Q}_h(s,a) = \left\{[\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a) - B'_h(s,a)\right\}^+. \tag{4.7}$$

After the construction of $\hat{\Gamma}_h$ and $\hat{Q}_h$, the algorithm is ready to return the *safe* policy $\hat{\pi} = \{\hat{\pi}_h\}_{h=1}^H$, where $\hat{\pi}_h(.|s) = \arg\max_{\theta(.|s) \in \hat{\Gamma}_h(s)} \mathbb{E}_{a \sim \theta(.|s)}\left[\hat{Q}_h(s,a)\right]$, as its output. In the following theorem, we characterize the safeness and suboptimality gap of Safe-DPVI.

Theorem 7. *Fix $\delta \in (0, 0.5)$. Let B and B' be δ-safety uncertainty quantifier and δ-Bellman uncertainty quantifier, respectively, $\hat{\pi}$ be the output of Algorithm 5, $\alpha_h = 2 + \frac{2H}{\tau - \max_{s\in\mathcal{S}} \tau_h(s)}$ and $\bar{B}_h(s,a) := \max\left\{B_h(s,a), B'_h(s,a)\right\}$. Then, under Assumption 9, if $\mathbb{E}_{a\sim\pi_h^0(.|s)}\left[B_h(s,a)\right] \leq \frac{\tau - \tau_h(s)}{2}$ for all $(s,h) \in \mathcal{S}\times[H]$, then with probability at least $1-2\delta$, it holds that 1. $\hat{\Pi}$ includes π^0 and therefore it is non-empty and $\hat{\pi}$ is safe; 2. $\Delta(\hat{\pi}; s) \leq \max\left\{\sum_{h=1}^H \alpha_h \mathbb{E}\left[\bar{B}_h(s_h,a_h)|s_1 = s, \pi^*\right], \sum_{h=1}^H \alpha_h \mathbb{E}\left[\bar{B}_h(s_h,a_h)|s_1 = s, \pi^0\right]\right\}$.*

The complete proof is given in Appendix C.1.3.

Now, we comment on the suboptimality gap of Algorithm 5 and how it compares to that of its *unsafe* counterpart in [65]. The bound on PEVI's suboptimality gap in [65] is $2\sum_{h=1}^H \mathbb{E}\left[B'_h(s_h,a_h)|s_1 = s, \pi^*\right]$. We observe that our bound is comparable with that of PEVI with the following differences: 1) Instead of B'_h, our bound includes $\alpha_h \bar{B}_h$ to account for the uncertainty regarding the additional unknown safety constraints we have to deal with in our setting; 2) Moreover, we take the maximum of the expected value of the uncertainty of trajectories induced by both the optimal safe policy π^* and the known safe policy π^0, which once again highlights the role of the known safe policy in Safe-DPVI's performance.

4.4.2 Proof Sketch of Theorem 7

While point 1 is directly proven from the definition of δ-safety uncertainty quantifier B in Definition 2, the proof of point 2 is more intricate and challenging and uses the following two key lemmas whose proofs are given in Appendixes C.1.1 and C.1.2.

Lemma 5 (Suboptimality Gap's Upper Bound Decomposition)**.** *Consider a meta-algorithm that employs the dataset to construct an estimated Q-function $\hat{Q}_h : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ and an estimated value function $\hat{V}_h : \mathcal{S} \to \mathbb{R}$. Let $\iota_h(s,a) := [\mathbb{B}_h \hat{V}_{h+1}](s,a) - \hat{Q}_h(s,a)$ be the model evaluation error and $\hat{\pi}$ be a policy such that $\hat{V}_h(s) = \min\left\{\mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[\hat{Q}_h(s,a)\right], H\right\}$ for all $(s,h) \in \mathcal{S}\times[H]$. Then, it holds that*

$$\Delta(\hat{\pi}; s) \leq \underbrace{V_1^*(s) - \hat{V}_1(s)}_{\textit{Term I}} + \underbrace{\sum_{h=1}^H \mathbb{E}\left[-\iota_h(s_h,a_h)\middle| s_1 = s, \hat{\pi}\right]}_{\textit{Term II}}.$$

Recall that $\iota_h(s,a) := [\mathbb{B}_h \hat{V}_{h+1}](s,a) - \hat{Q}_h(s,a)$. Thus, ι_h and $\hat{\pi}$ are correlated as they both depend on the dataset $\mathcal{D}$ and thus, the expectation in Term II can be rather large. The definition of δ-Bellman uncertainty quantifier B' and the pessimism in computation of $\hat{Q}_h(s,a)$ helps us eliminate Term II. Note that if $[\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a) - B'_h(s,a) < 0$, then $\hat{Q}_h(s,a) = 0$ and therefore $-\iota_h(s,a) = -[\mathbb{B}_h \hat{V}_{h+1}](s,a) \leq 0$ as $\hat{V}_h(s) \geq 0$ for all $(s,h) \in \mathcal{S} \times [H]$. Now, suppose $[\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a) - B'_h(s,a) \geq 0$. Since B' is a δ-Bellman uncertainty quantifier, we have

$$-\iota_h(s,a) = \hat{Q}_h(s,a) - [\mathbb{B}_h \hat{V}_{h+1}](s,a) = [\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a) - B'_h(s,a) - [\mathbb{B}_h \hat{V}_{h+1}](s,a) \leq 0.$$

This concludes that for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, with probability at least $1-\delta$, it holds that $-\iota_h(s,a) \leq 0$, and therefore Term II $= \sum_{h=1}^{H} \mathbb{E}\left[-\iota_h(s_h,a_h) \middle| s_1 = s, \hat{\pi}\right] \leq 0$.

Our main technical contribution towards bounding $\Delta(\hat{\pi}; s)$ is given in Lemma 6 that together with Lemma 5 proves point 2 of Theorem 7.

Lemma 6. *Fix $\delta \in (0, 0.5)$. Let B and B' be δ-safety uncertainty quantifier and δ-Bellman uncertainty quantifier, respectively. Also, let $\bar{B}_h(s,a) = \max\left\{B_h(s,a), B'_h(s,a)\right\}$ and $F_h(s) := \max\left\{\sum_{h'=h}^{H} \alpha_{h'} \mathbb{E}\left[\bar{B}_{h'}(s_{h'}, a_{h'}) | s_h = s, \pi^*\right], \sum_{h'=h}^{H} \alpha_{h'} \mathbb{E}\left[\bar{B}_{h'}(s_{h'}, a_{h'}) | s_h = s, \pi^0\right]\right\}$. Then, under Assumption 9 and provided that $\alpha_h = 2 + \frac{2H}{\tau - \max_{s \in \mathcal{S}} \tau_h(s)}$, with probability at least $1-2\delta$, it holds that*

$$V_h^*(s) - \hat{V}_h(s) \leq F_h(s), \ \forall (s,h) \in \mathcal{S} \times [H]. \tag{4.8}$$

In safe off-policy optimization, the safe set Π^{safe} is not known. Therefore, at each time-step, the agent's policy must be chosen from a conservative inner approximation of Π^{safe}. Intuitively, the better this approximation is, the more likely that the output policy of Safe-DPVI leads to small suboptimality gap, ideally of the same order as that of PEVI proposed by [65] in the classical offline RL setting. In order to better highlight the challenging part of our analysis compared to classical setting without safety constraint, we observe that for all $(s,h) \in \mathcal{S} \times [H]$, with probability at least $1-\delta$, it holds that $V_h^*(s) - \hat{V}_h(s) \leq$ Term i + Term ii, where Term i $= \min\left\{\mathbb{E}_{a \sim \pi_h^*(.|s)}\left[\hat{Q}_h(s,a)\right], H\right\} - \min\left\{\mathbb{E}_{a \sim \hat{\pi}_h(.|s)}\left[\hat{Q}_h(s,a)\right], H\right\}$ and Term ii $=$

$\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[2B'_h(s,a) - \left[\mathbb{P}_h\left(\hat{V}_{h+1} - V^*_{h+1}\right)\right](s,a)\right]$. A key difference in the analysis of Safe-DPVI compared to the classical offline RL without safety constraint is that $\pi_h^*(.|s_h)$ may not lie within the estimated safe set $\hat{\Gamma}_h(s_h)$, which makes controlling Term i and Term ii more delicate. This complication lies at the heart of the new formulation with additional safety constraints. When safety constraints are absent, classical pessimistic offline RL algorithms such as PEVI in [65] guarantee that Term i is non-positive and by induction it can be shown that Term ii$\leq 2\sum_{h'=h}^{H}\mathbb{E}\left[B'_{h'}(s_{h'},a_{h'})|s_h=s,\pi^*\right]$. Unfortunately, this is not the case here as $\pi_h^*(.|s_h)$ does not necessarily belong to $\hat{\Gamma}_h(s_h)$, thus Term i can be positive, which also affects the bound on Term ii. This extra positive term in the suboptimality gap is the price paid by Safe-DPVI for choosing safe policies at each time-step $h \in [H]$.

4.5 Safe-DPVI: Linear MDP

In this section, we specialize Safe-DPVI and its theoretical guarantees to the case where the underlying MDP is linear [31, 142, 64]. We further determine sufficient conditions that allow us to derive finite sample complexity for Safe-DPVI with an underlying linear MDP.

Definition 3 (Linear MDP). *$M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, R, C)$ is a linear MDP with feature map $\boldsymbol{\phi} : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$, if for any $h \in [H]$, there exist d unknown measures $\boldsymbol{\mu}_h^* := [\mu_h^{*(1)}, \ldots, \mu_h^{*(d)}]^\top$ over $\mathcal{S}$, and unknown vectors $\boldsymbol{\theta}_h^*, \boldsymbol{\zeta}_h^* \in \mathbb{R}^d$ such that $\mathbb{P}_h(.|s,a) = \left\langle \boldsymbol{\mu}_h^*(.), \boldsymbol{\phi}(s,a)\right\rangle$, $R_h(s,a) = \left\langle \boldsymbol{\theta}_h^*, \boldsymbol{\phi}(s,a)\right\rangle$, and $C_h(s,a) = \left\langle \boldsymbol{\zeta}_h^*, \boldsymbol{\phi}(s,a)\right\rangle$.*

4.5.1 Overview

We introduce the quantities that Safe-DPVI constructs based on the dataset $\mathcal{D}$ when the underlying MDP is linear. Recall that $\hat{\Gamma}_h(s)$ in (4.6) depends on $\hat{C}_h(s,a)$, an approximation of $C_h(s,a)$, and $B_h(s,a)$. In particular, Safe-DPVI constructs $\hat{C}_h(s,a) = \left\langle \hat{\boldsymbol{\zeta}}_h, \boldsymbol{\phi}(s,a)\right\rangle$, where $\hat{\boldsymbol{\zeta}}_h := \arg\min_{\boldsymbol{\nu}\in\mathbb{R}^d}\sum_{k=1}^{K}\left(\left\langle \boldsymbol{\nu}, \boldsymbol{\phi}(s_h^k, a_h^k)\right\rangle - c_h^k\right)^2 + \lambda\|\nu\|_2^2$ is the least square estimator of $\boldsymbol{\zeta}_h^*$ with regularization parameter $\lambda > 1$ and has the closed form $\hat{\boldsymbol{\zeta}}_h := \boldsymbol{\Lambda}_h^{-1}\left(\sum_{k=1}^{K}\boldsymbol{\phi}(s_h^k, a_h^k).\ c_h^k\right)$,

where $\mathbf{\Lambda}_h = \lambda I + \sum_{k=1}^{K} \boldsymbol{\phi}(s_h^k, a_h^k)\boldsymbol{\phi}(s_h^k, a_h^k)^\top$. Moreover, Safe-DPVI computes

$$[\hat{\mathbb{B}}_h \hat{V}_{h+1}](s,a) = \left\langle \hat{\mathbf{w}}_h, \boldsymbol{\phi}(s,a) \right\rangle, \; B_h(s,a) = \beta \left\| \boldsymbol{\phi}(s,a) \right\|_{\mathbf{\Lambda}_h^{-1}}, \; B'_h(s,a) = \beta' \left\| \boldsymbol{\phi}(s,a) \right\|_{\mathbf{\Lambda}_h^{-1}}, \tag{4.9}$$

where $\hat{\mathbf{w}}_h$ is the minimizer of the empirical mean squared Bellman error (MSBE), with closed form

$$\hat{\mathbf{w}}_h := \mathbf{\Lambda}_h^{-1} \left(\sum_{k=1}^{K} \boldsymbol{\phi}(s_h^k, a_h^k). \left[r_h^k + \hat{V}_{h+1}(s_{h+1}^k) \right] \right), \tag{4.10}$$

and $\beta, \beta' > 0$ are scaling parameters that will be defined shortly in Theorem 8.

4.5.2 Theoretical Guarantees

Now, we specialize our results in Theorem 7 to the case of linear MDP and provide a sample complexity for Safe-DPVI when the underlying MDP is linear and certain conditions hold. First, we make the remaining necessary assumptions under which our proposed algorithm operates and achieves small suboptimality gap.

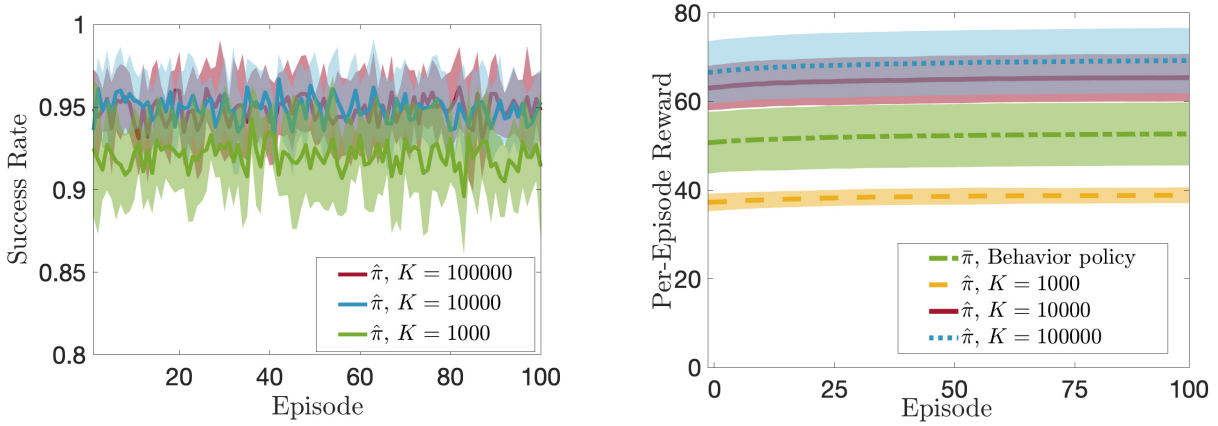
Assumption 10 (Subgaussian noise)**.** *For all $(h,k) \in [H] \times [K]$, η_h^k and ϵ_h^k are zero-mean σ-subGaussian random variables.*

Assumption 11 (Boundedness)**.** *Without loss of generality, $\left\| \boldsymbol{\phi}(s,a) \right\|_2 \leq 1$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $\max\left(\left\| \boldsymbol{\mu}_h^*(\mathcal{S}) \right\|_2, \|\boldsymbol{\theta}_h^*\|_2, \|\boldsymbol{\zeta}_h^*\|_2 \right) \leq \sqrt{d}$ for all $h \in [H]$.*

Assumption 12 (Well-explored dataset)**.** *There exists an absolute constant $\bar{c} > 0$ such that $\lambda_{min}(\Sigma_h) \geq \bar{c}, \; \forall h \in [H]$, where $\Sigma_h = \mathbb{E}_{\bar{\pi}}\left[\boldsymbol{\phi}(s_h, a_h)\boldsymbol{\phi}(s_h, a_h)^\top \right]$ and $\mathbb{E}_{\bar{\pi}}$ is the expectation taken with respect to the trajectory induced by behavior policy $\bar{\pi}$.*

Assumptions 10 and 11 are standard in linear MDP and bandit literature [64, 97, 11]. Assumption 12 is necessary to ensure that the data collecting process has sufficiently explored $\mathcal{A}$ and $\mathcal{S}$. This assumption is standard in the literature of offline policy optimization/evaluation; e.g., see [65, 45].

Given these assumptions, we are now ready to present the formal theoretical guarantees of Safe-DPVI, with underlying linear MDP defined in Definition 3, in the following theorem.



(a) Success Rate of $\hat{\pi}$ **(b)** Per-Episode Reward of $\hat{\pi}$, $\bar{\pi}$

Figure 4.1: Performance of Safe-DPVI with an underlying linear MDP on Inverted Pendulum. The shaded regions show standard deviation around the average over 100 realizations.

Theorem 8 (Suboptimality gap of Safe-DPVI: Linear MDP)**.** *Let the underlying MDP of Safe-DPVI be a linear MDP as stated in Definition 3, $\hat{\pi}$ be the output of Safe-DPVI and $\alpha_h = 2 + \frac{2H}{\tau - \max_{s\in\mathcal{S}} \tau_h(s)}$. Under Assumptions 9, 10, 11, and 12, if $K \geq \max\left\{\frac{8}{\bar{c}}\log(\frac{dH}{\delta}), \frac{8\beta^2}{\bar{c}\left(\tau - \max_{(s,h)\in\mathcal{S}\times[H]} \tau_h(s)\right)^2}\right\}$ and we set $\beta = \sigma\sqrt{d\log\left(\frac{2+\frac{2T}{\lambda}}{\delta}\right)} + \sqrt{\lambda d}$, $\beta' = cdH\sqrt{\log(\frac{dT}{\delta})}$ for an absolute constant $c > 0$, $\bar{\beta} = \max\{\beta, \beta'\}$, then for any fixed $\delta \in (0, 1/3)$, for all $s \in \mathcal{S}$, with probability at least $1 - 3\delta$, $\hat{\pi} \in \Pi^{safe}$ and $\Delta(\hat{\pi}; s) \leq \frac{\sqrt{2}\bar{\beta}\sum_{h=1}^{H}\alpha_h}{\sqrt{2\lambda + \bar{c}K}}$*

We observe that under the same wide-coverage assumption (Assumption 12), Safe-DPVI with underlying linear MDP achieves an upper bound on the suboptimality gap of the safe policy $\hat{\pi}$, which is nearly of the same order as $\frac{\sqrt{2}\beta' H}{\sqrt{2\lambda + \bar{c}K}}$ obtained for the state-of-the-art unsafe algorithm PEVI in [65]. The complete proof is reported in the Appendix C.2.

4.6 Experiments

In this section, we present numerical simulations to complement and confirm our theoretical findings. We apply Safe-DPVI to the control of a simulated *Inverted Pendulum* environment

from OpenAI Gym [32]. We consider a pendulum with mass $m = 1$, length $l = 1$, which is actuated by torque $u \in [-15, 15]$. The environment's state is described by the pendulum's angular position $\theta \in [-\pi, \pi]$ and its angular rate $\dot{\theta} \in [-5, 5]$. The system dynamics are defined as follows

$$\theta_{h+1} = \theta_h + \dot{\theta}_h \delta h + \frac{3g}{2l} \sin(\theta_t) \delta h^2 + \frac{3}{ml^2} u \delta h^2, \dot{\theta}_{h+1} = \dot{\theta}_h + \frac{3g}{2l} \sin(\theta_h) \delta h + \frac{3}{ml^2} u \delta h, \quad (4.11)$$

where $g = 9.8$ is the gravity constant and δh is the simulation step and we set it to 1.

For real numbers a and b and positive integer number n, let $\text{Disc}([a, b], n)$ be a discretized set formed of uniformly dividing $[a, b]$ into n intervals. We discretize the continuous state and action spaces and consider that $\mathcal{S} = \text{Disc}([-\pi, \pi], 10) \times \text{Disc}([-5, 5], 5)$ and $\mathcal{A} = \text{Disc}([-15, 15], 15)$. Thus, $|\mathcal{S}| = 50$ and $|\mathcal{A}| = 15$. For any $s \in \mathcal{S}$, let $s(1)$ and $s(2)$ be the corresponding pendulum's angular position and pendulum's angular rate.

We consider that the transition probability $\mathbb{P}$, the reward R, and the cost C do not vary during an episode. In order to induce stochasticity and parametrize $\mathbb{P}(s'|s, a)$, we assumed that when a torque a is chosen, an additive random torque affects it. In particular, we considered that $\mathbb{P}(s'|s, a) = 0.8$ for s' being the closest element of $\mathcal{S}$ to the next state of playing torque a at pendulum's angular position $s(1)$ and pendulum's angular rate $s(2)$ according to system's dynamics in (4.11). Moreover, $\mathbb{P}(s'|s, a) = 0.05$ for s' being the closest element of $\mathcal{S}$ to the next state of playing torque $a + i,\ i \in \{-6, -3, 3, 6\}$ at pendulum's angular position $s(1)$ and pendulum's angular rate $s(2)$ according to (4.11). We also let $R(s, a) = c - s(1)^2 + 0.1s(2)^2 + 0.001a^2$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, where c is a constant that makes the rewards positive, and divided them by $\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} R(s, a)$. This definition for the reward function encourages learning a controller that keeps the pendulum upright. We further defined the set of unsafe states as $\mathcal{S}^{\text{unsafe}} = \left\{s \in \mathcal{S} : s(1) \notin [-\pi/3, \pi/3]\right\}$ and specified $C(s, a) = \sum_{s'\in\mathcal{S}^{\text{unsafe}}} \mathbb{P}(s'|s, a)$, and $\tau = 0.01$. Therefore a safe policy ensures that the expected value of the probability of moving to an unsafe state is a small value ($\tau = 0.01$). Note that any tabular MDP with finitely many states and actions can be represented by a linear MDP. In particular, if we let $d = |\mathcal{S}||\mathcal{A}|$ and index each coordinate by an state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, then a linear MDP with $\boldsymbol{\phi}(s, a) = \mathbf{e}_{(s,a)}$, $\mathbb{P}_h(.|s, a) = \left\langle \boldsymbol{\mu}_h^*(.), \mathbf{e}_{(s,a)} \right\rangle$,

$R_h(s,a) = \left\langle \boldsymbol{\theta}_h^*, \mathbf{e}_{(s,a)} \right\rangle$, and $C_h(s,a) = \left\langle \boldsymbol{\zeta}_h^*, \mathbf{e}_{(s,a)} \right\rangle$, recovers the tabular MDP. As such, for the inverted pendulum environment with the above-stated tabular MDP, we considered an underlying linear MDP with standard basis vectors of dimension $d = |\mathcal{S}||\mathcal{A}|$ as its feature maps and episode length $H = 100$.

The performance of Batch RL algorithms can vary greatly from one dataset to another. To properly assess Safe-DPVI, we repeated the following for 100 times: 1) fixed a randomly selected safe behavior policy, $\bar{\pi}$, used in the data collecting process, and created datasets with size $K = 1000$, $K = 10000$, and $K = 100000$, on Inverted Pendulum environment discussed above; 2) implemented Safe-DPVI on each of these three datasets, and employed the output policies for 100 episodes with randomly selected initial state; 3) reported the per-episode reward and success rate, which is the number of time-steps the pendulum was in safe states during an episode divided by the duration of each episode $H = 100$, for each of the output policies. The results shown in Figure 4.1 depict averages over these 100 realizations, for which we have chosen $\delta = 0.01$, $\sigma = 0.05$, $\lambda = 1$. In this figure, we have numerically confirmed the result of Theorem 8. Figure 4.1a showcases that the rate of unsafe states visits is low (success rate is high) and therefore the output policy $\hat{\pi}$ is safe with high probability. Figure 4.1b confirms that $\hat{\pi}$, for sufficiently large datasets that satisfy wide-coverage assumption (see Assumption 12), performs near-optimally and better than the behavior policy $\bar{\pi}$.

4.7 Summary

In this chapter, we developed Safe-DPVI, a safe offline RL algorithm in the setting of episodic MDPs, that performs in a pessimistic manner when 1) it constructs a conservative set of safe policies; and 2) when it selects a good policy from that conservative set in the value iteration step. We guaranteed that Safe-DPVI outputs a policy $\hat{\pi}$ which is strictly safe in the sense that it respects the safety constraint at each time-step that it suggests an action to be played with high probability. Without assuming the sufficient coverage of the dataset or any structure for the underlying MDPs, we first established a data-dependent upper bound on the suboptimality gap of the *safe* policy Safe-DPVI returns. Then, we specialized our results to

linear MDPs with appropriate assumptions on dataset being well-explored and proved a high probability upper bound on the suboptimality gap of $\hat{\pi}$, i.e., $\Delta(\hat{\pi}; s) \leq \frac{\sqrt{2}\bar{\beta}\sum_{h=1}^{H}\alpha_h}{\sqrt{2\lambda+\bar{c}K}},\ \forall s \in \mathcal{S}$, which is order-wise comparable to those of its unsafe counter-parts. Finally, we implemented Safe-DPVI on Inverted Pendulum environment to empirically confirm our theoretical findings.

CHAPTER 5

Provably Efficient Lifelong Reinforcement Learning with Linear Representation

5.1 Introduction

Recently, there has been a surging interest in designing *lifelong learning* agents that can continuously learn to solve multiple sequential decision making problems in their lifetimes [124, 69, 108, 140]. This scenario is in particular motivated by building multi-purpose embodied intelligence, such as robots working in a weakly structured environment [102]. Typically, curating all tasks beforehand for such problems is nearly infeasible, and the problems the agent is tasked with may be adaptively selected based on the agent's past behaviors. Consider a household robot as an example. Since each household is unique, it is difficult to anticipate upfront all scenarios the robot would encounter. Moreover, the tasks the robot faces are not independent and identically distributed (i.i.d.). Instead, what the robot has done before can affect the next task and its starting state; e.g., if the robot fails to bring a glass of water and breaks it, then the user is likely to command the robot to clean up the mess. Thus, it is critical that the agent continuously improves and generalizes learned abilities to different tasks, regardless of their order.

In this work, we theoretically study lifelong reinforcement learning (RL) in a regret minimization setting [124, 19], where the agent needs to solve a sequence of tasks using rewards in an unknown environment while balancing exploration and exploitation. Motivated by the embodied intelligence scenario, we suppose that tasks differ in rewards, but share the same state and action spaces and transition dynamics [140]. To be realistic, we make *no*

assumptions on how the tasks and initial states are selected[1]; generally we allow them to be chosen from a continuous set by an adversary based on the agent's past behaviors. Once a task is specified and revealed, the agent has one chance (i.e., executing one rollout from its current state) to complete the task and then it moves to the next task.

The agent's goal is to perform near optimally for the tasks it faces, despite the online nature of the problem. This means that the accumulated regret of the learner compared with the best policy for each task should be sublinear in its lifetime. We assume that there is no memory constraint; this is usually the case for robotics applications where real-world interactions are the main bottleneck [140]. Nonetheless, we require that the agent eventually learns to make decisions without frequent deliberate planning, because planning is time consuming and creates undesirable wait time for user-interactive scenarios. In other words, the agent needs to learn a multi-task policy, generalizing from not only past samples but also past computation, to solve new tasks.

Formally, we consider an episodic setup based on the framework of contextual Markov decision process (CMDP) [1, 57]. It repeats the following steps: 1)At the beginning of an episode, the agent is set to an initial state and receives a context specifying the task reward, both of which can be arbitrarily chosen. 2) When needed, the agent uses its past experiences to plan for the current task. 3) The agent runs a policy in the environment for a fixed horizon in an attempt to solve the assigned task and gains experience from its policy execution. The agent's performance is measured as the regret with respect to the optimal policy of the corresponding task. We require that, for *any* task sequence, *both* the agent's overall regret and number of planning calls to be sublinear in the number of episodes.

While lifelong RL is not new, the realistic need of *simultaneously* achieving 1) sublinear regret and 2) sublinear number of planning calls for 3) a potentially adversarial sequence of tasks and initial states makes the setup considered here particularly challenging. To our knowledge, existing works only address a strict subset of these requirements; especially, the computation aspect is often ignored. Most provable works in lifelong RL make the assumption

[1]We adopt a stricter definition of lifelong RL here to distinguish it from multi-task RL, while there are existing works on lifelong RL (e.g. [34, 79]) assuming i.i.d. tasks.

that the tasks are finitely many [19, 151, 35], or are i.i.d. [18, 34, 4, 5, 79], while others considering similar setups to ours do not provide regret guarantees [63, 140]. On the technical side, the closest lines of work are [92, 1, 57, 91, 66] for contextual MDP and [137, 6] for the dynamic setting of multi-objective RL, which study the sample complexity for arbitrary task sequences; however, they either assume the problem is tabular or require a model-based planning oracle with unknown complexity. Importantly, none of the existing works properly addresses the need of sublinear planning calls, which creates a large gap between the abstract setup and practice need.

In this chapter, we aim to establish a foundation for designing agents meeting these three practically important requirements, a problem which has been overlooked in the literature. As the first step, here we study lifelong RL with linear representation. We suppose that the contextual MDP is linearly parameterized [142, 64] and the agent needs to learn a multi-task policy based on this linear representation. To make this possible, we introduce a new completeness-style assumption on the representation which is sufficient to ensure the optimal multi-task policy is realizable under the linear representation. Under these assumptions, we propose the first provably efficient lifelong RL algorithm, **U**pper **C**onfidence **B**ound **L**ifelong **V**alue **D**istillation (UCBlvd, pronounced as "UC Boulevard"), that possesses all three desired qualities. Specifically, for K episodes of horizon H, we prove a regret bound $\tilde{\mathcal{O}}(\sqrt{(d^3 + d'd)H^4K})$ using $\tilde{\mathcal{O}}(dH\log(K))$ planning calls, where d and d' are the feature dimensions of the dynamics and rewards, respectively.

From a high-level viewpoint, UCBlvd uses a linear structure to identify what to transfer and operates by interleaving 1) independent planning for a set of representative tasks and 2) distilling the planned results into a multi-task value-based policy. UCBlvd also constantly monitors whether the new experiences it gained are sufficiently significant, based on a doubling schedule, to avoid unnecessary planning. On the technical side, UCBlvd's design is inspired by single-task LSVI-UCB [64], however, we introduce a novel distillation step based on QCQP, along with a new completeness assumption, to enable computation sharing across tasks; we also extend the low-switching cost technique [2, 52, 132] for single-task RL to the lifelong setup to achieve sublinear number of planning calls.

5.2 Preliminaries

We formulate lifelong RL as a regret minimization problem in contextual MDP [1, 57] with adversarial context and initial state sequences. We suppose that a context determines the task reward but does not affect the dynamics. Such a context dependency is common for the lifelong learning scenario where an embodied agent consecutively solves multiple tasks. Below we give the formal problem definition.

Finite-horizon contextual MDP. We consider a finite-horizon contextual MDP denoted by $M = (\mathcal{S}, \mathcal{A}, \mathcal{W}, H, \mathbb{P}, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{W}$ is the task context space, H is the horizon (length of each episode), $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ are the transition probabilities, and $r = \{r_h\}_{h=1}^H$ are the reward functions. We allow $\mathcal{S}$ and $\mathcal{W}$ to be continuous or infinitely large, while we assume $\mathcal{A}$ is finite such that $\max_{a\in\mathcal{A}}$ can be performed easily. For $h \in [H]$, $r_h(s, a, w)$ denotes the reward function whose range is assumed to be in $[0, 1]$, and $\mathbb{P}_h(s'|s, a)$ denotes the probability of transitioning to state s' upon playing action a at state s. In short, a contextual MDP can be viewed as an MDP with state space $\mathcal{S} \times \mathcal{W}$ and action space $\mathcal{A}$ where the context part of the state remains constant in an episode.[2] To simplify the notation, for any function f, we write $\mathbb{P}_h[f](s, a) \coloneqq \mathbb{E}_{s'\sim\mathbb{P}_h(.|s,a)}[f(s')]$.

Policy and value functions. In a finite-horizon contextual MDP, a policy $\pi = \{\pi_h\}_{h=1}^H$ is a sequence where $\pi_h : \mathcal{S} \times \mathcal{W} \to \mathcal{A}$ determines the agent's action at time-step h. Given π, we define its state value function as $V_h^\pi(s, w) \coloneqq \mathbb{E}[\sum_{h'=h}^H r_{h'}\left(s_{h'}, \pi_{h'}(s_{h'}, w), w\right)|s_h = s]$ and its action-value function as $Q_h^\pi(s, a, w) \coloneqq r_h(s, a, w) + \mathbb{P}_h[V_{h+1}^\pi(., w)](s, a)$, where $Q_{H+1}^\pi = 0$. We denote the optimal policy as $\pi_h^*(s, w) \coloneqq \sup_\pi V_h^\pi(s, w)$, and let $V_h^* \coloneqq V_h^{\pi^*}$ and $Q_h^* \coloneqq Q_h^{\pi^*}$ denote the optimal value functions. Lastly, we recall the Bellman equation of the optimal policy:

$$Q_h^*(s, a, w) = r_h(s, a, w) + \mathbb{P}_h[V_{h+1}^*(., w)](s, a), \quad V_h^*(s, w) = \max_{a\in\mathcal{A}} Q_h^*(s, a, w). \tag{5.1}$$

[2] In general, a context-dependent dynamics would take the form $\mathbb{P}_h(s'|s, a, w)$.

Interaction protocol of lifelong RL. The agent interacts with a contextual MDP M in episodes. For presentation simplicity, we assume that the reward functions r are known, while the transition probabilities $\mathbb{P}$ are *unknown* and must be learned online; we will discuss how reward learning can be naturally incorporated in Section 5.4.3. At the beginning of episode k, the agent receives a task context $w^k \in \mathcal{W}$ and is set to an initial state s_1^k, both of which can be adversarially chosen. The agent can use past experiences to plan for the current task, if needed. Then the agent executes its policy π^k: at each time-step $h \in [H]$, it observes the state s_h^k, plays an action $a_h^k = \pi_h^k(s_h^k, w^k)$, observes a reward $r_h^k \coloneqq r_h(s_h^k, a_h^k, w^k)$, and goes to the next state s_{h+1}^k according to $\mathbb{P}_h(.|s_h^k, a_h^k)$. Let K be the total number of episodes. The agent's goal is to achieve sublinear regret, where the regret is defined as

$$R_K \coloneqq \sum_{k=1}^{K} V_1^*(s_1^k, w^k) - V_1^{\pi^k}(s_1^k, w^k). \tag{5.2}$$

As the comparator policy above (namely π^* that defines V_1^*) also knows the task context, achieving sublinear regret implies that the agent would attain near task-specific optimal performance on average.

Linear model representation. We focus on MDPs with linear transition kernels and reward functions [64, 142] that are encapsulated in the following assumption.

Assumption 13 (Linear MDPs)**.** *$M = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, \mathcal{W})$ is a linear MDP with feature maps $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\boldsymbol{\psi} : \mathcal{S} \times \mathcal{A} \times \mathcal{W} \to \mathbb{R}^{d'}$. That is, for any $h \in [H]$, there exist a vector $\boldsymbol{\eta}_h$ and d measures $\boldsymbol{\mu}_h \coloneqq [{\mu_h}^{(1)}, \ldots, {\mu_h}^{(d)}]^\top$ over $\mathcal{S}$ such that $\mathbb{P}_h(.|s,a) = \big\langle \boldsymbol{\mu}_h(.), \boldsymbol{\phi}(s,a) \big\rangle$ and $r_h(s,a,w) = \big\langle \boldsymbol{\eta}_h, \boldsymbol{\psi}(s,a,w) \big\rangle$, for all $(s,a,w) \in \mathcal{S} \times \mathcal{A} \times \mathcal{W}$. Without loss of generality, $\big\|\boldsymbol{\phi}(s,a)\big\|_2 \leq 1$, $\big\|\boldsymbol{\psi}(s,a,w)\big\|_2 \leq 1$, $\big\|\boldsymbol{\mu}_h(s)\big\|_2 \leq \sqrt{d}$, and $\|\boldsymbol{\eta}_h\|_2 \leq \sqrt{d'}$ for all $(s,a,w,h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{W} \times [H]$.*

In real-world problems, we can use the context to model the task specification of a problem. For example, if we want to design household robots to assist humans with a series of tasks like cooking, cleaning, washing dishes, lawn mowing, vacuuming, we can treat the the context as a natural language instruction that the human user would give to the robot, and we can

view the representations $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ as the embedding of a deep neural network model that has been pre-trained.

Example 1 (Weighted rewards)**.** *An interesting and common special case is* $\boldsymbol{\psi}(s,a,w) = \boldsymbol{\phi}(s,a) \otimes \boldsymbol{\rho}(w)$*, for some mapping* $\boldsymbol{\rho} : \mathcal{W} \to \mathbb{R}^m$*. In this case, it holds that* $d' = md$ *and* $r_h(s,a,w) = \langle \boldsymbol{\rho}(w), \mathbf{r}_h(s,a) \rangle$*, where* $\mathbf{r}_h(s,a) = \mathbf{A}_h \boldsymbol{\phi}(s,a) \in \mathbb{R}^m$*, for some* $\mathbf{A}_h \in \mathbb{R}^{m \times d}$*, is the vector reward functions at time-step* h*. We can view* $r_h(s,a,w)$ *as a weighted reward with weights* $\boldsymbol{\rho}(w)$ *that depend on task* w*. This setting is closely related to Multi-Objective RL studied for tabular case in [137], which studies the case where* $\boldsymbol{\rho}(w) = w \in \mathbb{R}^m$ *along with tabular* $\mathcal{S}$ *and* $\mathcal{A}$*.*

5.3 A Warm-up Algorithm for Lifelong RL

We first present a warm-up algorithm based on linear representation, termed Lifelong Least-Squares Value Iteration (Lifelong-LSVI), in Algorithm 6, which is a straightforward extension of the single-task LSVI-UCB algorithm proposed by [64] to the lifelong learning setting. The motivation of this warm-up algorithm is to give intuitions on how the problem structure in Assumption 13 can be used to achieve small regret and discuss the computational difficulty in lifelong learning.

We will show that Lifelong-LSVI has a sublinear regret bound, which matches the minimax optimal rate in the special case studied by [137] in terms of number of objectives, m (see Example 1). However, we will also show that Lifelong-LSVI is *not* computationally efficient, in the sense that the number of planning calls it requires grows linearly with the number of episodes, which would mean the overall computational complexity grows quadratically. This high computation cost is because the agent never learns to internalize the task solving skills but requires going though all past experiences for planning every time a new task arrives. Importantly, we will discuss why it cannot be made computationally efficient in an easy manner without further assumptions on the representation. This drawback motivates our new completeness assumption and our main algorithm, UCBlvd, which is provably efficient in terms of both regret and number of planning calls, in Section 5.4.

We remark that Lifelong-LSVI is only a warm-up algorithm that guides the reader to understand the mechanisms used for addressing the problem, motivates the need for UCBlvd, and shows what regret bound is possible when computational complexity is not a concern (though being impractical).

5.3.1 Algorithmic Notations

To begin, we introduce the template and the notations that will be used commonly in presenting the warm-up algorithm, Lifelong-LSVI, and later our main algorithm, UCBlvd. For each algorithm, first we will define an algorithm-specific action-value function $Q_h^k : \mathcal{S} \times \mathcal{A} \times \mathcal{W} \to \mathbb{R}$, which determines the agent's policy at time-step h in episode k; then we present the full algorithm and its analysis using the quantities below, which are defined with respect to each algorithm's definition of Q_h^k.

Given $\{Q_h^k\}_{h\in[H]}$, we define state value functions and their backups as

$$V_h^k(s,w) \coloneqq \min\left\{\max_{a\in\mathcal{A}} Q_h^k(s,a,w), H\right\}, \quad \boldsymbol{\theta}_h^k(w) \coloneqq \int_{\mathcal{S}} V_{h+1}^k(s',w) d\boldsymbol{\mu}_h(s'), \tag{5.3}$$

Thanks to the linear MDP structure in Assumption 13, it holds that

$$\mathbb{P}_h\left[V_{h+1}^k(.,w)\right](s,a) = \left\langle \boldsymbol{\theta}_h^k(w), \boldsymbol{\phi}(s,a)\right\rangle. \tag{5.4}$$

Let $\lambda > 0$ be a constant. We define the λ-regularized least squares estimator of $\boldsymbol{\theta}_h^k(w)$ as

$$\tilde{\boldsymbol{\theta}}_h^k(w) \coloneqq \left(\boldsymbol{\Lambda}_h^k\right)^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau V_{h+1}^k(s_{h+1}^\tau, w), \text{ where } \boldsymbol{\Lambda}_h^k \coloneqq \lambda \mathbf{I}_d + \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau \boldsymbol{\phi}_h^{\tau\top}, \tag{5.5}$$

and $\tilde{\boldsymbol{\theta}}_h^k(w)$ is the solution to $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{\tau=1}^{k-1}(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(s_h^\tau, a_h^\tau)\rangle - V_{h+1}^k(s_{h+1}^\tau, w))^2 + \lambda\|\boldsymbol{\theta}\|_2^2$, $\boldsymbol{\phi}_h^\tau \coloneqq \boldsymbol{\phi}(s_h^\tau, a_h^\tau)$, and $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is the identity matrix.

5.3.2 Details of Lifelong-LSVI and its Theoretical Guarantees

We define the upper confidence bound (UCB) style action-value function of Lifelong-LSVI as follows:

$$Q_h^k(s,a,w) \coloneqq r_h(s,a,w) + \left\langle \tilde{\boldsymbol{\theta}}_h^k(w), \boldsymbol{\phi}(s,a)\right\rangle + \beta\left\|\boldsymbol{\phi}(s,a)\right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}}, \tag{5.6}$$

Algorithm 6 Lifelong-LSVI

1: **Set:** $Q^k_{H+1}(.,.,.) = 0,\ \forall k \in [K]$
2: **for** episodes $k = 1, \ldots, K$ **do**
3: Observe the initial state s^k_1 and the task context w^k.
4: **for** time-steps $h = H, \ldots, 1$ **do**
5: Compute $\tilde{\boldsymbol{\theta}}^k_h(w^k)$ as in (5.5) using Q^k_{h+1} defined in (5.6).
6: **end for**
7: **for** time-steps $h = 1, \ldots, H$ **do**
8: Compute $Q^k_h(s^k_h, a, w^k)$ for all $a \in \mathcal{A}$ as in (5.6).
9: Play $a^k_h = \arg\max_{a \in \mathcal{A}} Q^k_h(s^k_h, a, w^k)$ and observe s^k_{h+1} and r^k_h.
10: **end for**
11: **end for**

where $Q^k_{H+1} = 0$ and $\tilde{\boldsymbol{\theta}}^k_h(w)$ and $\boldsymbol{\Lambda}^k_h$ are defined in (5.5). Here, β is an exploration factor that will be appropriately chosen in Theorem 9. At episode k, given w^k, Lifelong-LSVI first performs planning backward in time based on past data to compute $\tilde{\boldsymbol{\theta}}^k_h(w^k)$ in (5.5) using Q^k_{h+1} defined in (5.6) (Lines 4- 5). Then, in execution, it uses $\tilde{\boldsymbol{\theta}}^k_h(w^k)$ to compute $Q^k_h(s^k_h, a, w^k)$ for the current state and all $a \in \mathcal{A}$ (Line 8) and executes the action with the highest value (Line 9).

We show that Lifelong-LSVI achieves sublinear regret for our lifelong RL setup. The complete proof is reported in Appendix D.1, which follows the ideas of LSVI-UCB [64].

Theorem 9. *Let $T = KH$. Under Assumption 13, there exists an absolute constant $c > 0$ such that for any fixed $\delta \in (0, 0.5)$, if we set $\lambda = 1$ and $\beta = cH\left(d + \sqrt{d'}\right)\sqrt{\log(dd'T/\delta)}$ in Algorithm 6, then with probability at least $1 - 2\delta$, it holds that $R_K \leq \tilde{\mathcal{O}}\left(\sqrt{(d^3 + dd')H^3T}\right)$.*

Before introducing our main algorithm in Section 5.4, we make a few remarks on the regret and number of planning calls of Lifelong-LSVI. First, Theorem 9 implies that for the special case studied by [137] (Example 1), the regret bound of Lifelong-LSVI becomes $\tilde{\mathcal{O}}(\sqrt{md^3H^3T})$. This rate is optimal in terms of its dependency on m, as shown in [137]. Furthermore, this rate matches the LSVI-UCB's regret dependencies on d and H for the

single-task setting [64].

While Lifelong-LSVI has a decent regret guarantee, it requires computing $\tilde{\boldsymbol{\theta}}_h^k(w^k)$ for all $h \in [H]$, whenever a distinct new task w^k arrives. Since the number of unique tasks may be as large as K, the total number of planning calls required in Lifelong-LSVI is K in the worst case.

Unfortunately, the number of planning calls of Lifelong-LSVI cannot be easily improved, because under Assumption 13 alone, the optimal Q-function $Q_h^*(s, a, w)$ of the CMDP can be *nonlinear* in the representation $\boldsymbol{\psi}$. As a result, for any algorithm that represents its policy linearly based on both $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$, in general it is necessary to recompute the coefficients for every new w to be optimal. For Lifelong-LSVI specifically, this nonlinear dependency shows up in $\tilde{\boldsymbol{\theta}}_h^k(w)$ of $Q_h^k(s, a, w)$ in (5.6).

In the next section, we discuss how placing a completeness-style assumption, which ensures $Q_h^*(s, a, w)$ can be linearly parameterized by $\boldsymbol{\psi}$, would circumvent the issue of non-linear dependency of the action-value functions on w, and consequently would enable computation sharing to decrease the number of planning calls to $\mathcal{O}(dH \log(K))$.

5.4 UCB Lifelong Value Distillation (UCBlvd)

In this section, we present our main algorithm, **UCB L**ifelong **V**alue **D**istillation (UCBlvd), in Algorithm 7. Under new completeness-style assumption that we will introduce in Section 5.4.1, we show that UCBlvd shares the same regret bound as Lifelong-LSVI but significantly reduces the number of planning calls to be logarithmic in K. In contrast to Lifelong-LSVI which learns individual action-value function for each w^k, UCBlvd learns a single action-value function for all $w \in \mathcal{W}$ based on $\boldsymbol{\psi}(s, a, w)$ to enable computation sharing across tasks, which is made possible by the extra completeness-style assumption. In general, in order to directly extend Lifelong-LSVI to only use feature $\boldsymbol{\psi}(s, a, w) \in \mathbb{R}^{d'}$ with $d' \geq d$, we need a context-dependent dynamics structure, which would eventually increase the regret. UCBlvd maintains the same order of regret as Lifelong-LSVI by separating the planning into a novel two-step process: 1) independent planning with $\boldsymbol{\phi}$ for a set of representative task contexts

and 2) distilling the planned results into a multi-task value function parameterized by $\boldsymbol{\psi}$. In addition, UCBlvd runs a doubling schedule to decide whether replanning is necessary, which makes the total number of planning calls logarithmic in K.

5.4.1 Enabling Computation Sharing

As lifelong RL with Assumption 13 alone would require replanning in every episode in general (see Section 5.3), here we introduce new structural assumptions on $\boldsymbol{\psi}$ to enable computation sharing across tasks. First, we define the following class of functions

$$\mathcal{F} = \left\{ f : f(s,w) = \min\left\{ \max_{a\in\mathcal{A}} \left\{ \langle \boldsymbol{\nu}, \boldsymbol{\psi}(s,a,w)\rangle + \beta \left\|\boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Lambda}^{-1}} \right\}^{+}, H \right\}, \boldsymbol{\nu} \in \mathbb{R}^{d'}, \boldsymbol{\Lambda} \in \mathbf{S}^{d}_{++}, \beta \geq 0 \right\},$$

where $\mathbf{S}^{d}_{++}$ denotes the set of symmetric positive definite matrices. We now state our main completeness-style assumption.

Assumption 14 (Completeness)**.** *For any $f \in \mathcal{F}$ and $h \in [H]$, there exists a vector $\boldsymbol{\xi}^{f}_{h} \in \mathbb{R}^{d'}$ with $\left\|\boldsymbol{\xi}^{f}_{h}\right\| \leq H\sqrt{d'}$ such that $\mathbb{P}_h\left[f(.,w)\right](s,a) = \langle \boldsymbol{\xi}^{f}_{h}, \boldsymbol{\psi}(s,a,w)\rangle$.*

This assumption says that the backups of functions in $\mathcal{F}$ are captured by the feature $\boldsymbol{\psi}$ with bounded parameters. The definition of $\mathcal{F}$ closely models the structure of action-value function used by Lifelong-LSVI in (5.6), except $\langle \tilde{\boldsymbol{\theta}}^{k}_{h}(w), \boldsymbol{\phi}(s,a)\rangle$ there is replaced by functions linear in $\boldsymbol{\psi}(s,a,w)$. We will see that the action-value function used by UCBlvd defined in the next section is contained in $\mathcal{F}$. In addition, by setting $\beta = 0$ in $\mathcal{F}$ and (5.1), we see $Q^{*}_{h}(s,a,w)$ is linearly realizable by $\boldsymbol{\psi}$ under Assumption 14. We note that a similar notion of this assumption is mentioned in previous work for single-task settings under the name of "optimistic closure" [133].

Inspired by Example 1, we now introduce the next assumption on the structure of $\boldsymbol{\psi}$.

Assumption 15 (Mappings)**.** *We assume $\boldsymbol{\psi}(s,a,w) = \boldsymbol{\phi}(s,a) \otimes \boldsymbol{\rho}(w)$, for some mapping $\boldsymbol{\rho} : \mathcal{W} \to \mathbb{R}^{m}$, i.e., $d' = md$. We assume that there is a* known *set $\{w^{(1)}, w^{(2)}, \ldots, w^{(n)}\}$ of $n \leq m$ task contexts such that $\boldsymbol{\rho}(w) \in Span(\{\boldsymbol{\rho}(w^{(j)})\}_{j\in[n]})$ for all $w \in \mathcal{W}$. That is, for any $w \in \mathcal{W}$, there exist coefficients $\{c_j(w)\}_{j\in[n]}$ such that $\boldsymbol{\rho}(w) = \sum_{j\in[n]} c_j(w)\boldsymbol{\rho}(w^{(j)})$. We assume $\sum_{j\in[n]}\left|c_j(w)\right| \leq L$ for all $w \in \mathcal{W}$ and some $L < \infty$.*

Note that, for finite-dimensional representations, such set $\{\boldsymbol{\rho}(w^{(j)})\}_{j\in[n]}$ always exists. We assume that this set $\{w^{(1)}, w^{(2)}, \ldots, w^{(n)}\}$ is known to the algorithm

5.4.2 Details of UCBlvd

We define the UCB style action-value function of UCBlvd as follows:

$$Q_h^k(s,a,w) \coloneqq \left\{ r_h(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle + 2L\beta \left\| \boldsymbol{\phi}(s,a) \right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}} \right\}^+. \tag{5.7}$$

The parameter $\hat{\boldsymbol{\xi}}_h^k$ is computed by solving the convex quadratically constrained quadratic program (QCQP) in (5.8), which is defined on a set of representative task contexts $\{w^{(1)}, w^{(2)}, \ldots, w^{(n)}\}$ in Assumption 15 and state-action pairs $\mathcal{D} \coloneqq \big\{(s,a) : \boldsymbol{\phi}(s,a) \text{ are } d \text{ linearly independent vectors.}\big\}$.

$$\begin{aligned} \hat{\boldsymbol{\xi}}_h^k, \{\hat{\boldsymbol{\theta}}_h^{k(j)}\}_{j\in[n]} = \underset{\boldsymbol{\xi}, \{\boldsymbol{\theta}^{(j)}\}_{j\in[n]}}{\arg\min} & \sum_{j\in[n]} \sum_{(s,a)\in\mathcal{D}} \Big(\langle \boldsymbol{\theta}^{(j)}, \boldsymbol{\phi}(s,a) \rangle - \langle \boldsymbol{\xi}, \boldsymbol{\psi}(s,a,w^{(j)}) \rangle \Big)^2 \\ \text{s.t. } & \left\| \boldsymbol{\theta}^{(j)} - \tilde{\boldsymbol{\theta}}_h^k(w^{(j)}) \right\|_{\boldsymbol{\Lambda}_h^k} \le \beta, \ \forall j \in [n] \quad \text{and} \quad \|\boldsymbol{\xi}\|_2 \le H\sqrt{md}, \end{aligned} \tag{5.8}$$

where $\tilde{\boldsymbol{\theta}}_h^k(w)$ and $\boldsymbol{\Lambda}_h^k$ are defined in (5.5). In Appendix D.2.3, we will show that the action-value function in (5.7) is an optimistic estimate of the optimal action-value function.

UCBlvd also uses the linear dependency of Q_h^k on $\boldsymbol{\psi}$ to reduce calls of the planning step in (5.8). The agent triggers replanning only when it has gathered enough new information compared to the last update at episode $\tilde{k}$. This is measured by tracking the variations in Gram matrices $\{\boldsymbol{\Lambda}_h^k\}_{h\in[H]}$ (Line 4 for Algorithm 7). Finally, when executing the policy at episode k, the agent chooses the action according to $Q_h^{\tilde{k}}$ in Line 12.

5.4.3 Theoretical analysis of UCBlvd

We present our main theoretical result which shows UCBlvd achieves sublinear regret in lifelong RL using sublinear number of planning calls, for any sequence of tasks. The proof is given in Appendix D.2.

Theorem 10. *Let $T = KH$. Under Assumptions 13, 14, and 15, the number of planning*

Algorithm 7 UCBlvd (UCB Lifelong Value Distillation)

Set: $Q^k_{H+1}(.,.,.) = 0,\ \forall k \in [K],\ \tilde{k} = 1$
for episodes $k = 1, \ldots, K$ **do**
  Observe the initial state s^k_1 and the task context w^k.
  if $\exists h \in [H]$ such that $\log\det \mathbf{\Lambda}^k_h - \log\det \mathbf{\Lambda}^{\tilde{k}}_h > 1$ **then**
    $\tilde{k} = k$
    for time-steps $h = H, \ldots, 1$ **do**
      Compute $\hat{\boldsymbol{\xi}}^{\tilde{k}}_h$ as in (5.8).
    end for
  end if
  for time-steps $h = 1, \ldots, H$ **do**
    Compute $Q^{\tilde{k}}_h(s^k_h, a, w^k)$ for all $a \in \mathcal{A}$ as in (5.7).
    Play $a^k_h = \arg\max_{a\in\mathcal{A}} Q^{\tilde{k}}_h(s^k_h, a, w^k)$ and observe s^k_{h+1} and r^k_h.
  end for
end for

calls in Algorithm 7 is at most $dH\log(1+\frac{K}{d\lambda})$*, and there exists an absolute constant* $c>0$ *such that for any fixed* $\delta\in(0,0.5)$*, if we set* $\lambda=1$ *and* $\beta=cH(d+\sqrt{md})\sqrt{\log(mdT/\delta)}$ *in Algorithm 7, then with probability at least* $1-2\delta$*, it holds that* $R_K\leq\tilde{\mathcal{O}}\left(L\sqrt{(d^3+md^2)H^3T}\right)$.

Theorem 10 shows that UCBlvd has the same regret bound as Lifelong-LSVI in Theorem 9, but reduces the number of planning calls from K to $dH\log(1+K/d\lambda)$. As we discussed before, this is made possible by the unique QCQP-based distillation step of UCBlvd in (5.8). If we were to simply perform least-squares regression to fit $\langle\boldsymbol{\psi}(s,a,w),\hat{\boldsymbol{\xi}}_h^k\rangle$ to $\{\langle\boldsymbol{\phi}(s,a),\tilde{\boldsymbol{\theta}}_h^k(w^{(j)})\}_{j\in[n]}$ for distillation, we cannot guarantee the required optimism, because $\langle\phi(s,a),\tilde{\boldsymbol{\theta}}_h^k(w)\rangle$ computed based on finite samples can be an irregular function that cannot be modelled by $\boldsymbol{\psi}(s,a,w)$.

Remark 2. *If the rewards are unknown, we can adopt a slightly different completeness assumption with an extra bonus in terms of* $\boldsymbol{\psi}$*, and then combine tools from linear bandits [2] and our proof of Theorem 10. Because reward learning affects the radius of the confidence intervals for* $\boldsymbol{\theta}_h^k(w)$*, the number of planning calls and regret would increase by factors of* $\mathcal{O}(m)$ *and* $\mathcal{O}(\sqrt{m})$ [3]*, respectively, compared to those in Theorem 10. See Appendix D.3 for details.*

Remark 3. *It is possible to eliminate the assumption that* $\boldsymbol{\psi}(s,a,w)=\boldsymbol{\phi}(s,a)\otimes\boldsymbol{\rho}(w)$*. In this case, our analysis would instead require a set* $\{w^{(1)},w^{(2)},\ldots,w^{(n)}\}$ *of* n *tasks such that* $\boldsymbol{\psi}(s,a,w)\in Span(\{\boldsymbol{\psi}(s,a,w^{(j)})\}_{j\in[n]})$ *for all* $(s,a,w)\in\mathcal{S}\times\mathcal{A}\times\mathcal{W}$*. In Appendix D.4, we provide details of this relaxation, and show that this version still enjoys the same planning calls and regret as in Theorem 10.*

Remark 4. *We can eliminate Assumptions 13 and 15 and instead design a computation-sharing version of Lifelong-LSVI under a sightly different completeness assumption with a class* $\mathcal{F}$*, whose exploration bonus is* $\beta\big\|\boldsymbol{\psi}(s,a,w)\big\|_{\tilde{\boldsymbol{\Lambda}}^{-1}}$*. This assumption naturally includes settings with linear MDP in which dynamics also change with task context, i.e., for all* $h\in[H]$*, it holds that* $\mathbb{P}_h(.|s,a,w)=\langle\boldsymbol{\mu}_h(.),\boldsymbol{\psi}(s,a,w)\rangle$ *for* d' *unknown measures* $[\mu_h^{(1)},\ldots,\mu_h^{(d')}]^\top$*. Under this assumption, a slightly modified version of Lifelong-LSVI would use* $Q_h^k(s,a,w)=\{r_h(s,a,w)+\langle\tilde{\boldsymbol{\nu}}_h^k,\boldsymbol{\psi}(s,a,w)\rangle+\beta\big\|\boldsymbol{\psi}(s,a,w)\big\|_{(\tilde{\boldsymbol{\Lambda}}_h^k)^{-1}}\}^+$*, where*

[3] *While for both settings in this remark and Remark 4, the action-value functions contain exploration bonus in terms of* $\boldsymbol{\psi}$*, the regret here is better by a factor of* $\sqrt{m}$ *and this is because the multiplicative factor* β *here saves a factor* $\sqrt{m}$ *compared to that in Remark 4.*

$\tilde{\boldsymbol{\nu}}_h^k = (\tilde{\boldsymbol{\Lambda}}_h^k)^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\psi}_h^\tau . \min\{\max_{a\in\mathcal{A}} Q_{h+1}^k(s_{h+1}^\tau, a, w^\tau), H\}$, $\tilde{\boldsymbol{\Lambda}}_h^k = \lambda \mathbf{I}_{d'} + \sum_{\tau=1}^{k-1} \boldsymbol{\psi}_h^\tau \boldsymbol{\psi}_h^{\tau\top}$, $\boldsymbol{\psi}_h^\tau = \boldsymbol{\psi}(s_h^\tau, a_h^\tau, w^\tau)$, *and* $\beta = \tilde{\mathcal{O}}(d')$. *However, in Appendix D.5, we show how these new algorithm and assumption result in* $\tilde{\mathcal{O}}(mdH)$ *number of planning calls and a regret scaling with* $\tilde{\mathcal{O}}(\sqrt{m^3 d^3})$ *for settings with* $\boldsymbol{\psi}(s,a,w) = \boldsymbol{\phi}(s,a) \otimes \boldsymbol{\rho}(w)$. *These are worse than the number of planning calls and regret in Theorem 10 of UCBlvd by a factor of* $\mathcal{O}(m)$.

Remark 5. *A natural follow-up relaxation of Assumption 14 is when the equality holds up to an error of* ζ. *In Appendix D.6, we show that this relaxation results in a regret* $\tilde{\mathcal{O}}\left(\sqrt{md}T\zeta + \sqrt{\lambda(d^3 + md^2)H^3 T}\right)$ *and the same number of planning calls as that in Theorem 10. When* ζ *is sufficiently small, i.e.,* $\zeta = \mathcal{O}(\sqrt{d^2H^3/mT})$, *UCBlvd will still enjoy a regret of the same order as that in Theorem 10.*

5.4.4 Proof Sketch of Theorem 10

Because the proof of planning calls' upper bound follows standard arguments in low switching cost analysis of [2], in this section, we focus on the proof sketch for the regret bound. We start by introducing the high probability event $\mathcal{E}_1$, which is the foundation of our analysis:

$$\mathcal{E}_1(w) \coloneqq \left\{ \left\| \boldsymbol{\theta}_h^k(w) - \tilde{\boldsymbol{\theta}}_h^k(w) \right\|_{\boldsymbol{\Lambda}_h^k} \leq \beta, \forall (h,k) \in [H] \times [K] \right\}. \tag{5.9}$$

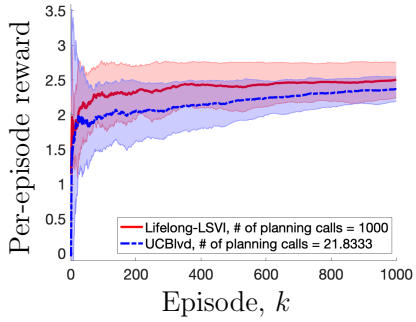
The following lemma highlights the importance of the carefully designed planning step in (5.8), which ensures good estimators for $\boldsymbol{\xi}_h^{V_{h+1}^*}$ without the need of bonus term $\left\|\boldsymbol{\psi}(s,a,w)\right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}}$. This step saves a factor $\mathcal{O}(m)$ in planning calls and regret.

Lemma 7. *Let* $\widetilde{\mathcal{W}} = \{w^\tau : \tau \in [K]\} \cup \{w^{(j)} : j \in [n]\}$. *Under the setting of Theorem 10 and conditioned on events* $\{\mathcal{E}_1(w)\}_{w\in\widetilde{\mathcal{W}}}$ *defined in* (5.9), *for all* $(s,a,w,h,k) \in \mathcal{S} \times \mathcal{A} \times \widetilde{\mathcal{W}} \times [H] \times [K]$, *it holds that* $\left| \langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w) \rangle - \mathbb{P}_h[V_{h+1}^k(.,w)](s,a) \right| \leq 2L\beta \left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}}$.

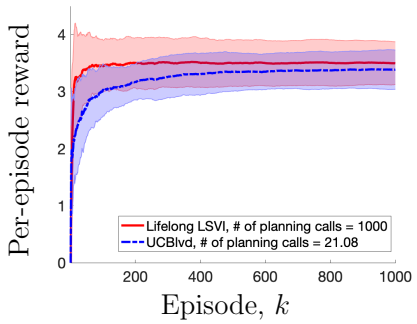
As the final step in the regret analysis, we use Lemma 7 to prove the optimistic nature of UCBlvd, i.e., $Q_h^k(s,a,w^k) \geq Q_h^*(s,a,w^k)$ for all $(s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$. Then following the standard analysis of single-task LSVI-UCB we derive the regret bound in Theorem 10.

5.4.5 Experiments

We implemented our main algorithm UCBlvd on synthetic environments and compared its performance with the warm-up algorithm Lifelong-LSVI, which is viewed as an idealized baseline ignoring the computational complexity. In all the experiments, the same setting, task sequences and feature mappings were used for both UCBlvd and Lifelong-LSVI. Figure 5.1a depicts per-episode rewards for the main setup considered throughout the chapter, and Figure 5.1b shows those for the setup in Remark 3. The plots verify that Lifelong-LSVI and UCBlvd statistically perform almost the same while UCBlvd uses much smaller numbers of planning calls (1000 vs ~ 20). We remark that Lifelong-LSVI has an overall computation complexity of $\mathcal{O}(K^2)$, which makes it not practical for the lifelong learning setting, as its planning complexity increases linearly with the number of samples. The details on the parameters of simulations are deferred to Appendix D.8.



(a) Setting of Theorem 10, $d = 5$, $m = 5$, $d' = 25$



(b) Setting of Remark 3, $d = 5$, $d' = 10$

Figure 5.1: UCBlvd vs Lifelong-LSVI. The experimental results include 50 seeds.

5.5 Related Work

We consider the regret minimization setup of lifelong RL under the contextual MDP framework, where the agent receives tasks specified by contexts in sequence and needs to achieve a sublinear regret for any task sequence. Below, we contrast our work with related work in the literature.

Lifelong RL. Generally lifelong RL studies how to learn to solve a streaming sequence of tasks using rewards. While it was originally motivated by the need of endless learning of robots [124], historically many works on lifelong RL [18, 34, 4, 5, 79] assume that the tasks are i.i.d. (similar to multi-task RL; see below). There are works for adversarial sequences, but most of them assume finite number of tasks [35, 19, 151] or are purely empirical [140]. The work by [63] uses contexts to enable zero-shot learning like here, but it (as well as most works above) does not provide formal regret guarantees.[4] [35] and [140] assume the task identity is latent, which requires additional exploration; in this sense, their problem is harder than the setup here where the task context is revealed. Extending the setup here to consider latent context is an important future direction.

Contextual MDP and Multi-objective RL. Our setup is closely related to the exploration problem studied in the contextual MDP literature, though contextual MDP is originally not motivated from the lifelong learning perspective. A similar mathematical problem appears in the dynamic setup of multi-objective RL [137, 6], which can be viewed as a special case of contextual MDP where the context linearly determines the reward function but not the dynamics. Most contextual MDP works allow adversarial contexts and initial states, but a majority of them focuses on the tabular setup [1, 57, 91, 92, 80, 137], whereas our setup allows continuous states. [66] and [44] allow continuous state and action spaces, but the former assumes a planning oracle with unclear computational complexity and the latter focuses on only LQG problems. While generally contextual MDP allows both the reward and the dynamics to vary with contexts, we focus on the effects of context-independent dynamics similar to [66, 137]. In particular, the recent work of [137] is the closest to ours, but they study the sample complexity in the tabular setup with linearly parameterized rewards. In view of Example 1, their proposed algorithm has a regret bound $\tilde{\mathcal{O}}(\sqrt{\min\{m,|\mathcal{S}|\}H|\mathcal{S}||\mathcal{A}|\,K})$. However, they need linear number of planning calls. On the contrary, our algorithm, UCBlvd, allows continuous states, nonlinear context dependency, and has both sublinear regret and

[4] [19] give regret bounds but only for linearized value difference; [35] show regret bounds only for finite number of tasks.

number of planning calls.

Multi-task RL. Another closely related line of work is multi-task RL. Compared to our setting, multi-task RL assumes that there are beforehand known finite tasks and/or they are i.i.d. samples from a fixed distribution. For example, in [144, 61, 33, 49, 152, 109], tasks are assumed to be chosen from a known finite set, and in [144, 135, 33, 116], tasks are sampled from a fixed distribution. By contrast, our setting provides guarantees on regret and number of planning calls for adversarial task sequences.

5.6 Discussion

In this chapter, we frame lifelong RL as contextual MDPs and identify a new completeness-style assumption to enable provably efficient lifelong RL with linear representation. We propose UCBlvd, an algorithm that *simultaneously* satisfies the practical need of achieving 1) sublinear regret and 2) sublinear number of planning calls for 3) any sequence of tasks and initial states. Specifically, for K task episodes of horizon H, we prove that UCBlvd has a regret bound $\tilde{\mathcal{O}}(\sqrt{(d^3 + d'd)H^4K})$ based on $\tilde{\mathcal{O}}(dH\log(K))$ number of planning calls, where d and d' are the feature dimensions of the dynamics and rewards, respectively. We believe that our results would inspire new research directions in the literature of CMDP and multi-objective RL, as existing work to our knowledge does not cover the computation-sharing aspect of lifelong RL. That said, our work's limitations motivate further investigations in the following directions: 1) extension to more general class of MDPs, potentially using general function approximation/representation tools, 2) establishing an information-theoretic lower bound on the number of planning calls/computation complexity.

CHAPTER 6

Distributed Contextual Linear Bandits with Minimax Optimal Communication Cost

6.1 Introduction

In the contextual bandit problem, a learning agent repeatedly makes decisions based on contextual information, with the goal of learning a policy that maximizes their total reward over time. This model captures simple reinforcement learning tasks in which the agent must learn to make high-quality decisions in an uncertain environment, but does not need to engage in long-term planning. Contextual bandit algorithms are deployed in online personalization systems such as medical trials and product recommendation in e-commerce [8, 121]. For example, by modelling personalized recommendation of articles as a contextual bandit problem, a learning algorithm sequentially selects articles to be recommended to users based on contextual information about the users and articles, while continuously updating its article-selection strategy based on user-click feedback to maximize total user clicks [82].

Distributed cooperative learning is a paradigm where multiple agents collaboratively learn a shared prediction model. More recently, researchers have explored the potential of contextual bandit algorithms in distributed systems, such as in robotics, wireless networks, the power grid and medical trials [84, 24, 28, 115]. For example, in sensor/wireless networks [24] and channel selection in radio networks [85, 86, 87], a collaborative behavior is required for decision-makers/agents to select better actions as individuals.

While a distributed nature is inherent in certain systems, distributed solutions might also be preferred in broader settings, as they can lead to speed-ups of the learning process. This

calls for extensions of the traditional single-agent bandit setting to networked systems. In addition to speeding up the learning process, another desirable goal of each distributed learning algorithm is *communication efficiency*. In particular, keeping the communication as rare as possible in collaborative learning is of importance. The notion of communication efficiency in distributed learning paradigms is directly related to the issue of efficient environment queries made in single-agent settings. In many practical single-agent scenarios, where the agent sequentially makes active queries about the environment, it is desirable to limit these queries to a small number of rounds of interaction, which helps to increase the parallelism of the learning process and reduce the management cost. In recent years, to address such scenarios, a surge of research activity in the area of batch online learning has shown that in many popular online learning tasks, a very small number of batches may achieve minimax optimal learning performance, and therefore it is possible to enjoy the benefits of both adaptiveity and parallelism [103, 58, 53]. In light of the connection between communication cost in distributed settings and the number of environment queries in single-agent settings, a careful use of batch learning methods in multi-agent learning scenarios may positively affect the communication efficiency by limiting the number of necessary communication rounds. In this chapter, we first prove an information-theoretic lower bound on the communication cost of distributed contextual linear bandits, and then leverage such batch learning methods to design an algorithm with a small communication cost that matches this lower bound while guaranteeing optimal regret.

6.1.1 Problem Formulation

We consider a network of N agents acting cooperatively to efficiently solve a K-armed stochastic linear bandit problem. Let T be the total number of rounds. At each round $t \in [T]$, each agent i is given a decision set $\mathcal{X}_t^i = \{\mathbf{x}_{t,a}^i : a \in [K]\} \subset \mathbb{R}^d$, drawn independently from a distribution $\mathcal{D}_t^i$. We assume that $\mathcal{D}_t^i = \mathcal{D}$ for all $(i,t) \in [N] \times [T]$. Here, $\mathbf{x}_{t,a}^i$ is a mapping from action a and the contextual information agent i receives at round t to the d-dimensional space. We call $\mathbf{x}_{t,a}^i$ the feature vector associated with action a and agent i at round t. Agent i selects action $a_{i,t} \in [K]$, and observes the reward $y_t^i = \langle \boldsymbol{\theta}, \mathbf{x}_{t,a_{i,t}}^i \rangle + \eta_t^i$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ is an

Setting	Algorithm	Regret	Communication cost	Communication cost lower bound
Contexts are fixed over time horizon and agents	DELB with server [134]	$\mathcal{O}\left(d\sqrt{NT\log T}\right)$	$\mathcal{O}\left((dN + d\log\log d)\log T\right)$	
Contexts adversarially vary over time horizon and agents	DisLinUCB with server [134]	$\mathcal{O}\left(d\sqrt{NT}\log^2 T\right)$	$\mathcal{O}\left(d^3N^{1.5}\right)$	
	FedUCB with server [46]	$\mathcal{O}\left(d\sqrt{NT}\log^2 T\right)$	$\mathcal{O}\left(d^3N^{1.5}\right)$	
Contexts adversarially vary over agents	Fed-PE with server [62]	$\mathcal{O}(\sqrt{dNT\log(KNT)})$	$\mathcal{O}\left((d^2 + dK)N\log T\right)$	
Contexts stochastically vary over time horizon and agents (**this work**)	DisBE-LUCB with server	$\mathcal{O}(\sqrt{dNT\log d\log^2(KNT)})$	$\mathcal{O}\left(dN\log\log(NT)\right)$	$\Omega(dN)$
	DecBE-LUCB without server	$\mathcal{O}(NS + \sqrt{dN(T+S)\log d\log^2(KNT)})$	$\mathcal{O}\left(S\delta_{\max}dN\log\log(NT)\right)$	

Table 6.1: N: number of agents; K: number of arms; T: time horizon; d: dimension of the feature vectors; $S = \frac{\log(dN)}{\sqrt{1/|\lambda_2|}}$; $|\lambda_2|$: the second largest eigenvalue of communication matrix in absolute value; $\delta_{\max}$ is the maximum degree of the graph representing agents' network. The lower bound for the communication cost is interpreted as follows: For any algorithm with expected communication cost less than $\frac{dN}{64}$, there exists a contextual linear bandit instance with stochastic contexts, for which the algorithm's regret is $\Omega(N\sqrt{dT})$. See Theorem 11.

unknown vector and η_t^i is an independent zero-mean additive noise. The agents are also allowed to communicate with each other. Both the action selection and the communicated information of each agent may only depend on previously played actions, observed rewards, decision sets, and communication received from other agents. Throughout the chapter, we rely on the following assumption.

Assumption 16. *Without loss of generality,* $\|\boldsymbol{\theta}\|_2 \leq 1$, $\|\mathbf{x}_{t,a}^i\|_2 \leq 1$, $\left|y_t^i\right| \leq 1$ *for all* $(a, i, t) \in [K] \times [N] \times [T]$. *Also, the distribution* $\mathcal{D}$ *is known to the agents.*

The boundedness assumption is standard in the linear bandit literature [39, 41, 62]. Moreover, our results can be readily extended to the settings where the assumption on the boundedness of y_t^i is relaxed by assuming the noise variables η_t^i are conditionally σ-subGaussiam for a constant $\sigma \geq 0$. As such, a high probability bound on η_t^i and consequently y_t^i can be established, which is desired in our analysis for establishing confidence intervals in Appendix E.2.1.

Our assumption on the knowledge of $\mathcal{D}$ is fairly well-motivated. A standard argument is based on having loads of unsupervised data in real-world scenarios. For example, Google, Amazon, Netflix, etc, have collected massive amounts of data about users, products, and queries, sufficiently describing the joint distributions. Given this, even if the features change (for a given user or product, etc.), their distributions can be computed/sampled from as the features are computed via a deterministic feature map. In light of this, [59] recently studied contextual linear bandits with known context distribution. We further relax this assumption in Remark 8 in Section 6.4.2.

Goal. The performance of the network is measured via the cumulative regret of all agents in T rounds, defined as

$$R_T := \mathbb{E}[\textstyle\sum_{t=1}^{T} \sum_{i=1}^{N} \langle \boldsymbol{\theta}, \mathbf{x}_{*,t}^i \rangle - \langle \boldsymbol{\theta}, \mathbf{x}_t^i \rangle], \tag{6.1}$$

where the expectation is taken over the random variables $\mathcal{X}_t^i, (i, t) \in [N] \times [T]$ with joint distribution $\bigotimes_{i,t=1}^{N,T} \mathcal{D}_t^i$, $\mathbf{x}_t^i$ and $\mathbf{x}_{*,t}^i \in \arg\max_{\mathbf{x} \in \mathcal{X}_t^i} \langle \boldsymbol{\theta}, \mathbf{x} \rangle$ are the feature vectors associated with the action chosen by agent i at round t and the best possible action, respectively.

For simplicity, in our algorithms the communication cost is measured as the number of communicated real numbers *over the course of* T *rounds*. In Section 6.3, we also discuss variants of our methods where the communication cost is measured as the number of communicated bits.

The goal is to design a distributed collaborative algorithm that minimizes the cumulative regret, while maintaining an efficient coordination protocol with a small communication cost. Specifically, we wish to achieve a regret close to $\tilde{\mathcal{O}}(\sqrt{dNT})$ that is incurred by an optimal *single-agent algorithm for NT rounds* (the total number of arm pulls) while the communication cost is $\tilde{\mathcal{O}}(dN)$ with only a mild (logarithmic) dependence on T.

A motivating example. In news article recommendation, the candidate actions correspond to K news articles. At round t, an individual user visits an online news platform that has N servers employing the same recommender systems to recommend news articles from an article pool. The contextual information of the user, the articles and the servers at round t is modeled by $\mathcal{X}_t^i = \{\mathbf{x}_{t,a}^i : a \in [K]\}$, characterizing user's reaction to each recommended article a (e.g., click/not click) by server i, and the probability of clicking on a is modeled by $\langle \boldsymbol{\theta}, \mathbf{x}_{t,a}^i \rangle$, which corresponds to the expected reward. On the distributed side, these N servers collaborate with each other by sharing information about the feedback they receive from the users after recommending articles in an attempt to speed up learning the users' preferences. In this example, the individual users and articles can often be viewed as independent samples from the population which is characterized by distribution $\mathcal{D}$.

6.1.2 Contributions

We establish a lower bound on the communication cost of distributed contextual linear bandits. We propose algorithms with optimal regret and communication cost matching our lower bound (up to logarithmic factors) and growing linearly with d and N while those of previous best-known algorithms scale super linearly either in d or N. Below, we elaborate more on our contributions:

Minimax lower bound for the communication cost. As our main technical contribution, in Section 6.3, we prove the first information-theoretic lower bound on the communication cost (measured in bits) of any algorithm achieving an optimal regret rate for the distributed contextual linear bandit problem with stochastic contexts. In particular, we prove that for

any distributed algorithm with expected communication cost less than $\frac{dN}{64}$, there exists a contextual linear bandit problem instance with stochastic contexts for which the algorithm's regret is $\Omega(N\sqrt{dT})$.

DisBE-LUCB. We propose a distributed batch elimination contextual linear bandit algorithm (DisBE-LUCB): the time steps are grouped into M pre-defined batches and at each time step, each agent first constructs confidence intervals for each action's reward, and the actions whose confidence intervals completely fall below those of other actions are eliminated. Throughout each batch, each agent uses the same policy to select actions from the surviving action sets. At the end of each batch, the agents share information through a central server and update the policy they use in the next batch. We prove that while the communication cost of DisBE-LUCB is only $\tilde{\mathcal{O}}(dN)$, it achieves a regret $\tilde{\mathcal{O}}(\sqrt{dNT})$, which is of the same order as that incurred by a near optimal *single-agent algorithm for NT rounds* . This shows that DisBE-LUCB is nearly minimax optimal in terms of *both regret and communication cost*. We highlight that while DisBE-LUCB is inspired by the single-agent batch elimination style algorithms [103] in an attempt to save on communication as much as possible, a direct use of confidence intervals used in such algorithms would fail to guarantee optimal communication cost $\tilde{\mathcal{O}}(dN)$ and require more communication by a factor of $\mathcal{O}(d)$. We address this issue by introducing new confidence intervals in Lemma 9. Details are given in Section 6.4.

DecBE-LUCB. Finally, we propose a fully decentralized variant of DisBE-LUCB without a central server, where the agents can only communicate with their *immediate neighbors* given by a communication graph. Our algorithm, called decentralized batch elimination linear UCB (DecBE-LUCB), runs a carefully designed consensus procedure to spread information throughout the network. For this algorithm, we prove a regret bound that captures both the degree of selected actions' optimality and the inevitable delay in information-sharing due to the network structure while the communication cost still grows linearly with d and N. See Section 6.4.4.

We complement our theoretical results with numerical simulations under various settings

in Section 6.5.

6.2 Related Work

Distributed MAB. Multi-armed bandit (MAB) in multi-agent distributed settings has received attention from several academic communities. In the context of the classical K-armed MAB, [90, 73, 74, 75] proposed decentralized algorithms for a network of N agents that can share information only with their immediate neighbors, while [119] studied the MAB problem on peer-to-peer networks.

Distributed contextual linear bandits. The most closely related works on distributed linear bandits are those of [134, 46, 62, 70, 60]. In particular, [134] investigate communication-efficient distributed linear bandits, where the agents can communicate with a server by sending and receiving packets. They propose two algorithms, namely, DELB and DisLinUCB, for fixed and time-varying action sets, respectively. The works of [46, 62] consider the federated linear contextual bandit model and the former focuses on federated differential privacy. In the latter, the contexts denote the specifics of the agents and are different but fixed during the entire time horizon for each agent. In the former, however, the contexts contain the information about both the environment and the agents, in the sense that contexts associated with different agents are different and vary during the time horizon. To put these in the context of an example, consider a recommender system. Both [46] and [62] consider a multi-agent model, where each agent is associated with a different user profile. [62] fix a user profile for an agent, while [46] consider a time-varying user profile. Therefore, [62] capture the variation of contexts over agents, whereas it is captured over both agents and time horizon in [46]. A regret and communication cost comparison between DisBE-LUCB, DecBE-LUCB and other baseline algorithms is given in Table 6.1.

Batch elimination in distributed bandits. An important line of work related to communication efficiency in distributed bandits studies practical single-agent scenarios using

batch elimination methods, in which a very small number of batches achieve minimax optimal learning performance [103, 58, 53]. Our proposed algorithms are inspired by the single-agent BatchLinUCB-DG proposed in [103] in an attempt to save on communication as much as possible. That said, a direct use of confidence intervals in [103] would fail to guarantee optimal communication cost $\tilde{\mathcal{O}}(dN)$ and require more communication by a factor of $\mathcal{O}(d)$. We address this issue by introducing new confidence intervals, used in our algorithms, in Lemma 9.

Minimax lower bound on communication cost. We are unaware of any lower bound on the communication cost scaling with both d and N for contextual linear bandits in the distributed/federated learning setting. To the best of our knowledge, our work is the first to establish such a minimax lower bound and to propose algorithms with optimal regret and communication cost matching this lower bound up to logarithmic factors. Recently, [81] proved a $\Omega(N)$ communication lower bound for asynchronous federated contextual linear bandits. However, their lower bound does not include the dependency on d, which is of importance in our work and emphasizes how our proposed algorithm optimally improves the communication cost of existing methods. In addition, [134] previously proved a $\Omega(N)$ communication lower bound for distributed MAB.

6.3 Lower Bound on Communication Cost

In this section, we derive an information-theoretic lower bound on the communication cost of the distributed contextual linear bandits with stochastic contexts. In particular, we prove that for any distributed contextual linear bandit algorithm with stochastic contexts that achieves the optimal regret rate $\tilde{\mathcal{O}}(\sqrt{dNT})$, the expected amount of communication must be at least $\Omega(dN)$. This is formally stated in the following theorem.

Theorem 11. *Let $T \geq \max\{4d\log(8), d^2/500\}$. For any algorithm with expected communication cost (measured in bits) less than $\frac{dN}{64}$, there exists a contextual linear bandit instance with stochastic contexts, for which the algorithm's regret is $\Omega(N\sqrt{dT})$.*

6.3.1 Proof of Theorem 11

We start with a lower bound for a single-agent Bayesian two-armed bandit problem where the agent is given side information that contains a small amount of information about the optimal action.

Lemma 8. *Let $\boldsymbol{\mu}_1 = (\Delta, 0)$ and $\boldsymbol{\mu}_2 = (-\Delta, 0)$ and consider the single-agent Bayesian two-armed Gaussian bandit with mean $\boldsymbol{\mu}$ uniformly sampled from $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$ and $a_* = \arg\max_{a \in \{1,2\}} \boldsymbol{\mu}_a$, which is a random variable. Suppose additionally that the agent has access to a random element M with $I(M; a_*) \leq 1/16$. Then, for any policy π,*

$$BR_T(\pi) \geq \Delta T \left(\frac{1}{2} - \sqrt{\frac{1}{2}\left(\frac{1}{16} + 4T\Delta^2\right)} \right),$$

where $BR_T(\pi) = \mathbb{E}_{\boldsymbol{\mu} \sim \ Unif\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}}[R_T(\pi, \boldsymbol{\mu})]$ and $R_T(\pi, \boldsymbol{\mu})$ is the regret suffered by policy π in the Gaussian two-armed bandit with means $\boldsymbol{\mu}$.

Remark 6. *We assume in Lemma 8 that the agent has access to the message M from the beginning. The same bound continues to hold in the strictly harder problem where the agent has sequential access to a sequence of messages $M_1, \ldots, M_T$ with $I(\{M_t\}_{t=1}^T; a_*) \leq 1/16$.*

The proof is presented in Appendix E.1. This lemma emphasizes the role of extra information a single agent might receive throughout the learning process on its performance, and therefore, it is key in proving Theorem 11. Specifically, since Lemma 8 makes no assumption on how the agent receives the extra information about the learning environment, we can prove Theorem 11 by employing this lemma and a reduction from single-agent bandit to multi-agent bandit as explained in what follows.

The construction. We consider a bandit instance where $K = 2$ and the decision sets are drawn uniformly from $\big\{(\mathbf{e}_1, \mathbf{e}_2), (\mathbf{e}_3, \mathbf{e}_4), \ldots, (\mathbf{e}_{d-1}, \mathbf{e}_d)\big\}$. Let $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^d : (\boldsymbol{\theta}_{2j-1}, \boldsymbol{\theta}_{2j}) \in \{(\Delta, 0), (-\Delta, 0)\},\ \forall j \in [\frac{d}{2}]\}$. We call $(\boldsymbol{\theta}_{2j-1}, \boldsymbol{\theta}_{2j})$ by j-th block of reward vector.

Bayesian regret. As in Lemma 8, we prove the minimax-style lower bound using the Bayesian regret. Let $\boldsymbol{\theta}$ be sampled uniformly from Θ and π be a fixed multi-agent policy.

The multi-agent Bayesian regret is

$$BR_T = \mathbb{E}[\textstyle\sum_{t=1}^{T} \sum_{i=1}^{N} \langle \boldsymbol{\theta}, \mathbf{x}_{*,t}^i \rangle - \langle \boldsymbol{\theta}, \mathbf{x}_t^i \rangle] \,,$$

where the expectation integrates over the randomness in both $\boldsymbol{\theta}$ and the corresponding history induced by the interaction between π and the environment determined by $\boldsymbol{\theta}$. By Yao's minimax principle, there exists a $\boldsymbol{\theta} \in \Theta$ such that the expected regret is at least BR_T, so it suffices to lower bound the Bayesian regret. For the remainder of the proof $\mathbb{E}[\cdot]$ and $\mathbb{P}(\cdot)$ correspond to the expectation and probability measure on $\boldsymbol{\theta}$ and the history. For technical reasons, we assume that these probability spaces are defined to include an infinite interaction between the agents and environment. Of course, this is only used in the analysis.

Reduction from single-agent to multi-agent. Let M_{ij} be the mutual information between messages agent i receives in T rounds and $(\boldsymbol{\theta}_{2j-1}, \boldsymbol{\theta}_{2j})$. By assumption,

$$\sum_{i=1}^{N} \sum_{j=1}^{\frac{d}{2}} M_{ij} \leq \sum_{i=1}^{N} \mathbb{E}[\text{Total number of bits agent } i \text{ receives}] \leq \frac{dN}{64}. \tag{6.2}$$

Let $\mathcal{S}$ be the set of $\frac{dN}{4}$ pairs $(i,j) \in [N] \times [\frac{d}{2}]$ with smallest M_{ij}. From (6.2) and the definition of $\mathcal{S}$, we observe that for every pair $(i,j) \in \mathcal{S}$, we have

$$M_{ij} \leq \frac{dN}{64 \frac{dN}{4}} = \frac{1}{16}.$$

Let B_{ijt} be the indicator that the context is such that agent i interacts with j-th block in round t, which is

$$B_{ijt} = \mathbf{1}(\mathbf{x}_{t,1}^i = \mathbf{e}_{2j-1}) \,.$$

Note that $\{B_{ijt}\}_{t=1}^{\infty}$ are independent and $\mathbb{E}[B_{ijt}] = 2/d$. Let $\mathcal{T}_{ij} = \{t : B_{ijt} = 1\}$ and $\mathcal{T}_{ij}^{\circ}$ be the first $T_{\circ}$ elements of $\mathcal{T}_{ij}$ with $T_{\circ} = T/d$. Let

$$R_{ij} = \sum_{t \in \mathcal{T}_{ij}^{\circ}} \langle \boldsymbol{\theta}, \mathbf{x}_{*,t}^i \rangle - \langle \boldsymbol{\theta}, \mathbf{x}_t^i \rangle$$

be the regret of agent i during the rounds in $\mathcal{T}_{ij}^{\circ}$ in bandit instance $\boldsymbol{\theta}$. Note that $\mathcal{T}_{ij}^{\circ}$ may contain rounds larger than T. Nevertheless,

$$\begin{aligned} BR_T &\geq \sum_{i=1}^{N} \sum_{j=1}^{d/2} \mathbb{E}[R_{ij} \mathbf{1}(\mathcal{T}_{ij}^{\circ} \subset \{1, \dots, T\})] \\ &\geq \sum_{(i,j) \in \mathcal{S}} \mathbb{E}[R_{ij} \mathbf{1}(\mathcal{T}_{ij}^{\circ} \subset \{1, \dots, T\})] \\ &= \sum_{(i,j) \in \mathcal{S}} \mathbb{E}[R_{ij}] - \mathbb{E}[R_{ij} \mathbf{1}(\mathcal{T}_{ij}^{\circ} \not\subset \{1, \dots, T\})] \,. \end{aligned}$$

Suppose that $(i, j) \in \mathcal{S}$. Now, $\mathbb{E}[R_{ij}]$ is exactly the Bayesian regret of some policy interacting with the Bayesian two-armed bandit defined in Lemma 8 for $T_{\circ}$ rounds. Furthermore, the mutual information between the optimal action in this bandit and the messages passed to the agent is at most $M_{ij} \leq 1/16$. Hence, by Lemma 8 and Remark 6,

$$\mathbb{E}[R_{ij}] \geq \Delta T_{\circ} \left(\frac{1}{2} - \sqrt{\frac{1}{2} \left(\frac{1}{16} + 4 T_{\circ} \Delta^2 \right)} \right) \,.$$

On the other hand,

$$\mathbb{E}[R_{ij} \mathbf{1}(\mathcal{T}_{ij}^{\circ} \not\subset \{1, \dots, T\})] \leq 2 \Delta T_{\circ} \mathbb{P}(\mathcal{T}_{ij}^{\circ} \not\subset \{1, \dots, T\}) = 2 \Delta T_{\circ} \mathbb{P} \left(\textstyle\sum_{t=1}^{T} B_{ijt} < T_{\circ} \right) .$$

By Chernoff's bound, $T \geq 4d \log(8)$ and $\mathbb{E}[B_{ijt}] = 2/d$,

$$2 \mathbb{P} \left(\textstyle\sum_{t=1}^{T} B_{ijt} < T_{\circ} \right) = 2 \mathbb{P} \left(\textstyle\sum_{t=1}^{T} B_{ijt} < T/d \right) \leq 2 \exp\left(-T/(4d) \right) \leq \frac{1}{4} \,.$$

Therefore, with $\Delta = 0.0695 \sqrt{\frac{d}{T}}$, we have

$$BR_T \geq \frac{d N T_{\circ} \Delta}{4} \left(\frac{1}{4} - \sqrt{\frac{1}{2} \left(\frac{1}{16} + 4 T_{\circ} \Delta^2 \right)} \right) \geq \frac{N \sqrt{dT}}{1250} = \Omega \left(N \sqrt{dT} \right) \,,$$

which concludes the proof of Theorem 11. Also, note that the fact that $T \geq \frac{d^2}{500}$ ensures that $\|\boldsymbol{\theta}\|_2 = \Delta \sqrt{\frac{d}{2}} \leq 1$ which is compatible with Assumption 16.

6.4 An Optimal Algorithm

Following the communication cost lower bound in previous section, we now present an algorithm called, *Distributed Batch Elimination Linear Upper Confidence Bound* (DisBE-LUCB), whose communication cost matches the lower bound up to logarithmic factors while

achieving an optimal regret rate. DisBE-LUCB employs a central server to which, the agents send *local* updates and it then aggregates and broadcasts the updated *global* values of interest. We also discuss *Decentralized Batch Elimination Linear Upper Confidence Bound* (DecBE-LUCB), a modified version of DisBE-LUCB in the absence of a central server, where each agent can only communicate with its *immediate neighbors*.

6.4.1 Overview of DisBE-LUCB

Before describing how DisBE-LUCB operates for every agent $i \in [N]$, we note that all agents run DisBE-LUCB concurrently. In DisBE-LUCB, the time steps are grouped into M pre-defined batches by a grid $\mathcal{T} = \{\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_M\}$, where $0 = \mathcal{T}_0 \leq \mathcal{T}_1 \leq \ldots \leq \mathcal{T}_M$, $T \leq \mathcal{T}_M$ and $T_m = \mathcal{T}_m - \mathcal{T}_{m-1}$ is the length of batch m. Our choice of grid implies that for any $m \geq 3$, we have $T_m = (a^{2^{m-1}-1} d^{\frac{1}{2}} / N^{\frac{1}{2}})^{\frac{1}{2^{m-2}}}$. Parameter a is chosen such that $T_M = T$ and $\mathcal{T}_M = \sum_{m \in [M]} T_m \geq T_M = T$, and therefore our choice of grid $\mathcal{T}$ is valid. At rounds $t \in [\mathcal{T}_{m-1} + 1 : \mathcal{T}_m]$ during batch $m \in [M]$, agent i first constructs confidence intervals for each action's reward, and the actions whose confidence intervals completely fall below those of other actions are eliminated. We denote the set of feature vectors associated with the surviving actions by $\mathcal{X}_t^{i(m)} = \cap_{k=0}^{m-1} \mathcal{E}(\mathcal{X}_t^i; (\mathbf{\Lambda}_k^i, \boldsymbol{\theta}_k^i, \beta))$, where

$$\mathcal{E}(\mathcal{X}_t^i; (\mathbf{\Lambda}_k^i, \boldsymbol{\theta}_k^i, \beta)) := \{\mathbf{x} \in \mathcal{X}_t^i : \langle \boldsymbol{\theta}_k^i, \mathbf{x} \rangle + \beta \|\mathbf{x}\|_{(\mathbf{\Lambda}_k^i)^{-1}} \geq \langle \boldsymbol{\theta}_k^i, \mathbf{y} \rangle - \beta \|\mathbf{y}\|_{(\mathbf{\Lambda}_k^i)^{-1}}, \ \forall \mathbf{y} \in \mathcal{X}_t^i\}.$$

Here, $\{\mathbf{\Lambda}_k^i\}_{k=0}^{m-1}$ and $\{\boldsymbol{\theta}_k^i\}_{k=0}^{m-1}$ are agent i's statistics used in computation of $\mathcal{X}_t^{(i)m}$ for $t \in [\mathcal{T}_{m-1} + 1 : \mathcal{T}_m]$. They are initialized to $\lambda \mathbf{I}$ and $\mathbf{0}$ and will be updated at the end of each batch (will be specified how shortly). Let π_0^i be an arbitrary initial policy used in the first batch. Throughout batch $m \in [M]$, agent i uses the same policy π_{m-1}^i to select actions from the surviving actions set. At the end of batch $m \in [M]$, agent $i \in [N]$ sends $\mathbf{u}_m^i = \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \mathbf{x}_t^i y_t^i$ to the server who broadcasts $\sum_{i=1}^N \mathbf{u}_m^i$ to all the agents. Then, agent i updates policy π_m^i (used in the next batch) and the following components that are key in

the construction of the surviving actions set in the next batch as follows:

$$\mathbf{\Lambda}_m^i = \lambda \mathbf{I} + \frac{NT_m}{2} \mathbb{E}_{\mathcal{X} \sim \mathcal{D}_m^i} \mathbb{E}_{\mathbf{x} \sim \pi_{m-1}^i(\mathcal{X})} [\mathbf{x}\mathbf{x}^\top], \tag{6.3}$$

$$\boldsymbol{\theta}_m^i = (\mathbf{\Lambda}_m^i)^{-1} \sum_{j=1}^N \mathbf{u}_m^j, \tag{6.4}$$

where $\lambda > 0$ is a regularization constant and when conditioned on the first $(m-1)$ batches, $\mathcal{D}_m^i$ is the distribution based on which the sets of surviving feature vectors $\mathcal{X}_t^{i(m)}$ for all $t \in [\mathcal{T}_{m-1} + 1 : \mathcal{T}_m]$ are generated.

Algorithm 8 DisBE-LUCB for agent i

Input: N, d, δ, T, M, λ

Initialization: $a = \sqrt{T}(NT/d)^{\frac{1}{2(2^{M-1}-1)}}$, $T_1 = T_2 = a\sqrt{d/N}$, $T_m = \lfloor a\sqrt{T_{m-1}} \rfloor$, $\boldsymbol{\theta}_0^i = \mathbf{0}$, $\mathbf{\Lambda}_0^i = \lambda \mathbf{I}$, $\mathcal{T}_0 = 0$, $\mathcal{T}_m = \mathcal{T}_{m-1} + T_m$, $\lambda = 5\log(4dT/\delta)$, $\beta = 6\sqrt{\log(2KNT/\delta)} + \sqrt{\lambda}$, arbitrary policy π_0^i

for $m = 1, \ldots, M$ **do**

  for $t = \mathcal{T}_{m-1} + 1, \ldots, \min\{\mathcal{T}_m, T\}$ **do**

    Construct $\mathcal{X}_t^{i(m)} = \cap_{k=0}^{m-1} \mathcal{E}\left(\mathcal{X}_t^i; (\mathbf{\Lambda}_k^i, \boldsymbol{\theta}_k^i, \beta)\right)$.

    Play arm $a_{i,t}$ associated with feature vector $\mathbf{x}_t^i \sim \pi_{m-1}^i\left(\mathcal{X}_t^{i(m)}\right)$ and observe y_t^i.

  end for

  Send $\mathbf{u}_m^i = \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \mathbf{x}_t^i y_t^i$ to the server.

  Receive $\sum_{j=1}^N \mathbf{u}_m^j$ from the server.

  Compute/construct $\mathbf{\Lambda}_m^i$ and $\boldsymbol{\theta}_m^i$ as in (6.3) and (6.4), respectively, $\mathcal{S}_m^i$ as in (6.5), and $\pi_m^i =$ ExplorationPolicy $\left(\frac{2\lambda}{NT_m}, \mathcal{S}_m^i\right)$, where ExplorationPolicyis presented in Appendix E.4.

end for

Statistics $\mathbf{\Lambda}_m^i$ and $\boldsymbol{\theta}_m^i$ are used in defining *new* confidence intervals in Lemma 9. We highlight that a direct use of existing standard confidence intervals in the literature such as the ones in [103] would fail to guarantee optimal communication cost $\tilde{\mathcal{O}}(dN)$ and require more communication by a factor of d [1]. Using matrix concentration inequalities,

[1] $d^2 + d$ values per agent, i.e., $\mathbf{u}_m^i$ and $\sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \mathbf{x}_t^i \mathbf{x}_t^{i\top}$.

we address this issue by replacing matrix $\lambda\mathbf{I} + \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{i=1}^{N} \mathbf{x}_t^i \mathbf{x}_t^{i\top}$, which would have been used if Algorithm 5 in [103] had been directly extended to a multi-agent one, with $\lambda\mathbf{I} + (NT_m/2)\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i}\mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]$. This allows agent i to communicate only d values ($\mathbf{u}_m^i$) while achieving $\tilde{\mathcal{O}}(\sqrt{dNT})$ regret as will be shown in Theorem 12. As the final step of batch m, agent i implements ExplorationPolicywith inputs $\frac{2\lambda}{NT_m}, \mathcal{S}_m^i$, where

$$\mathcal{S}_m^i = \{\mathcal{X}_t^{i(m+1)}\}_{t=\mathcal{T}_{m-1}+T_m/2+1}^{\mathcal{T}_m}. \tag{6.5}$$

ExplorationPolicy, which is presented in Algorithm 14 in Appendix E.4 and is inspired by Algorithm 3 in [103], computes policy π_m^i that will be used to select actions from the sets of surviving actions in the next batch. This choice of policy coupled with the definition of $\mathbf{\Lambda}_m^i$ in (6.3) guarantees that at all rounds $t \in [\mathcal{T}_1 + 1 : T]$, the length of the longest confidence interval in the surviving sets, which is an upper bound on the instantaneous regret of agent i at round t, can be bounded by $\mathcal{O}(\sqrt{d/NT})$. This allows us to achieve the optimal $\mathcal{O}(\sqrt{dNT})$ regret, while other exploration policies, such as the G-optimal design results in a $\mathcal{O}(d\sqrt{NT})$ regret.

6.4.2 Theoretical Results for DisBE-LUCB

We present our theoretical results for DisBE-LUCB, showing that it is nearly minimax optimal in terms of *both regret and communication cost*. The proof is given in Appendix E.2.

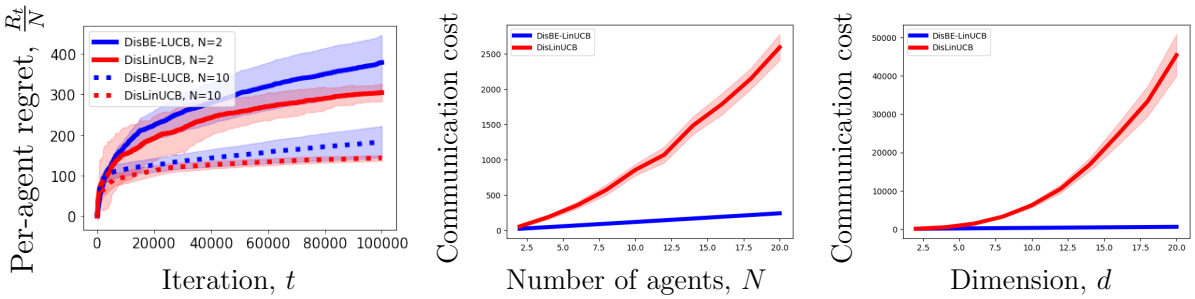
Theorem 12. *Fix $M = 1 + \log(\log(NT/d)/2 + 1)$ in Algorithm 8. Suppose Assumption 16 holds. If $T \geq \Omega(d^{22}\log^2(NT/\delta)\log^2 d\log^2(d\lambda^{-1}))$, then with probability at least $1-2\delta$, it holds that $R_T \leq \mathcal{O}\left(\sqrt{dNT\log d\log^2(KNT/\delta\lambda)}\log\log\left(NT/d\right)\right)$, and Communication Cost $\leq \mathcal{O}(dN\log\log(NT/d))$, where the communication cost is measured by the number of real numbers communicated by the agents.*

We remark that simple tricks may significantly reduce the exponent constant in constraint $T \geq d^{\mathcal{O}(1)}$. For example, first running a simpler version of DisBE-LUCB, in which the exploration policy is the G-optimal design $\pi^{\mathrm{G}}(\mathcal{X}_t^{i(m)})$, for $\sqrt{T/dN}$ rounds and then switching to DisBE-LUCB would reduce the exponent to 10.

Remark 7. *For the sake of Algorithm 8's presentation, we find it instructive to consider the communication cost as the number of real numbers communicated in the network. However, it is more realistic if we translate it into the total number of communicated bits. It would also allow us to make a fair comparison with the lower bound in Theorem 11 as it is stated in terms of number of communicated bits. Therefore, if we slightly modify Algorithm 8 such that instead of communicating vectors $\mathbf{u}_m^i$ in Line 8, agent i first rounds each entry of $\mathbf{u}_m^i$ with precision ϵ_0 and then sends the rounded vector to the server, then $\mathcal{O}(\log(1/\epsilon_0))$ number of bits is sufficient to communicate each entry of the rounded vectors $\mathbf{u}_m^i$. Our analysis in Appendix E.2.3 shows that compared to bounds in Theorem 12, by selecting $\epsilon_0 = \mathcal{O}(1/(N\sqrt{dT}))$, the communication cost of this slightly modified version of DisBE-LUCB, which is measured in bits, is $\mathcal{O}\left(dN \log\log\left(NT/d\right)\log(dNT)\right)$ and its regret is same as DisBE-LUCB's.*

Remark 8. *As mentioned in Section 6.4.1, a direct use of confidence intervals in [103] would fail to guarantee optimal communication cost $\tilde{\mathcal{O}}(dN)$ and require more communication by a factor of d. Thus, we use new confidence intervals (see Lemma 9) so that DisBE-LUCB would enjoy an optimal communication rate. The assumption on the knowledge of $\mathcal{D}$ is required in the computation of $\mathbf{\Lambda}_m^i$ in (6.3) used in these new confidence intervals. However, in practice, distribution $\mathcal{D}$ is not fully known and can only be estimated; therefore, $\mathbf{\Lambda}_m^i$ cannot be computed without any error. We relax this assumption and consider more realistic settings where each agent i can estimate matrix $\mathbf{\Lambda}_m^i$ in batch m up to an ϵ_m error, i.e., $(1-\epsilon_m)\mathbf{\Lambda}_m^i \preceq \tilde{\mathbf{\Lambda}}_m^i \preceq (1+\epsilon_m)\mathbf{\Lambda}_m^i$, where $\tilde{\mathbf{\Lambda}}_m^i$ is an estimation of $\mathbf{\Lambda}_m^i$ and $\epsilon_m \in (0,1)$[2]. In Appendix E.2.4, we show that for sufficiently small values of $\epsilon_m \leq 1/\sqrt{NT_m}$, a multiplicative factor $(1-\max_{m\in[M]}\epsilon_m)^{-1}$ appears in the regret bound while the communication cost remains unchanged.*

[2]*This is a weaker condition compared to the component-wise condition $(1-\epsilon_m)\mathbf{\Lambda}_m^i \leq \tilde{\mathbf{\Lambda}}_m^i \leq (1+\epsilon_m)\mathbf{\Lambda}_m^i$.*



(a) $d = 4$ **(b)** $d = 4$ **(c)** $N = 10$

Figure 6.1: The shaded regions show standard deviation around the mean. Standard deviation for communication cost of DisBE-LUCB is zero, because communication cost $= dNM$ and parameters determining M are known upfront (see Theorem 12).

6.4.3 Proof Sketch of Theorem 12

We first introduce the following lemma that constructs confidence intervals for the expected rewards.

Lemma 9 (Confidence intervals for DisBE-LUCB)**.** *Suppose Assumption 16 holds. For* $\delta \in (0,1)$*, let* $\beta = 6\sqrt{\log(2KNT/\delta)} + \sqrt{\lambda}$*. Then for all* $\mathbf{x} \in \mathcal{X}_t^i, i \in [N], t \in [T], m \in [M]$*, with probability at least* $1-\delta$*, it holds that* $\left|\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta}\rangle\right| \leq \beta \|\mathbf{x}\|_{(\boldsymbol{\Lambda}_m^i)^{-1}}$.

We prove this lemma by first employing appropriate matrix concentration inequalities to lower bound $\boldsymbol{\Lambda}_m^i$ by matrix $\frac{1}{2}\sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2}\sum_{i=1}^{N}\mathbf{x}_t^i{\mathbf{x}_t^i}^\top$. Carefully replacing $\boldsymbol{\Lambda}_m^i$ with its lower bound and using Azuma's inequality, we establish confidence intervals stated in the lemma. This lemma is key in ensuring an optimal communication rate $\tilde{\mathcal{O}}(dN)$, as a direct use of confidence intervals in [103] fails to guarantee optimal communication cost and requires $\tilde{\mathcal{O}}(d^2N)$ communication. See Appendix E.2.1 for proof.

Thanks to our choice of T_1 and T_2, and the fact that expected value of the rewards are bounded in $[-1,1]$, the regret of first two batches is bounded by $\mathcal{O}(\sqrt{dNT})$. For each batch $m \geq 3$, the confidence intervals imply that for all $t \in [\mathcal{T}_{m-1}+1 : \mathcal{T}_m]$, $\mathbf{x}_{t,*}^i \in \mathcal{X}_t^{i(m)}$ with high probability, and allow us to bound the instantaneous regret $r_t^i = \mathbb{E}[\langle \boldsymbol{\theta}, \mathbf{x}_{*,t}^i\rangle - \langle \boldsymbol{\theta}, \mathbf{x}_t^i\rangle]$ by $4\beta\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}[\max_{\mathbf{x}\in\mathcal{X}}\sqrt{\mathbf{x}^\top(\boldsymbol{\Lambda}_{m-1}^i)^{-1}\mathbf{x}}]$. Note that learning of

$\boldsymbol{\theta}_m^i$ and π_m^i are done through disjoint sets of samples, i.e., $\mathcal{A} = [\mathcal{T}_{m-1} + 1 : \mathcal{T}_{m-1} + T_m/2]$ and $\mathcal{B} = [\mathcal{T}_{m-1} + T_m/2 + 1 : \mathcal{T}_m]$, respectively. This is because $\mathcal{D}_m^i$ depends on $\boldsymbol{\theta}_m^i$, which is learned from $\mathcal{A}$, and we have to make $\mathcal{B}$ disjoint from $\mathcal{A}$ so as to ensure that elements in $\mathcal{S}_m^i$ are independently sampled from $\mathcal{D}_m^i$. Therefore, Theorem 5 in [103] guarantees that $\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}[\max_{\mathbf{x}\in\mathcal{X}} \sqrt{\mathbf{x}^\top(\boldsymbol{\Lambda}_{m-1}^i)^{-1}\mathbf{x}}] \leq \tilde{\mathcal{O}}(\sqrt{d/(NT_{m-1})})$. Finally, these combined with our choice of grid $\mathcal{T} = \{\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_M\}$ and $M = 1 + \log(\log(NT/d)/2 + 1)$ lead us to a regret bound $\tilde{\mathcal{O}}(\sqrt{dNT})$. Moreover, communications happen only at the end of each batch, whose number is M, and agents only share d-dimensional vectors $\mathbf{u}_m^i$. Therefore, communication cost is $dNM = \mathcal{O}(dN \log\log(NT/d))$.

6.4.4 Fully Decentralized Batch Elimination LUCB

In a scenario where there is no server and the agents are allowed to communicate *only* with their immediate neighbors, they can be represented by nodes of a graph. Applying a carefully designed consensus procedure that guarantees sufficient information mixing among the entire network, in Appendix E.3, we propose a fully decentralized version of DisBE-LUCB, called DecBE-LUCB. Communication cost of DecBE-LUCB is greater than DisBE-LUCB's by an extra multiplication factor $S = \log(dN)\delta_{\max}/\sqrt{1/|\lambda_2|}$, where $\delta_{\max}$ is the maximum degree of the network's graph and $|\lambda_2|$ is the second largest eigenvalue of the communication matrix in absolute value characterizing the graph's connectivity level. This is because at the last S rounds of each batch m, agents communicate each entry of their estimations of vector $\sum_{j=1}^N \mathbf{u}_m^j$ with their neighbors, whose number is at most $\delta_{\max}$, to ensure enough information mixing. Moreover, this results in DecBE-LUCB having no control over the regret of the mixing rounds, and therefore an additional term $\log(dN)NM/\sqrt{1/|\lambda_2|}$, which we call the *delay effect*, in the regret bound. Note that the more connected the graph is, the smaller $|\lambda_2|$ is. This aligns with the fact that the more connected the graph is, the less number of mixing rounds S is required. For example, fixing $N = 20$, for chain, ring, star, random Erdős–Renyi graph with parameter $p = 0.5$, and complete graphs, the values of $|\lambda_2|$ are 0.9918, 0.9674, 0.97, 0.67 (average over 100 instances), and 0, respectively. As expected, for less connected graphs (Chain, Ring, Star), $|\lambda_2|$ is close to 1 and for the fully connected graph $|\lambda_2| = 0$

and for a random graph $|\lambda_2|$ is not too small nor too large. The theoretical guarantees of DecBE-LUCB are summarized in table 6.1 and a detailed discussion is given in Appendix E.3.

6.5 Experiments

In this section, we present numerical simulations to confirm our theoretical findings. We evaluate the performance of DisBE-LUCB on synthetic data and compare it to that of DisLinUCB proposed by [134] that study the most similar setting to ours. The results shown in Figure 6.1 depict averages over 20 realizations, for which we have chosen $K = 20$, $\delta = 0.01$ and $T = 100000$. For each realization, the parameter $\boldsymbol{\theta}$ is drawn from $\mathcal{N}(0, I_d)$ and then normalized to unit norm and noise variables are zero-mean Gaussian random variables with variance 0.01. The decision set distribution $\mathcal{D}$ is chosen to be uniform over $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{X}}_2, \ldots, \tilde{\mathcal{X}}_{100}\}$, where each $\tilde{\mathcal{X}}_i$ is a set of K vectors drawn from $\mathcal{N}(0, I_d)$ and then normalized to unit norm. While implementing DisBE-LUCB, in order to compute $\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i}\mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]$ for agent i at batch m, we followed these steps: 1) for each $j \in [100]$, we built $\tilde{\mathcal{X}}_j^{i(m)} = \cap_{k=0}^{m-1}\mathcal{E}(\tilde{\mathcal{X}}_j; (\boldsymbol{\Lambda}_k^i, \boldsymbol{\theta}_k^i, \beta))$; 2) we took average over all 100 matrices $\frac{1}{100}\sum_{j\in[100]}\mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\tilde{\mathcal{X}}_j^{i(m)})}[\mathbf{x}\mathbf{x}^\top]$ as $\mathcal{D}$ is a uniform distribution over $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{X}}_2, \ldots, \tilde{\mathcal{X}}_{100}\}$. In Figure 6.1a, fixing $d = 4$, we compare the per-agent regret R_t/N of DisBE-LUCB and DisLinUCB for $t \in [T]$ and for different values of $N = 2$ and $N = 10$, where $R_t = \sum_{s=1}^t \sum_{i=1}^N \langle \boldsymbol{\theta}, \mathbf{x}_{*,s}^i \rangle - \langle \boldsymbol{\theta}, \mathbf{x}_s^i \rangle$. Figure 6.1b compares the communication cost of DisBE-LUCB and DisLinUCB over T rounds when both algorithms are implemented for fixed $d = 4$, and N varying from 2 to 20. Finally, Figure 6.1c compares the communication cost of DisBE-LUCB and DisLinUCB over T rounds when both algorithms are implemented for fixed $N = 10$, and d varying from 2 to 20. From these three comparisons, we conclude that DisBE-LUCB achieves a regret comparable with DisLinUCB, at a significantly smaller communication rate. The curves in Figures 6.1b and 6.1c verify the linear dependency of DisBE-LUCB's communication cost on N and d while communication cost of DisLinUCB grows super-linearly with N and d (see Table 6.1 for theoretical comparisons). Moreover, Figure 6.1a emphasizes the value of collaboration in speeding up the learning process. As the

number of agents increases, each agent learns the environment faster as an individual.

6.6 Conclusion

We proved an information-theoretic lower bound on the communication cost of any algorithm achieving an optimal regret rate for the distributed contextual linear bandit problem with stochastic contexts. We then proposed DisBE-LUCB with optimal regret $\tilde{\mathcal{O}}(\sqrt{dNT})$ and communication cost $\tilde{\mathcal{O}}(dN)$ which (nearly) matches our lower bound and improves upon the previous best-known algorithms whose communication cost scale super linearly either in d or N. Finally, we proposed DecBE-LUCB, a fully decentralized variant of DisBE-LUCB, without a central server where the agents can only communicate with their immediate neighbors given by a communication graph. We showed that the structure of the network affects the regret performance via a small additive term that depends on the spectral gap of the underlying graph, while the communication cost still grows linearly with d and N. As shown in Table 6.1, the best communication cost achieved for settings with *adversarially* varying contexts over time horizon and agents is $\mathcal{O}(d^3N^{1.5})$. There is no formal theory proving such bounds are optimal for the adversarial context case. While our work provides optimal theoretical guarantees for stochastically varying contexts, it is not clear how to generalize these *optimal* results to settings with adversarially varying contexts. Therefore, an important future direction is to design optimal algorithms and prove communication cost lower bounds for scenarios with adversarial contexts.

APPENDIX A

Proofs for Chapter 2

A.1 Proof of Lemma 1

In order to bound the minimum eigenvalue of the Gram matrix at round $T'+1$, we use the Matrix Chernoff Inequality [125, Thm. 5.1.1].

Theorem 13 (Matrix Chernoff Inequality, [125])**.** *Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, symmetric matrices in $\mathbb{R}^d$. Assume that $\lambda_{min}(\mathbf{X}_k) \geq 0$ and $\lambda_{max}(\mathbf{X}_k) \leq L$ for each index k. Introduce the random matrix $\mathbf{Y} = \sum_k \mathbf{X}_k$. Let μ_{min} denote the minimum eigenvalue of the expectation $\mathbb{E}[Y]$,*

$$\mu_{min} = \lambda_{min}\left(\mathbb{E}[\mathbf{Y}]\right) = \lambda_{min}\left(\sum_k E[\mathbf{X}_k]\right).$$

Then, for any $\epsilon \in (0,1)$, it holds,

$$\Pr\left(\lambda_{min}(\mathbf{Y}) \leq \epsilon\mu_{min}\right) \leq d \cdot \exp\left(-(1-\epsilon)^2\frac{\mu_{min}}{2L}\right).$$

Proof of Lemma 1. Let $\mathbf{X}_t = \mathbf{x}_t\mathbf{x}_t^\top$ for $t \in [T']$, such that each $\mathbf{X}_t$ is a symmetric matrix with $\lambda_{\min}(\mathbf{X}_t) \geq 0$ and $\lambda_{\max}(\mathbf{X}_t) \leq L^2$. In this notation, $\mathbf{A}_{T'+1} = \lambda\mathbf{I} + \sum_{t=1}^{T'}\mathbf{X}_t$. In order to apply Theorem 13, we compute:

$$\mu_{\min} := \lambda_{\min}\left(\sum_{t=1}^{T'}\mathbb{E}[\mathbf{X}_t]\right) = \lambda_{\min}\left(\sum_{t=1}^{T'}\mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top]\right) = \lambda_{\min}\left(T'\Sigma\right) = \lambda_- T'.$$

Thus, the theorem implies the following for any $\epsilon \in [0,1)$:

$$\Pr\left[\lambda_{\min}(\sum_{t=1}^{T'}\mathbf{X}_t) \leq \epsilon\lambda_- T'\right] \leq d \cdot \exp\left(-(1-\epsilon)^2\frac{\lambda_- T'}{2L^2}\right). \tag{A.1}$$

To complete the proof of the lemma, simply choose $\epsilon = 0.5$ (say) and $T' \geq \frac{8L^2}{\lambda_-}\log(\frac{d}{\delta})$ in (A.1). This gives $\Pr\left[\lambda_{\min}(\mathbf{A}_{T'+1}) \geq \lambda + \frac{\lambda_- T'}{2}\right] \geq 1-\delta$, as desired. □

A.2 Proof of Theorems 2 and 3

In this section, we present the proofs of Theorems 2 and 3.

A.2.1 Preliminaries

Conditioning on $\boldsymbol{\mu} \in \mathcal{C}_t,\ \forall t > 0$. Consider the event

$$\mathcal{E} := \{\boldsymbol{\mu} \in \mathcal{C}_t,\ \forall t > 0\}, \tag{A.2}$$

that $\boldsymbol{\mu}$ is inside the confidence region for all times t. By Theorem 1 the event holds with probability $1-\delta$. Onwards, we condition on this event, and we make repeated use of the fact that $\boldsymbol{\mu} \in \mathcal{C}_t$ for all $t > 0$, without further explicit reference.

Decomposing the regret in two terms. Recall the decomposition of the instantaneous regret in two terms in (2.10) as follows:

$$r_t = \boldsymbol{\mu}^\top \mathbf{x}_t - \boldsymbol{\mu}^\top \mathbf{x}^* = \underbrace{\boldsymbol{\mu}^\top \mathbf{x}_t - \tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}_t}_{\text{Term I}} + \underbrace{\tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}_t - \mu^\top \mathbf{x}^*}_{\text{Term II}}. \tag{A.3}$$

As discussed in Section 2.5.1, we control the two terms separately.

A.2.2 Bounding Term I

The results in this subsection are by now rather standard in the literature (see for example [2,]). We provide the necessary details for completeness.

We start with the following chain of inequalities, that hold for all $t \geq T' + 1$:

$$\begin{aligned}
\text{Term I} &:= \boldsymbol{\mu}^\top \mathbf{x}_t - \tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}_t = (\boldsymbol{\mu}^\top \mathbf{x}_t - \hat{\boldsymbol{\mu}}_t^\top \mathbf{x}_t) + (\hat{\boldsymbol{\mu}}_t^\top \mathbf{x}_t - \tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}_t) \\
&\leq \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_t\|_{\mathbf{A}_t} \|\mathbf{x}_t\|_{\mathbf{A}_t^{-1}} + \|\hat{\boldsymbol{\mu}}_t - \tilde{\boldsymbol{\mu}}_t\|_{\mathbf{A}_t} \|\mathbf{x}_t\|_{\mathbf{A}_t^{-1}} \\
&\leq 2\beta_t \|\mathbf{x}_t\|_{\mathbf{A}_t^{-1}}.
\end{aligned} \tag{A.4}$$

The last inequality (A.4) follows from Theorem 1 and the fact that μ and $\tilde{\boldsymbol{\mu}}_t \in \mathcal{C}_t$. Recall, from Assumption 2, the trivial bound on the instantaneous regret

$$r_t = \boldsymbol{\mu}^\top \mathbf{x}_t - \boldsymbol{\mu}^\top \mathbf{x}^* \leq 2.$$

Thus, we conclude with the following

$$\text{Term I} \leq 2 \min\left(\beta_t \|\mathbf{x}_t\|_{\mathbf{A}_t^{-1}}, 1\right). \tag{A.5}$$

The next lemma bounds the total contribution of the (squared) terms in (A.4) across the entire horizon $t = T' + 1, \ldots, T$.

Lemma 10 (Term I)**.** *Let Assumptions 1 and 2 hold. Fix any $\delta \in (0, 0.5)$ and assume that T' is such that $T' \geq \frac{8L^2}{\lambda_-} \log\left(\frac{d}{\delta}\right)$. Then, with probability at least $1 - \delta$, it holds*

$$\sum_{t=T'+1}^{T} \min\left(\|\mathbf{x}_t\|^2_{\mathbf{A}_t^{-1}}, 1\right) \leq 2d \log\left(\frac{2TL^2}{d(2\lambda + \lambda_- T')}\right).$$

Thus, with probability at least $1 - 2\delta$, it holds

$$\sum_{t=T'+1}^{T} \left(\boldsymbol{\mu}^\top \mathbf{x}_t - \tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}_t\right) \leq 2\beta_t \sqrt{2d\,(T - T')\, \log\left(\frac{2TL^2}{d\,(2\lambda + \lambda_- T')}\right)}. \tag{A.6}$$

Proof. The proof is mostly adapted from [40, Lem. 9] but we also exploit the bound on $\lambda_{\min}(\mathbf{A}_{T'+1})$ thanks to Lemma 1. We present the details for the reader's convenience.

With probability at least $1 - \delta$, we find that for all $t \geq T' + 1$:

$$\begin{aligned}
\det(\mathbf{A}_{t+1}) &= \det(\mathbf{A}_t + \mathbf{x}_t \mathbf{x}_t^\top) = \det(\mathbf{A}_t) \det(I + (\mathbf{A}_t^{-\frac{1}{2}} \mathbf{x}_t)(\mathbf{A}_t^{-\frac{1}{2}} \mathbf{x}_t)^\top) = \det(\mathbf{A}_t)(1 + \|\mathbf{x}_t\|^2_{\mathbf{A}_t^{-1}}) \\
&= \ldots = \det(\mathbf{A}_{T'+1}) \prod_{\tau=T'+1}^{t} (1 + \|\mathbf{x}_\tau\|^2_{\mathbf{A}_\tau^{-1}}) \\
&\geq \left(\lambda + \frac{\lambda_- T'}{2}\right)^d \prod_{\tau=T'+1}^{t} (1 + \|\mathbf{x}_\tau\|^2_{\mathbf{A}_\tau^{-1}}),
\end{aligned}$$

where the last inequality follows form Lemma 1 and the fact that $\det(A) = \prod_{i=1}^d \lambda_i(A) \geq (\lambda_{\min}(A))^d$. Furthermore, by the AM-GM inequality applied to the eigenvalues of A_{t+1}, if holds

$$\det(\mathbf{A}_{t+1}) = \prod_{i=1}^{d} \lambda_i(\mathbf{A}_{t+1}) \leq \left(\frac{tL^2}{d}\right)^d,$$

where we also used the fact that $\|\mathbf{x}_t\|_2 \leq L$ for all t. These combined yield,

$$\prod_{\tau=T'+1}^{t} (1 + \| x_\tau \|^2_{\mathbf{A}_\tau^{-1}}) \leq \left(\frac{2tL^2}{d(2\lambda + \lambda_- T')} \right)^d .$$

Next, using the fact that for any $0 \leq y \leq 1$, $\log(1+y) \geq y/2$, we have

$$\begin{aligned} \sum_{t=T'+1}^{T} \min \left(\|\mathbf{x}_t\|^2_{\mathbf{A}_t^{-1}}, 1 \right) &\leq 2 \sum_{t=T'+1}^{T} \log \left(\|\mathbf{x}_t\|^2_{\mathbf{A}_t^{-1}} + 1 \right) = 2 \log \Big(\prod_{t=T'+1}^{T} \left(\|\mathbf{x}_t\|^2_{\mathbf{A}_t^{-1}} + 1 \right) \Big) \\ &\leq 2d \log \left(\frac{2TL^2}{d(2\lambda + \lambda_- T')} \right). \end{aligned}$$

It remains to prove (A.6). Recall from (A.5) that for any $T' < t \leq T$, with probability at least $1 - \delta$ (note that we have conditioned in the event $\mathcal{E}$ in (A.2)),

$$\left(\boldsymbol{\mu}^\top \mathbf{x}_t - \tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}_t \right) \leq 2 \min \left(\beta_t \|\mathbf{x}_t\|_{\mathbf{A}_t^{-1}}, 1 \right) \leq 2\beta_t \min \left(\|\mathbf{x}_t\|_{\mathbf{A}_t^{-1}}, 1 \right),$$

where for the inequality we have used the fact that $\beta_t \leq \beta_t$ (and assumed for simplicity that T large enough such that $\beta_t > 1$). Thus, the desired bound in (A.6) follows from applying Cauchy-Schwartz inequality to the above. □

A.2.3 Bounding Term II

As discussed in Section 2.5.2, the challenge in bounding Term II in (2.10) is that , in general, $\mathcal{D}_t^{\mathrm{s}} \neq \mathcal{D}^{\mathrm{s}}$, so $\mathbf{x}^*$ might not belong in $\mathcal{D}_t^{\mathrm{s}}$. Bounding Term II amounts to bounding a certain "distance" of the set $\mathcal{D}_t^{\mathrm{s}}$ from the set $\mathcal{D}_0$. In order to accomplish this task, we proceed as follows. First, we define a shrunk version $\widetilde{\mathcal{D}}_t^{\mathrm{s}}$ of $\mathcal{D}_t^{\mathrm{s}}$, for which we have a more convenient characterization, compared to the original $\widetilde{\mathcal{D}}_t^{\mathrm{s}}$. Then, we select the point z_t in $\widetilde{\mathcal{D}}_t^{\mathrm{s}}$ that is in the direction of $\mathbf{x}^*$ and is as close to it as possible. Finally, we are able to bound the distance of z_t to $\mathbf{x}^*$.

A shrunk safe region $\widetilde{\mathcal{D}}_t^{\mathrm{s}}$. Consider an enlarged confidence region $\tilde{\mathcal{C}}_t$ centered at μ defined as follows:

$$\tilde{\mathcal{C}}_t := \{ v \in \mathbb{R}^d : \|\boldsymbol{\nu} - \boldsymbol{\mu}\|_{\mathbf{A}_t} \leq 2\beta_t \} \supseteq \mathcal{C}_t. \tag{A.7}$$

The inclusion property above holds since $\boldsymbol{\mu} \in \mathcal{C}_t$, and, by triangle inequality, for all $v \in \mathcal{C}_t$, one has that $\|\boldsymbol{\nu} - \boldsymbol{\mu}\|_{\mathbf{A}_t} \leq \|\boldsymbol{\nu} - \hat{\boldsymbol{\mu}}_t\|_{\mathbf{A}_t} + \|\hat{\boldsymbol{\mu}}_t - \boldsymbol{\mu}\|_{\mathbf{A}_t} \leq 2\beta_t$.

The definition of the enlarged confidence region in (A.7) naturally leads to the definition of a corresponding shrunk safe decision set $\widetilde{\mathcal{D}}_t^{\mathrm{s}}$. Namely, let

$$\begin{aligned}\widetilde{\mathcal{D}}_t^{\mathrm{s}} &:= \{\mathbf{x} \in \mathcal{D}_0 : \boldsymbol{\nu}^\top \mathbf{B}\mathbf{x} \leq c,\ \forall v \in \tilde{\mathcal{C}}_t\} = \{\mathbf{x} \in \mathcal{D}_0 : \max_{v \in \tilde{\mathcal{C}}_t} \boldsymbol{\nu}^\top \mathbf{B}\mathbf{x} \leq c\} \\ &= \{\mathbf{x} \in \mathcal{D}_0 : \boldsymbol{\mu}^\top \mathbf{B}\mathbf{x} + 2\beta_t \|\mathbf{B}\mathbf{x}\|_{\mathbf{A}_t^{-1}} \leq c\}, \end{aligned} \tag{A.8}$$

and observe that $\widetilde{\mathcal{D}}_t^{\mathrm{s}} \subseteq \mathcal{D}_t^{\mathrm{s}}$. Note here that since by Assumption 3 zero is in the interior of $\mathcal{D}_0$, the sets $\tilde{\mathcal{D}}_t^{\mathrm{s}}$ and $\mathcal{D}_t^{\mathrm{s}}$ have a nonempty interior.

A point $z_t \in \widetilde{\mathcal{D}}_t^{\mathrm{s}}$ close to $\mathbf{x}^*$. Let z_t be a vector in the direction of $\mathbf{x}^*$ that belongs in $\widetilde{\mathcal{D}}_t^{\mathrm{s}}$ and is closest to $\mathbf{x}^*$. Formally, $z_t := \alpha_t \mathbf{x}^*$, where

$$\alpha_t := \max\left\{\alpha \in [0,1] \,|\, z_t = \alpha \mathbf{x}^* \in \widetilde{\mathcal{D}}_t^{\mathrm{s}}\right\}.$$

Since both 0 and $\mathbf{x}^* \in \mathcal{D}_0$, and, $\mathcal{D}_0$ is convex by assumption, it follows in view of (A.8) that

$$\alpha_t := \max\left\{\alpha \in [0,1] \,|\, \alpha \cdot \left(\boldsymbol{\mu}^\top \mathbf{B}\mathbf{x}^* + 2\beta_t \|\mathbf{B}\mathbf{x}^*\|_{\mathbf{A}_t^{-1}}\right) \leq c\right\}. \tag{A.9}$$

Recall that $C > 0$, thus (A.9) can be simplified to the following:

$$\alpha_t = \begin{cases} 1 & \text{, if} \quad \boldsymbol{\mu}^\top \mathbf{B}\mathbf{x}^* + 2\beta_t \|\mathbf{B}\mathbf{x}^*\|_{\mathbf{A}_t^{-1}} \leq c, \\ \min\left(\frac{c}{\boldsymbol{\mu}^\top \mathbf{B}\mathbf{x}^* + 2\beta_t \|\mathbf{B}\mathbf{x}^*\|_{\mathbf{A}_t^{-1}}}, 1\right) & \text{, otherwise.} \end{cases} \tag{A.10}$$

Bounding Term II in terms of α_t. Due to the fact that $\widetilde{\mathcal{D}}_t^{\mathrm{s}} \subseteq \mathcal{D}_t^{\mathrm{s}}$, it holds that $z_t \in \mathcal{D}_t^{\mathrm{s}}$. Using this, and optimality of $(\tilde{\mu}, \mathbf{x}_t)$ in the minimization in Step 10 of Algorithm 1, we can bound Term II as follows:

$$\begin{aligned} \text{Term II} &:= \tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}_t - \mu^\top \mathbf{x}^* \\ &\leq \boldsymbol{\mu}^\top z_t - \mu^\top \mathbf{x}^* = \alpha_t \boldsymbol{\mu}^\top \mathbf{x}^* - \boldsymbol{\mu}^\top \mathbf{x}^* \\ &\leq |\alpha_t - 1|\, |\boldsymbol{\mu}^\top \mathbf{x}^*| \\ &\leq |\alpha_t - 1| = (1 - \alpha_t). \end{aligned} \tag{A.11}$$

The inequality in the last line uses Assumption 2. For the last equality recall that $\alpha_t \in [0,1]$

To proceed further from (A.11) we consider separately the two cases $\Delta > 0$ and $\Delta = 0$ that lead to Theorems 2 and 3, respectively.

A.2.3.1 Bound for the Case $\Delta > 0$

Here, assuming that $\Delta > 0$, we prove that if the duration T' of the pure exploration phase of Safe-LUCB is chosen appropriately, then $\alpha_t = 1$, and equivalently, $\mathbf{x}^* \in \mathcal{D}_t^{\mathrm{s}}$. The precise statement is given in Lemma 11 below, which is a restatement of Lemma 2, given here for the reader's convenience.

Lemma 11 ($\Delta > 0 \implies \mathbf{x}^* \in \mathcal{D}_t^{\mathrm{s}}$)**.** *Let Assumptions 1, 2 and 3 hold for all $t \in [T]$. Fix any $\delta \in (0, 0.5)$ and assume a positive safety gap $\Delta > 0$. Initialize Safe-LUCB with*

$$T' \geq \left(\frac{8L^2 \|\mathbf{B}\|^2 \beta_T^2}{\lambda_- \, \Delta^2} - \frac{2\lambda}{\lambda_-} \right) \vee t_\delta. \tag{A.12}$$

Then, with probability at least $1 - 2\delta$, for all $t = T' + 1, \ldots, T$ it holds that

$$\textit{Term II} := \tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}_t - \boldsymbol{\mu}^\top \mathbf{x}^* \leq 0.$$

Thus, with the same probability

$$\sum_{t=T'+1}^{T} (\tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}_t - \boldsymbol{\mu}^\top \mathbf{x}^*) \leq 0. \tag{A.13}$$

Proof. Recall from (A.11), that for any $T' < t \leq T$, with probability at least $1 - \delta$ (note that we have conditioned in the event $\mathcal{E}$ in (A.2)), Term II $= 1 - \alpha_t$. Thus, in view of (A.10), it suffices to prove that for any $T' < t \leq T$, with probability at least $1 - \delta$, it holds $\alpha_t = 1$, or equivalently,

$$\boldsymbol{\mu}^\top \mathbf{B}\mathbf{x}^* + 2\beta_t \|\mathbf{B}\mathbf{x}^*\|_{\mathbf{A}_t^{-1}} \leq c \quad \Leftrightarrow \quad \beta_t \|\mathbf{B}\mathbf{x}^*\|_{\mathbf{A}_t^{-1}} \leq \Delta/2. \tag{A.14}$$

For any $T' < t \leq T$, we have

$$\beta_t \|\mathbf{B}\mathbf{x}^*\|_{A_t^{-1}} \leq \frac{\beta_t \|\mathbf{B}\mathbf{x}^*\|_2}{\sqrt{\lambda_{\min}(\mathbf{A}_t)}} \leq \frac{\beta_T \|\mathbf{B}\mathbf{x}^*\|_2}{\sqrt{\lambda_{\min}(\mathbf{A}_{T'+1})}} \leq \frac{\beta_T \|B\| L}{\sqrt{\lambda_{\min}(\mathbf{A}_{T'+1})}}, \tag{A.15}$$

where, in the second inequality we used $\beta_t \leq \beta_t$ and $\lambda_{\min}(\mathbf{A}_t) \geq \lambda_{\min}(\mathbf{A}_{T'+1})$, and in the last inequality we used Assumption 2. Next, since $t_\delta \leq T'$, we may apply Lemma 1 to find from (A.15), that for all $T'+1 \leq t \leq T$, with probability at least $1-\delta$:

$$\beta_t \|\mathbf{B}\mathbf{x}^*\|_{A_t^{-1}} \leq \frac{\sqrt{2}\|B\|L\beta_T}{\sqrt{2\lambda + \lambda_- T'}}. \tag{A.16}$$

To complete the proof of the lemma note that the assumption $T' \geq \frac{8\|B\|^2 L^2 \beta_T^2}{\lambda_- \Delta^2} - \frac{2\lambda}{\lambda_-}$ when combined with (A.16), it guarantees (A.14), as desired. □

Remark 9. *We remark on a simple tweak in the algorithm that results in a constant T' (i.e., independent of T) in Lemma 11. However, this does* not *change the final order of regret bound in Theorem 2. In particular, we modify Safe-LUCB to use the nested (as is called in [68]) confidence region $\mathcal{B}_t = \cap_{\tau=1}^t \mathcal{C}_\tau$ at round t such that $\ldots \subseteq \mathcal{B}_{t+1} \subseteq \mathcal{B}_t \subseteq \mathcal{B}_{t-1} \subseteq \ldots$. According to Theorem 1, it is guaranteed that for all $t > 0$, $\boldsymbol{\mu} \in \mathcal{B}_t$, with high probability. Applying these nested confidence regions in creating safe sets, results in $\cdots \subseteq \mathcal{D}_{t-1}^s \subseteq \mathcal{D}_t^s \subseteq \mathcal{D}_{t+1}^s \subseteq \ldots$. Thanks to this, it is now guaranteed that once $\mathbf{x}^* \in \mathcal{D}_t^s$, the optimal action $\mathbf{x}^*$ will remain inside the safe decision sets for all rounds after t. Thus, it is sufficient to find the first round t, such that $\mathbf{x}^* \in \mathcal{D}_t^s$. This leads to a shorter duration T' for the pure exploration phase. In particular, following the arguments in Lemma 11, it can be shown that T' becomes the smallest value satisfying $2\sqrt{2}\|B\|L\beta_{T'} \leq \Delta\sqrt{2\lambda + \lambda_- T'}$, which is now a constant independent of T.*

A.2.3.2 Bound for the Case $\Delta = 0$

Lemma 12 (Term II for $\Delta = 0$)**.** *Let Assumptions 1, 2 and 3 hold. Fix any $\delta \in (0, 0.5)$ and assume that T' is such that $T' \geq t_\delta$. Then, with probability at least $1-\delta$, it holds*

$$\sum_{t=T'+1}^{T} 1 - \alpha_t \leq \frac{2\sqrt{2}\|B\|L\beta_t(T-T')}{c\sqrt{2\lambda + \lambda_- T'}}. \tag{A.17}$$

Therefore, with probability at least $1-2\delta$, it holds

$$\sum_{t=T'+1}^{T} (\tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}_t - \boldsymbol{\mu}^\top \mathbf{x}^*) \leq \frac{2\sqrt{2}\|B\|L\beta_t(T-T')}{c\sqrt{2\lambda + \lambda_- T'}}. \tag{A.18}$$

Proof. Recall from (A.11), that for any $T' < t \leq T$, with probability at least $1-\delta$ (note that we have conditioned in the event $\mathcal{E}$ in (A.2)), Term II $= 1-\alpha_t$. Thus, (A.18) directly follows once we show (A.17). In what follows, we prove (A.17).

The definition of α_t in (A.10) and the fact that $\boldsymbol{\mu}^\top \mathbf{B}\mathbf{x}^* \leq c$ imply that

$$\alpha_t = \begin{cases} 1 & \text{, if } \boldsymbol{\mu}^\top \mathbf{B}\mathbf{x}^* + 2\beta_t \|\mathbf{B}\mathbf{x}^*\|_{\mathbf{A}_t^{-1}} \leq c, \\ \frac{c}{\boldsymbol{\mu}^\top \mathbf{B}\mathbf{x}^* + 2\beta_t \|\mathbf{B}\mathbf{x}^*\|_{\mathbf{A}_t^{-1}}} \geq \frac{c}{c+2\beta_t \|\mathbf{B}\mathbf{x}^*\|_{\mathbf{A}_t^{-1}}} & \text{, otherwise.} \end{cases}$$

Thus, for all $t \geq T'+1$:

$$\alpha_t \geq \frac{c}{c + 2\beta_t \|\mathbf{B}\mathbf{x}^*\|_{\mathbf{A}_t^{-1}}},$$

from which it follows,

$$1-\alpha_t \leq \frac{2\beta_t \|\mathbf{B}\mathbf{x}^*\|_{\mathbf{A}_t^{-1}}}{c + 2\beta_t \|\mathbf{B}\mathbf{x}^*\|_{\mathbf{A}_t^{-1}}} \leq \frac{2\beta_t}{c} \|\mathbf{B}\mathbf{x}^*\|_{\mathbf{A}_t^{-1}} \leq \frac{2\beta_t \|\mathbf{B}\mathbf{x}^*\|_2}{c\sqrt{\lambda_{\min}(\mathbf{A}_t)}} \leq \frac{2\beta_t \|\mathbf{B}\| L}{c\sqrt{\lambda_{\min}(\mathbf{A}_{T'+1})}}.$$

The last two inequalities follow as in (A.15). To complete the proof, note that since $T' \geq t_\delta$, we can apply Lemma 1. Thus, with probability at least $1-\delta$ it holds,

$$\sum_{t=T'+1}^{T} 1-\alpha_t \leq \frac{2\beta_t \|B\| L (T-T')}{c\sqrt{\lambda_{\min}(\mathbf{A}_{T'+1})}} \leq \frac{2\sqrt{2}\|\mathbf{B}\| L \beta_t (T-T')}{c\sqrt{2\lambda + \lambda_- T'}},$$

as desired. □

A.2.4 Completing the Proof of Theorem 2

We are now ready to complete the proof of Theorem 2. Let T sufficiently large such that

$$T > T' \geq \left(\frac{8L^2 \|\mathbf{B}\|^2 \beta_T^2}{\lambda_- \Delta^2} - \frac{2\lambda}{\lambda_-} \right) \vee t_\delta. \tag{A.19}$$

We combine Lemma 10 (specifically, Eqn. (A.6)), Lemma 11 (specifically, Eqn. (A.13)), and, the decomposition in (A.3), to conclude that

$$R_T = \sum_{t=1}^{T'} r_t + \sum_{t=T'+1}^{T} r_t \leq 2T' + 2\beta_t \sqrt{2d(T-T') \log\left(\frac{2TL^2}{d(2\lambda + \lambda_- T')} \right)}.$$

Specifically, choosing $T' = \left(\frac{8L^2\|B\|^2\beta_T^2}{\lambda_- \Delta^2} - \frac{2\lambda}{\lambda_-} \right) \vee t_\delta$ in the above, results in

$$R_T = \mathcal{O}\left(\frac{\|B\|^2}{\lambda_- \Delta^2} d\sqrt{T} \log T \right), \tag{A.20}$$

where the constant in the Big-O notation may only depend on L, S, R, λ and δ.

A.2.5 Completing the proof of Theorem 3

We are now ready to complete the proof of Theorem 3. Let T sufficiently large such that

$$T > T' \geq t_\delta.$$

We combine Lemma 10 (specifically, Eqn. (A.6)), Lemma 12 (specifically, Eqn. (A.18)), and, the decomposition in (A.3), to conclude that

$$R_T = \sum_{t=1}^{T'} r_t + \sum_{t=T'+1}^{T} r_t \leq 2T' + 2\beta_t \sqrt{2d(T-T') \log\left(\frac{2TL^2}{d(2\lambda + \lambda_- T')}\right)} + \frac{2\sqrt{2}\|B\| L \beta_t (T-T')}{c\sqrt{2\lambda + \lambda_- T'}}.$$

Specifically, choosing $T' = \left(\frac{\|\mathbf{B}\| L \beta_t T}{c\sqrt{2\lambda_-}}\right)^{\frac{2}{3}} \vee t_\delta$ in the above, results in

$$R_T = \mathcal{O}\left(\left(\frac{\|\mathbf{B}\|}{c}\right)^{\frac{2}{3}} \lambda_-^{-1/3} d\, T^{2/3} \log T\right), \tag{A.21}$$

where as in (A.20) the constant in the Big-O notation may only depend on L, S, R, λ and δ.

A.3 Extension to linear Contextual Bandits

In this section, we present an extension to the setting of K-armed contextual bandit. At each round $t \in [T]$, the learner observes a context consisting of K action vectors, $\{\mathbf{y}_{t,a} : a \in [K]\} \subset \mathbb{R}^d$ and chooses one action denoted by a_t and observes its associated loss, $\ell_t = \boldsymbol{\mu}^\top \mathbf{y}_{t,a_t} + \eta_t$. We consider the same constraint (2.1) which results in a *safe* set of actions at each round $\{\mathbf{y}_{t,a} \,|\, a \in [K], \boldsymbol{\mu}^\top \mathbf{B} \mathbf{y}_{t,a} \leq c\}$. The optimal action at round t is denoted by y_{t,a_t^*} where

$$a_t^* \in \operatorname*{arg\,min}_{a \in [K], \mu^\top \mathbf{B} \mathbf{y}_{t,a} \leq c} \mu^\top \mathbf{y}_{t,a}. \tag{A.22}$$

If the chosen action at round t is denoted by $\mathbf{x}_t := \mathbf{y}_{t,a_t}$ and the optimal one by $\mathbf{x}_t^* := \mathbf{y}_{t,a_t^*}$, the cumulative regret over total T rounds will be

$$R_T = \sum_{t=1}^{T} \boldsymbol{\mu}^\top \mathbf{x}_t - \boldsymbol{\mu}^\top \mathbf{x}_t^*.$$

We briefly discuss how Safe-LUCB extends to the K-armed contextual setting with provable regret guarantees under the following assumptions.

First, we need the standard Assumptions 1 and 2 that naturally extend to the linear contextual bandit setting. Beyond these, in order for the safe-bandit problem to be well-defined, we assume that safe actions exist at each round. Equivalently, the feasible set in (A.22) is nonempty and x_t^* is well-defined. Moreover, in order to be able to run the pure-exploration phase of Safe-LUCB with random actions (that guarantee Lemma 1 holds) we further require that at least one of these safe actions is randomly sampled at each round t (technically, we need this assumption to hold only for rounds $1, \dots, T'$). These two assumptions are both implied by Assumption 17 below.

Assumption 17 (Nonempty safe sets)**.** *Consider the set* $\mathcal{D}^w = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{B}\mathbf{x}\|_2 \leq \frac{c}{S}\}$. *Then, at each round* t, $N_t \geq 1$ *number of* K *action vectors lie within* $\mathcal{D}^w$.

Finally, in order to guarantee that Safe-LUCB has sub-linear regret for the K-armed linear setting we need that the safety gap at each round is strictly positive.

Assumption 18 (Nonzero Δ)**.** *The safety gap* $\Delta_t = c - \boldsymbol{\mu}^\top \mathbf{B}\mathbf{x}_t^*$ *at each round* t *is positive.*

Under these assumptions, Safe-LUCB naturally extends to the K-armed linear bandit setting. Specifically, at rounds $t \leq T'$, Safe-LUCB randomly selects x_t to be one of the available N_t action vectors that belong to the set $\mathcal{D}^w$. Assume that $\lambda_{\min}(\mathbb{E}[\mathbf{x}_t \mathbf{x}_t^\top]) \geq \lambda_- > 0$ for all $t \in [T']$.

After round T', Safe-LUCB implements the safe exploration-exploitation phase by choosing safe actions based on OFU principle as in (2.9). Therefore line 10 of Safe-LUCB changes to

$$a_t = \arg\min_{a \in \mathcal{A}_t^{\mathrm{s}}} \min_{\boldsymbol{\nu} \in \mathcal{C}_t} \boldsymbol{\nu}^\top \mathbf{y}_{t,a}, \tag{A.23}$$

where the safe set at rounds $t \geq T' + 1$ is defined by

$$\mathcal{A}_t^{\mathrm{s}} = \{a \in [K] : \boldsymbol{\nu}^\top \mathbf{B}\mathbf{y}_{t,a} \leq c, \forall \boldsymbol{\nu} \in \mathcal{C}_t\}. \tag{A.24}$$

With these and subject to Assumptions 1, 2, 17 and 18, it is straightforward to extend the results of Theorem 2 to the setting considered here. Namely, under these assumptions, Safe-LUCB achieves regret $\tilde{\mathcal{O}}(\sqrt{T})$ when T' is set to T_Δ as in (2.13) for $\Delta = \min_{t \in [T]} \Delta_t$.

A.4 Safe-LUCB with ℓ_1-confidence Region

In this section we briefly discussed a modified ℓ_1-confidence region (as in [40]), which is used in our numerical experiments.

Motivation. The minimization in (2.9) involves solving a bilinear optimization problem. In view of (2.6) and (2.8) it is not hard to show that (2.9) can be equivalently expressed as follows:

$$\tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}_t = \min_x \ \hat{\boldsymbol{\mu}}_t^\top x - \beta_t \|\mathbf{x}\|_{\mathbf{A}_t^{-1}} \quad \text{sub. to} \quad \hat{\boldsymbol{\mu}}_t^\top \mathbf{B}\mathbf{x} + \beta_t \|\mathbf{B}\mathbf{x}\|_{\mathbf{A}_t^{-1}} \leq c, \quad x \in \mathcal{D}_0 \ .$$

This is a non-convex optimization problem. Thus, we present a variant of Safe-LUCB (and its analysis) and we show that it can be efficiently implemented (particularly so, when the decision set is a polytope) [40]. We use this variant in our simulation results (see Appendix A.6).

Algorithm and guarantees. We adapt the procedure first presented in [40] to our new Safe-LUCB algorithm. The pure-exploration phase of the algorithm remains unaltered. In the safe exploration-exploitation phase, the only thing that changes is the definition of the confidence region in Line 8 in Algorithm 1. Specifically, we define the modified ℓ_1-confidence region as follows:

$$\mathcal{C}_t^{\ell_1} := \{\boldsymbol{\nu} \in \mathbb{R}^d : \|\boldsymbol{\nu} - \hat{\boldsymbol{\mu}}_t\|_{\mathbf{A}_{t,1}} \leq \beta_t \sqrt{d}\}. \tag{A.25}$$

Note that for any $v \in \mathcal{C}_t$ and all $t > 0$, $\|\mathbf{A}_t^{1/2}(\boldsymbol{\nu} - \hat{\boldsymbol{\mu}}_t)\|_1 \leq \sqrt{d}\|\mathbf{A}_t^{1/2}(\boldsymbol{\nu} - \hat{\boldsymbol{\mu}}_t)\|_2 \leq \sqrt{d}\beta_t$. Thus, $\mathcal{C}_t \subseteq \mathcal{C}_t^{\ell_1}, \quad \forall t > 0$. From this and Theorem 1, we conclude $\Pr(\boldsymbol{\mu} \in \mathcal{C}_t^{\ell_1}, \forall t > 0) \geq 1 - \delta$. Then, the natural modification of (2.9) becomes

$$\tilde{\boldsymbol{\mu}}_t^\top \mathbf{x}_t = \min_{\mathbf{x} \in \mathcal{D}_t^s, \boldsymbol{\nu} \in \mathcal{C}_t^{\ell_1}} \boldsymbol{\nu}^\top \mathbf{x} = \min_{\boldsymbol{\nu} \in \mathcal{C}_t^{\ell_1}} \quad f(\boldsymbol{\nu}), \tag{A.26}$$

where

$$f(\boldsymbol{\nu}) := \min_{\substack{\mathbf{x} \in \mathcal{D}_0 \\ \hat{\mathbf{m}}_t^\top \mathbf{B}\mathbf{x} + \sqrt{d}\beta_t \|\mathbf{B}\mathbf{x}\|_{\mathbf{A}_t^{-1}} \leq C}} \nu^\top \mathbf{x}. \tag{A.27}$$

From these, it is clear that all the results and theorems can be directly applied to the modified algorithm which uses ℓ_1-confidence region in (A.25), with $\beta_t\sqrt{d}$ instead of β_t. As noted in [40] the regret of the modified algorithm does not optimally scale with the dimension d (since there is an extra factor of $\sqrt{d}$ introduced by the substitution $\beta_t \leftarrow \beta_t\sqrt{d}$). However, as explained next, solving (A.26) is now computationally tractable.

On computational efficiency. Note that the minimization in (A.27) is a convex program that can be efficiently solved for fixed $\boldsymbol{\nu}$. In particular, if $\mathcal{D}_0$ is a polytope then the minimization in (A.27) is a quadratic program. Moreover, note that $f(\boldsymbol{\nu})$ is positive homogeneous of degree one, i.e., $f(\theta v) = \theta f(\boldsymbol{\nu})$ for any $\theta \geq 0$. Therefore, in order to solve (A.26) it suffices to evaluate the function $f(\boldsymbol{\nu})$ at the $2d$ vertices $v_1, \ldots, v_{2d}$ of $\mathcal{C}_t^{\ell_1}$ in (A.25) and choose the minimum $f_{\min} := \min_{\boldsymbol{\nu}_i,\ i\in[2d]} f(\boldsymbol{\nu}_i)$. In order to see this, let $\boldsymbol{\nu}^* \in \arg\min_{\boldsymbol{\nu}\in\mathcal{C}_t^{\ell_1}} f(\boldsymbol{\nu})$ and $\theta_1, \ldots, \theta_{2d} \geq 0, \sum_{i=1}^{d} \theta_i = 1$ such that $\boldsymbol{\nu}^* = \sum_{i=1}^{2d} \theta_i \boldsymbol{\nu}_i$. Then,

$$\min_{v\in\mathcal{C}_t^{\ell_1}} f(\boldsymbol{\nu}) = f(\boldsymbol{\nu}^*) = \sum_{i=1}^{2d} \theta_i f(\boldsymbol{\nu}_i) \geq f_{\min} \sum_{i=1}^{2d} \theta_i = f_{\min} \geq \min_{v\in\mathcal{C}_t^{\ell_1}} f(\boldsymbol{\nu}).$$

Thus,

$$\min_{v\in\mathcal{C}_t^{\ell_1}} f(\boldsymbol{\nu}) = \min_{v_i,\ i\in[2d]} f(\boldsymbol{\nu}_i). \tag{A.28}$$

To sum up, we see from (A.28) that solving (A.26) amounts to solving $2d$ quadratic programs (when $\mathcal{D}_0$ is a polytope).

A.5 On GSLUCB

Having no knowledge of the safety gap Δ, GSLUCB starts conservatively by setting the length of the pure exploration phase to its largest possible value, which is equal to T_0 defined in Theorem 3 (corresponding to $\Delta = 0$). The idea behind GSLUB is to generate at each round t of the pure-exploration phase a certain value Δ_t that serves as a lower bound for the unknown safety gap Δ. We discuss possible ways to do so next, but for now let us describe how these lower estimates of Δ can be useful. Owing to the result of Theorem 2, at each

round t, GSLUCB computes a pure exploration duration $T'_t = T_{\Delta_t}$, which is associated with the lower confidence bound Δ_t (Eqn. (2.13) for $\Delta = \Delta_t$). If at some round t, the computed T'_t becomes less than t, then Theorem 2 guarantees that $x^* \in \mathcal{D}^{\mathrm{s}}_t$ and the algorithm switches to the exploration-exploitation phase.

One way to compute the Δ_t's that guarantees $\Delta_t \leq \Delta$ is as follows. For each vector $\boldsymbol{\nu} \in \mathcal{C}_t$ denote $\mathbf{x}^*_{\boldsymbol{\nu}} \in \arg\min_{\mathbf{x} \in \mathcal{D}^{\mathrm{s}}(\boldsymbol{\nu})} v^\top x$, where $\mathcal{D}^{\mathrm{s}}(\boldsymbol{\nu}) := \{x \in \mathcal{D}_0 : \boldsymbol{\nu}^\top \mathbf{B}\mathbf{x} \leq c\}$ and define

$$\Delta_t := \min_{\boldsymbol{\nu} \in \mathcal{C}_t} \Delta_{\boldsymbol{\nu}}, \tag{A.29}$$

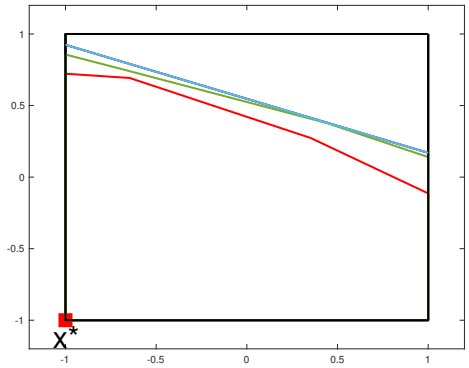
where $\Delta_{\boldsymbol{\nu}} := c - \boldsymbol{\nu}^\top \mathbf{B}\mathbf{x}^*_{\boldsymbol{\nu}}$. Since $\mu \in \mathcal{C}_t$ with high probability (cf. Theorem 1) and by definition of Δ, it can be seen that $\Delta_t \leq \Delta$. Unfortunately, solving (A.29) can be challenging and, in general, one has to resort to relaxed versions of the optimization involved, but ones that guarantee $\Delta_t \leq \Delta$ (at least after a few rounds). We leave the study of this general case to future work and we discuss here a special case in which this is possible. We have implemented this special case in the simulation results presented in Figure 2.1a (see Appendix A.6). Specifically, we consider a finite K-armed linear bandit setting with feature vectors denoted by $\mathbf{y}_1, \ldots, \mathbf{y}_K$. We produce lower estimates Δ_t as follows. For all $i \in [K]$, we form the following two sets. (i) The set $\mathcal{C}^i_t = \{\boldsymbol{\nu} \in \mathcal{C}_t \mid \boldsymbol{\nu}^\top B y_i \leq c\}$ of all vectors in the confidence region for which the action y_i is deemed safe; (ii) The set $\mathcal{Y}^i_t = \{\mathbf{y}_j, j \in [K] \mid \max_{v \in \mathcal{C}^i_t} \boldsymbol{\nu}^\top \mathbf{B}\mathbf{y}_j \leq c\}$ of all actions that are considered safe with respect to all $\boldsymbol{\nu} \in \mathcal{C}^i_t$. Then, we define

$$\Delta^i_t := \min_{\substack{\boldsymbol{\nu} \in \mathcal{C}^i_t \\ \boldsymbol{\nu}^\top y_i \leq \boldsymbol{\nu}^\top \mathbf{y}, \text{ for all } \mathbf{y} \in \mathcal{Y}^i_t}} c - \boldsymbol{\nu}^\top \mathbf{B}\mathbf{y}_i. \tag{A.30}$$

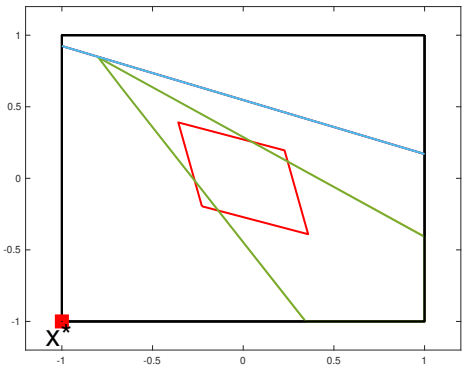
It can be checked that $\min_{i \in [K]} \Delta^i_t \leq \Delta$. Thus we rely on $\min_{i \in [K]} \Delta^i_t$ as our lower confidence bound on Δ. Note that computing $\min_{i \in [K]} \Delta^i_t$ is computationally tractable for finite K and an ℓ_1 confidence region.

A.6 Experiments

In this section, we provide the details of our numerical experiments. In view of our discussion in Appendix A.4, we implement a modified version of Safe-LUCB which uses 1-norms instead



(a) Safe-LUCB with pure exploration phase.



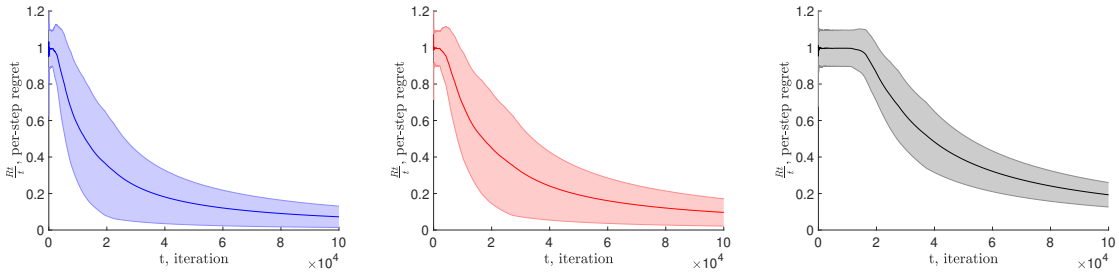
(b) Safe-LUCB without pure exploration phase

Figure A.1: Growth of $\mathcal{D}_t^{\mathrm{s}}$ with and without pure exploration phase. In both figures: $\mathcal{D}_0$ (in black) $\mathcal{D}^{\mathrm{s}}$ (in blue), $\mathcal{D}_{T'+1}^S$ (in red), $\mathcal{D}_{5e4}^S$ (in green). Also, shown the optimal action $\mathbf{x}^*$. Note that $\mathbf{x}^* \in \mathcal{D}_{T'+1}^S$ when pure exploration phase is used as suggested by Lemma 2.

of 2-norms (as in [40]; see also Appendix A.4 for details). We have taken $\delta = 0.01$, $\lambda = 1$, and $R = 0.1$ in all cases.

Figure 2.1a compares the average per-step regret of 1) Safe-LUCB with knowledge of Δ; 2) Safe-LUCB without knowledge of Δ (hence, assuming $\Delta = 0$); 3) GSLUCB without knowledge of Δ (the algorithm creates a lower confidence bound for Δ as the pure exploration phase runs). Figure A.2 highlights the sample standard deviation of regret around the average per-step regret for each of the above-mentioned cases. We considered a time independent decision set of 15 arms in $\mathbb{R}^4$ such that 5 of the feature vectors are drawn uniformly from $\mathcal{D}^w$ and the other 10 are drawn uniformly from unit ball in $\mathbb{R}^4$. Moreover, μ is drawn from $\mathcal{N}(0, I_4)$ and then normalized to unit norm. B and c are drawn uniformly from $[0, 0.5]^{4\times 4}$ and [0,1] respectively. The results shown depict averages over 20 realizations. It can be seen from the figure that GSLUCB performs significantly better than the worst case suggested by Theorem 3 (aka Safe-LUCB assuming $\Delta = 0$). In fact, it appears that it approaches the improved regret performance suggested by Theorem 2 of Safe-LUCB with knowledge of Δ.

Our second numerical experiment serves to showcase the value of the safe exploration phase



(a) Safe-LUCB, $T' = T_\Delta$ **(b)** GSLUCB **(c)** Safe-LUCB, $T' = T_0$

Figure A.2: Comparison of mean per-step regret for Safe-LUCB($T' = T_\Delta$), GSLUCB, and Safe-LUCB($T' = T_0$). The shaded regions show one standard deviation around the mean. The results are averages over 20 problem realizations.

as discussed in Section 2.5.3. We focus on an instance with positive safety gap $\Delta > 0$ to verify the validity of Lemma 2, namely that $\mathbf{x}^* \in \mathcal{D}_t^{\text{s}}$ for $t \geq T' + 1$, when T' is appropriately chosen. Furthermore, we compare the performance with a "naive" variation of Safe-LUCB that only implements the safe exploration-exploitation phase (aka, no pure exploration phase). The regret plots of the two algorithms (with and without pure exploration phase) shown in Figure 2.1b clearly demonstrate the value of the pure exploration phase for the simulated example. Specifically, for the simulation, we consider a horizon $T = 100000$ with decision set $\mathcal{D}_0$ the unit ℓ_∞-ball in $\mathbb{R}^2$, and, the following parameters: $\boldsymbol{\mu} = \begin{bmatrix} 0.9 \\ 0.044 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 0.6 & 1.8 \\ 1.8 & 0.4 \end{bmatrix}$, $c = 0.9$. We have chosen a low-dimensional instance, because we find it instructive to also depict the the growth of the safe sets for the two algorithms. This is done in Figures A.1a and A.1b, where we illustrate the safe sets of Safe-LUCB with and without pure exploration phase, respectively. Black lines denote the (border of) the polytope $\mathcal{D}_0$; blue lines denote the linear constraint in (2.1); red lines denote the (border of) $\mathcal{D}_{T'+1}^{\text{s}}$, where $T' = T_\Delta = 1054$ and $T' = 0$ for Figures A.1a and A.2c, respectively; and, green lines denote (the border of) safe sets $\mathcal{D}_{50000}^{\text{s}}$ at round 50000. Also depicted the optimal action $\mathbf{x}^*$ with coordinates $\{-1, -1\}$. As expected, Safe-LUCB starts the exploration-exploitation phase with a safe set that includes $\mathbf{x}^*$ while, without the pure exploration phase, the algorithm starts the exploration-exploitation phase with a smaller safe set which does not include $\mathbf{x}^*$ and as a

results, fails in expanding the safe set to include $\mathbf{x}^*$ even after $T = 50000$ rounds. This results in the bad regret performance in Figure 2.1b.

APPENDIX B

Proofs for Chapter 3

B.1 SLUCB-QVI Proofs

In this section, we prove the technical statements in Sections 3.4 and 3.5. First, recall the definitions of the following events that we repeatedly refer to throughout this section:

$$\mathcal{E}_1 := \left\{ \Phi_0^{\perp}\left(s, \boldsymbol{\gamma}_h^*\right) \in \mathcal{C}_h^k(s),\ \forall (s,h,k) \in \mathcal{S} \times [H] \times [K] \right\}, \tag{B.1}$$

$$\begin{aligned} \mathcal{E}_2 := \Bigg\{ & \left| \langle \mathbf{w}_h^k, \boldsymbol{\phi}(s,a) \rangle - Q_h^{\pi}(s,a) + [\mathbb{P}_h V_{h+1}^{\pi} - V_{h+1}^k](s,a) \right| \leq \beta \left\| \boldsymbol{\phi}(s,a) \right\|_{\left(\mathbf{A}_h^k\right)^{-1}} \\ & , \forall (a,s,h,k) \in \mathcal{A} \times \mathcal{S} \times [H] \times [K] \Bigg\}. \end{aligned} \tag{B.2}$$

B.1.1 Proof of Proposition 1

Let $a \in \mathcal{A}_h^k(s)$. Recall that $\Phi_0(s, \mathbf{x}) = \left\langle \mathbf{x}, \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right) \right\rangle \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)$ for any $\mathbf{x} \in \mathbb{R}^d$. By the definition of $\mathcal{A}_h^k(s)$ in (3.9), we have

$$\frac{\left\langle \Phi_0\left(s, \boldsymbol{\phi}(s,a)\right), \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right) \right\rangle}{\left\| \boldsymbol{\phi}\left(s, a_0\left(s\right)\right) \right\|_2} \tau_h(s) + \left\langle \boldsymbol{\gamma}_{h,s}^k, \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s,a)\right) \right\rangle + \beta \left\| \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s,a)\right) \right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}} \leq \tau \tag{B.3}$$

Moreover, using Cauchy-Schwarz inequality and conditioned on event $\mathcal{E}_1$ in (B.1), we get

$$\left| \left\langle \boldsymbol{\gamma}_{h,s}^k - \Phi_0^{\perp}\left(s, \boldsymbol{\gamma}_h^*\right), \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s,a)\right) \right\rangle \right| \leq \beta \left\| \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s,a)\right) \right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}, \tag{B.4}$$

and thus,

$$\left\langle \Phi_0^{\perp}\left(s, \boldsymbol{\gamma}_h^*\right), \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s,a)\right)\right\rangle \leq \left\langle \boldsymbol{\gamma}_{h,s}^k, \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s,a)\right)\right\rangle + \beta\left\|\Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s,a)\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}. \quad \text{(B.5)}$$

Note that $\left\langle \Phi_0^{\perp}\left(s, \boldsymbol{\gamma}_h^*\right), \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s,a)\right)\right\rangle = \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\phi}(s,a)\right\rangle - \left\langle \boldsymbol{\gamma}_{h,s}^*, \Phi_0\left(s, \boldsymbol{\phi}(s,a)\right)\right\rangle = \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\phi}(s,a)\right\rangle - \frac{\left\langle \Phi_0\left(s,\boldsymbol{\phi}(s,a)\right), \tilde{\boldsymbol{\phi}}\left(s,a_0(s)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s,a_0(s)\right)\right\|_2}\tau_h(s)$. Combining this fact with (B.3) and (B.5) concludes that

$$\begin{aligned}
\left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\phi}(s,a)\right\rangle &= \frac{\left\langle \Phi_0\left(s, \boldsymbol{\phi}(s,a)\right), \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)\right\|_2}\tau_h(s) + \left\langle \Phi_0^{\perp}\left(s, \boldsymbol{\gamma}_h^*\right), \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s,a)\right)\right\rangle \\
&\leq \frac{\left\langle \Phi_0\left(s, \boldsymbol{\phi}(s,a)\right), \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)\right\|_2}\tau_h(s) + \left\langle \boldsymbol{\gamma}_{h,s}^k, \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s,a)\right)\right\rangle + \beta\left\|\Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s,a)\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}} \\
&\qquad \text{(Eqn. (B.5))} \\
&\leq \tau, \qquad \text{(Eqn. (B.3))}
\end{aligned}$$

as desired.

B.1.2 Proof of Lemma 3

Before we start the main proof, we introduce vectors $\{\mathbf{w}_h^{\pi}\}_{h\in[H]}$ for any policy π:

$$\mathbf{w}_h^{\pi} := \boldsymbol{\theta}_h^* + \int_{\mathcal{S}} V_{h+1}^{\pi}(s')d\boldsymbol{\mu}(s'). \quad \text{(B.6)}$$

From the Bellman equation in (3.4) and the linearity of the MDP in Assumption 4, we have:

$$Q_h^{\pi}(s,a) := \left\langle \boldsymbol{\phi}(s,a), \mathbf{w}_h^{\pi}\right\rangle. \quad \text{(B.7)}$$

See Proposition 2.3 in [64] for the proof.

Now, we prove Lemma 3 by induction. First, we prove the base case at time-step $H+1$. The statement holds because $V_{H+1}^*(s) = V_{H+1}^k(s) = 0$. Now, suppose the statement holds for time-step $h+1$. We prove it also holds for time-step h. For all $(s,h,k) \in \mathcal{S}\times[H]\times[K]$, let

$$a_h^k(s) := \underset{a\in\mathcal{A}_h^k(s)}{\arg\max}\, Q_h^k(s,a) \quad \text{and} \quad a_h^*(s) := \underset{a\in\mathcal{A}_h^{\text{safe}}(s)}{\arg\max}\, Q_h^*(s,a). \quad \text{(B.8)}$$

We consider the following two cases:

1) If $a_h^*(s) \in \mathcal{A}_h^k(s)$, we have

$$\begin{aligned} V_h^k(s) = \max_{a \in \mathcal{A}_h^k(s)} Q_h^k(s,a) &\geq Q_h^k(s, a_h^*(s)) \\ &\geq Q_h^*(s, a_h^*(s)) + \mathbb{E}_{s' \sim \mathbb{P}(.|s, a_h^*(s))} \left[V_{h+1}^k(s') - V_{h+1}^*(s') \right] \\ &\qquad \text{(Conditioned on } \mathcal{E}_2 \text{ in (B.2))} \\ &\geq Q_h^*(s, a_h^*(s)) = V_h^*(s), \qquad \text{(Induction assumption)} \end{aligned}$$

as desired.

2) Now, we recall the definition of $\mathcal{A}_h^k(s)$ in (3.9) and focus on the other case when $a_h^*(s) \notin \mathcal{A}_h^k(s)$, which means

$$\frac{\left\langle \Phi_0\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right), \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right) \right\rangle}{\left\| \boldsymbol{\phi}\left(s, a_0\left(s\right)\right) \right\|_2} \tau_h(s) + \left\langle \boldsymbol{\gamma}_{h,s}^k, \Phi_0^\perp\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right) \right\rangle + \beta \left\| \Phi_0^\perp\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right) \right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}} > \tau. \tag{B.9}$$

Now, we observe that $a_0(s) \in \mathcal{A}_h^k(s)$. Recall that $\tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right) = \frac{\boldsymbol{\phi}\left(s, a_0(s)\right)}{\left\| \boldsymbol{\phi}\left(s, a_0(s)\right) \right\|_2}$ and note that $\Phi_0\left(s, \boldsymbol{\phi}\left(s, a_0\left(s\right)\right)\right) = \boldsymbol{\phi}\left(s, a_0\left(s\right)\right)$ and $\Phi_0^\perp\left(s, \boldsymbol{\phi}(s, a_0(s))\right) = \mathbf{0}$. Thus

$$\begin{aligned} &\frac{\left\langle \boldsymbol{\phi}\left(s, a_0\left(s\right)\right), \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right) \right\rangle}{\left\| \boldsymbol{\phi}\left(s, a_0\left(s\right)\right) \right\|_2} \tau_h(s) + \left\langle \boldsymbol{\gamma}_{h,s}^k, \Phi_0^\perp\left(s, \boldsymbol{\phi}(s, a_0(s))\right) \right\rangle + \beta \left\| \Phi_0^\perp\left(s, \boldsymbol{\phi}(s, a_0(s))\right) \right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}} = \tau_h(s) \\ &\qquad < \tau, \end{aligned} \tag{B.10}$$

which implies that $a_0(s) \in \mathcal{A}_h^k(s)$ or equivalently $\boldsymbol{\phi}\left(s, a_0\left(s\right)\right) \in \mathcal{D}_h^k(s) := \left\{ \boldsymbol{\phi}(s,a) : a \in \mathcal{A}_h^k(s) \right\}$. Now, let

$$\alpha_h^k(s) := \max \left\{ \alpha \in [0,1] : \alpha \boldsymbol{\phi}(s, a_h^*(s)) + (1-\alpha) \boldsymbol{\phi}\left(s, a_0\left(s\right)\right) \in \mathcal{D}_h^k(s) \right\}. \tag{B.11}$$

Assumption 5 guarantees that $\alpha_h^k(s)$ exists for all $(s,k) \in \mathcal{S} \times [H] \times [K]$. Note that $\Phi_0\left(s, \alpha \boldsymbol{\phi}(s, a_h^*(s)) + (1-\alpha) \boldsymbol{\phi}\left(s, a_0\left(s\right)\right)\right) = \alpha \Phi_0\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right) + (1-\alpha) \boldsymbol{\phi}\left(s, a_0\left(s\right)\right)$ and $\Phi_0^\perp\left(s, \alpha \boldsymbol{\phi}(s, a_h^*(s)) + (1-\alpha) \boldsymbol{\phi}\left(s, a_0\left(s\right)\right)\right) = \alpha \Phi_0^\perp\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)$. Thus, by the definition of

$\mathcal{D}_h^k(s)$, we have

$$\alpha_h^k(s) := \max\Bigg\{\alpha \in [0,1] : \frac{\Big\langle \alpha\Phi_0\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right) + (1-\alpha)\boldsymbol{\phi}\left(s, a_0\left(s\right)\right), \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)\Big\rangle}{\left\|\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)\right\|_2}\tau_h(s)$$
$$+ \alpha\left\langle \boldsymbol{\gamma}_{h,s}^k, \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\rangle + \alpha\beta\left\|\Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}} \leq \tau\Bigg\}. \tag{B.12}$$

For all $(s,k) \in \mathcal{S} \times [K]$, at time-step h, let $\mathbf{y}_h^k(s) := \alpha_h^k(s)\boldsymbol{\phi}(s, a_h^*(s)) + (1 - \alpha_h^k(s))\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)$. Thus, the definition of $\alpha_h^k(s)$ in (B.11) implies that $\mathbf{y}_h^k(s) \in \mathcal{D}_h^k(s)$, and thus

$$\begin{aligned}\max_{a\in\mathcal{A}_h^k(s)} Q_h^k(s,a) &\geq \min\left\{\left\langle \mathbf{w}_h^k, \mathbf{y}_h^k(s)\right\rangle + \kappa_h(s)\beta\left\|\mathbf{y}_h^k(s)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}, H\right\}\\ &= \min\left\{\left\langle \mathbf{w}_h^k - \mathbf{w}_h^*, \mathbf{y}_h^k(s)\right\rangle + \left\langle \mathbf{w}_h^*, \mathbf{y}_h^k(s)\right\rangle + \kappa_h(s)\beta\left\|\mathbf{y}_h^k(s)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}, H\right\}.\end{aligned} \tag{B.13}$$

Conditioned on event $\mathcal{E}_2$ in (B.2), and by the induction assumption, we have

$$\begin{aligned}-\beta\left\|\mathbf{y}_h^k(s)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}} &\leq \left\langle \mathbf{w}_h^k - \mathbf{w}_h^*, \mathbf{y}_h^k(s)\right\rangle + \mathbb{E}_{s'\sim\mathbb{P}(.|s,a_h^*(s))}\left[V_{h+1}^*(s') - V_{h+1}^k(s')\right]\\ &\leq \left\langle \mathbf{w}_h^k - \mathbf{w}_h^*, \mathbf{y}_h^k(s)\right\rangle.\end{aligned} \tag{B.14}$$

By combining (B.13) and (B.14), we conclude that

$$\begin{aligned}\max_{a\in\mathcal{A}_h^k(s)} Q_h^k(s,a) &\geq \min\left\{\left\langle \mathbf{w}_h^*, \mathbf{y}_h^k(s)\right\rangle + (\kappa_h(s)-1)\beta\left\|\mathbf{y}_h^k(s)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}, H\right\}\\ &\geq \min\Big\{\alpha_h^k(s)\left\langle \mathbf{w}_h^*, \boldsymbol{\phi}(s, a_h^*(s))\right\rangle\\ &\quad +(1-\alpha_h^k(s))\left\langle \mathbf{w}_h^*, \boldsymbol{\phi}\left(s, a_0\left(s\right)\right)\right\rangle + (\kappa_h(s)-1)\beta\left\|\Phi_0^{\perp}\left(s, \mathbf{y}_h^k(s)\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}, H\Big\}\\ &\geq \min\left\{\alpha_h^k(s)\left(\left\langle \mathbf{w}_h^*, \boldsymbol{\phi}(s, a_h^*(s))\right\rangle + (\kappa_h(s)-1)\beta\left\|\Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}\right), H\right\},\end{aligned} \tag{B.15}$$

where the second inequality holds because $\left\|\mathbf{y}_h^k(s)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}} \geq \left\|\Phi_0^{\perp}\left(s, \mathbf{y}_h^k(s)\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}$ (see Lemma 3 in [96] for a proof). The last inequality follows from the fact that $(1 -$

$\alpha_h^k(s))\left\langle \mathbf{w}_h^*, \boldsymbol{\phi}\left(s, a_0\left(s\right)\right)\right\rangle \geq 0$ as the reward is always positive, i.e., $r_h(s,a) \in [0,1]$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

Now, we show that $\alpha_h^k(s) \geq \frac{\tau - \tau_h(s)}{\tau - \tau_h(s) + 2\beta \left\| \Phi_0^\perp\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}}$, which eventually leads to a proper value for $\kappa_h(s) > 1$ that guarantees for all $(s,h,k) \in \mathcal{S} \times [H] \times [K]$ it holds that $V_h^*(s) \leq V_h^k(s)$ conditioned on $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$. Definitions of $\alpha_h^k(s)$ in (B.12) and the estimated safe set $\mathcal{A}_h^k(s)$ in (3.9) imply that for all $(s,h,k) \in \mathcal{S} \times [H] \times [K]$, we have

$$\frac{(1-\alpha_h^k(s))\left\langle \boldsymbol{\phi}\left(s, a_0\left(s\right)\right), \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)\right\|_2}\tau_h(s) + \alpha_h^k(s)\left[\frac{\left\langle \Phi_0\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right), \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)\right\|_2}\tau_h(s)\right.$$
$$\left. + \left\langle \boldsymbol{\gamma}_{h,s}^k, \Phi_0^\perp\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\rangle + \beta\left\|\Phi_0^\perp\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}\right] = \tau. \tag{B.16}$$

Let $M = \frac{\left\langle \Phi_0\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right), \tilde{\boldsymbol{\phi}}\left(s, a_0(s)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s, a_0(s)\right)\right\|_2}\tau_h(s) + \left\langle \boldsymbol{\gamma}_{h,s}^k, \Phi_0^\perp\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\rangle + \beta\left\|\Phi_0^\perp\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}$. Note that due to (B.9), $M > \tau$, and recall that $\tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right) = \frac{\boldsymbol{\phi}\left(s, a_0(s)\right)}{\left\|\boldsymbol{\phi}\left(s, a_0(s)\right)\right\|_2}$. Therefore, (B.16) gives that

$$0 < \alpha_h^k(s) = \frac{\tau - \tau_h(s)}{M - \tau_h(s)} < 1. \tag{B.17}$$

In order to lower bound $\alpha_h^k(s)$ (upper bound M), we first rewrite M as

$$M = \frac{\left\langle \Phi_0\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right), \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)\right\|_2}\tau_h(s) + \left\langle \boldsymbol{\gamma}_h^*, \Phi_0^\perp\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\rangle$$
$$+ \left\langle \boldsymbol{\gamma}_{h,s}^k - \boldsymbol{\gamma}_h^*, \Phi_0^\perp\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\rangle + \beta\left\|\Phi_0^\perp\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}, \tag{B.18}$$

and show that

(a) $\frac{\left\langle \Phi_0\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right), \tilde{\boldsymbol{\phi}}\left(s, a_0(s)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s, a_0(s)\right)\right\|_2}\tau_h(s) + \left\langle \boldsymbol{\gamma}_h^*, \Phi_0^\perp\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\rangle \leq \tau$ because

$$
\begin{aligned}
&\frac{\left\langle \Phi_0\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right), \tilde{\boldsymbol{\phi}}\left(s, a_0(s)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s, a_0(s)\right)\right\|_2}\tau_h(s) + \left\langle \boldsymbol{\gamma}_h^*, \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\rangle \\
&= \left\langle \boldsymbol{\gamma}_h^*, \left\langle \Phi_0\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right), \tilde{\boldsymbol{\phi}}\left(s, a_0(s)\right)\right\rangle \tilde{\boldsymbol{\phi}}\left(s, a_0(s)\right)\right\rangle \\
&+ \left\langle \boldsymbol{\gamma}_h^*, \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\rangle \\
&= \left\langle \boldsymbol{\gamma}_h^*, \Phi_0\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\rangle + \left\langle \boldsymbol{\gamma}_h^*, \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\rangle \\
&= \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\phi}(s, a_h^*(s))\right\rangle \\
&\leq \tau. \qquad \text{(B.19)}
\end{aligned}
$$

(b) $\left\langle \boldsymbol{\gamma}_{h,s}^k - \boldsymbol{\gamma}_h^*, \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\rangle \leq \beta\left\|\Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}$, because conditioned on $\mathcal{E}_1$ in (B.1), we have

$$
\begin{aligned}
\left\langle \boldsymbol{\gamma}_{h,s}^k - \boldsymbol{\gamma}_h^*, \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\rangle &= \left\langle \boldsymbol{\gamma}_{h,s}^k - \Phi_0^{\perp}\left(s, \boldsymbol{\gamma}_h^*\right), \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\rangle \\
&\leq \beta\left\|\Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}. \qquad \text{(B.20)}
\end{aligned}
$$

Now, we combine (B.18), (B.19) and (B.20) to conclude that

$$
M \leq \tau + 2\beta\left\|\Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}} \quad \Rightarrow \quad \alpha_h^k(s) \geq \frac{\tau - \tau_h(s)}{\tau - \tau_h(s) + 2\beta\left\|\Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}}. \qquad \text{(B.21)}
$$

This lower bound on $\alpha_h^k(s)$ combined with (B.15) gives

$$
\max_{a \in \mathcal{A}_h^k(s)} Q_h^k(s,a) \geq \min\left\{ \frac{(\tau - \tau_h(s))\left(\left\langle \mathbf{w}_h^*, \boldsymbol{\phi}(s, a_h^*(s))\right\rangle + (\kappa_h(s) - 1)\beta\left\|\Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}\right)}{\tau - \tau_h(s) + 2\beta\left\|\Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}}, H \right\} \qquad \text{(B.22)}
$$

Let $M_1 = \beta\left\|\Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_h^*(s))\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}$. We observe that Therefore $\max_{a \in \mathcal{A}_h^{\text{safe}}} Q_h^*(s,a) =$

$\max_{a\in\mathcal{A}_h^{\text{safe}}} \min\left\{Q_h^*(s,a), H\right\} = \min\left\{\max_{a\in\mathcal{A}_h^{\text{safe}}} Q_h^*(s,a), H\right\}$. Therefore

$$\begin{aligned}
\max_{a\in\mathcal{A}_h^k(s)} Q_h^k(s,a) \geq \max_{a\in\mathcal{A}_h^{\text{safe}}} Q_h^*(s,a) &\iff \big(\tau-\tau_h(s)\big)\left(\max_{a\in\mathcal{A}_h^{\text{safe}}} Q_h^*(s,a) + (\kappa_h(s)-1)M_1\right)\\
&\geq \big(\tau-\tau_h(s)+2M_1\big)\max_{a\in\mathcal{A}_h^{\text{safe}}} Q_h^*(s,a)\\
&\iff \big(\tau-\tau_h(s)\big)\left(\kappa_h(s)-1\right) \geq 2\max_{a\in\mathcal{A}_h^{\text{safe}}} Q_h^*(s,a)\\
&\iff \big(\tau-\tau_h(s)\big)\left(\kappa_h(s)-1\right) \geq 2H\\
&\iff \kappa_h(s) \geq \frac{2H}{\tau-\tau_h(s)} + 1, \qquad \text{(B.23)}
\end{aligned}$$

as desired.

B.1.3 Proof of Theorem 4

The key property of optimism in the face of safety constraint in SLUCB-QVI, which is proved in Appendix B.1.2 as our main technical allows us to follow the standard steps in establishing the regret bound of unsafe LSVI-UCB in [64] to complete the proof of Theorem 4.

Conditioned on event $\mathcal{E}_2$ in (B.2), for any $(a,s,h,k)\in\mathcal{A}\times\mathcal{S}\times[H]\times[K]$, we have

$$\begin{aligned}
Q_h^k(s,a) - Q_h^{\pi_k}(s,a) &= \min\left\{\left\langle \mathbf{w}_h^k, \boldsymbol{\phi}(s,a)\right\rangle + \kappa_h(s)\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}, H\right\} - Q_h^{\pi_k}(s,a)\\
&\leq \left\langle \mathbf{w}_h^k, \boldsymbol{\phi}(s,a)\right\rangle + \kappa_h(s)\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}} - Q_h^{\pi_k}(s,a)\\
&\leq \mathbb{E}_{s'\sim\mathbb{P}(.|s,a)}\left[V_{h+1}^k(s') - V_{h+1}^{\pi_k}(s')\right] + \big(1+\kappa_h(s)\big)\,\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}. \qquad \text{(B.24)}
\end{aligned}$$

Let $\delta_h^k := V_h^k(s_h^k) - V_h^{\pi_k}(s_h^k)$ and $\zeta_{h+1}^k := \mathbb{E}_{s'\sim\mathbb{P}(.|s_h^k,a_h^k)}\left[V_{h+1}^k(s') - V_{h+1}^{\pi_k}(s')\right] - \delta_{h+1}^k$. We can write

$$\begin{aligned}
\delta_h^k &= V_h^k(s_h^k) - V_h^{\pi_k}(s_h^k)\\
&= Q_h^k(s_h^k,a_h^k) - Q_h^{\pi_k}(s_h^k,a_h^k)\\
&\leq \mathbb{E}_{s'\sim\mathbb{P}(.|s_h^k,a_h^k)}\left[V_{h+1}^k(s') - V_{h+1}^{\pi_k}(s')\right] + \big(1+\kappa_h(s)\big)\,\beta\left\|\boldsymbol{\phi}_h^k\right\|_{\left(\mathbf{A}_h^k\right)^{-1}} \qquad \text{(Eqn. (B.24))}\\
&= \delta_{h+1}^k + \zeta_{h+1}^k + \big(1+\kappa_h(s)\big)\,\beta\left\|\boldsymbol{\phi}_h^k\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}. \qquad \text{(B.25)}
\end{aligned}$$

Now, conditioning on event $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$, we bound the cumulative regret as follows:

$$\begin{aligned} R_K &= \sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \leq \sum_{k=1}^{K} \delta_1^k && \text{(Lemma 3)} \\ &\leq \sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_h^k + \sum_{k=1}^{K}\sum_{h=1}^{H} \left(1+\kappa_h(s)\right)\beta \left\|\boldsymbol{\phi}_h^k\right\|_{\left(\mathbf{A}_h^k\right)^{-1}} \\ &\leq \sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_h^k + (1+\kappa)\,\beta \sum_{k=1}^{K}\sum_{h=1}^{H} \left\|\boldsymbol{\phi}_h^k\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}. && \text{(B.26)} \end{aligned}$$

We observe that $\{\zeta_h^k\}$ is a martingale difference sequence satisfying $|\zeta_h^k| \leq 2H$. Thus, thanks to Azuma-Hoeffding inequality, we have

$$\mathbb{P}\left(\sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_h^k \leq 2H\sqrt{T\log(dT/\delta)}\right) \geq 1-\delta. \tag{B.27}$$

In order to bound $\sum_{k=1}^{K}\sum_{h=1}^{H}\left\|\boldsymbol{\phi}_h^k\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}$, note that for any $h \in [H]$, we have

$$\begin{aligned} \sum_{k=1}^{K}\left\|\boldsymbol{\phi}_h^k\right\|_{\left(\mathbf{A}_h^k\right)^{-1}} &\leq \sqrt{K\sum_{k=1}^{K}\left\|\boldsymbol{\phi}_h^k\right\|^2_{\left(\mathbf{A}_h^k\right)^{-1}}} && \text{(Cauchy-Schwartz inequality)} \\ &\leq \sqrt{2K\log\left(\frac{\det\left(\mathbf{A}_h^K\right)}{\det\left(\mathbf{A}_h^1\right)}\right)} && \text{(B.28)} \\ &\leq \sqrt{2dK\log\left(1+\frac{K}{d\lambda}\right)}. && \text{(B.29)} \end{aligned}$$

In inequality (B.28), we used the standard argument in regret analysis of linear bandits [2] (Lemma 11) as follows:

$$\sum_{t=1}^{n}\min\left(\|\mathbf{y}_t\|^2_{\mathbf{V}_t^{-1}}, 1\right) \leq 2\log\frac{\det\mathbf{V}_{n+1}}{\det\mathbf{V}_1} \quad \text{where} \quad \mathbf{V}_n = \mathbf{V}_1 + \sum_{t=1}^{n-1}\mathbf{y}_t\mathbf{y}_t^\top. \tag{B.30}$$

In inequality (B.29), we used Assumption 7 and the fact that $\det(\mathbf{A}) = \prod_{i=1}^{d}\lambda_i(\mathbf{A}) \leq (\mathrm{trace}(\mathbf{A})/d)^d$. Combining (B.26), (B.27), and (B.29), we have with probability at least $1-2\delta$

$$R_K \leq 2H\sqrt{T\log(dT/\delta)} + (1+\kappa)\beta\sqrt{2dHT\log\left(1+\frac{K}{d\lambda}\right)}. \tag{B.31}$$

B.1.4 Unknown $\tau_h(s)$

In this section, we relax Assumption 5, and instead assume that we only have the knowledge of safe actions $a_0(s)$, and remove the assumption on the knowledge about their costs $\tau_h(s)$. Similar results are provided by [96].

Let k be the number of times the agent has played action $a_0(s)$ at time-step h, and $\hat{\tau}_h(s)$ be the empirical mean estimator of $\tau_h(s)$. Then, for any $\delta \in (0,1)$, we have

$$\mathbb{P}\left(\tau_h(s) \leq \hat{\tau}_h(s) + \sqrt{2\log(1/\delta)/k}\right) \geq 1-\delta. \tag{B.32}$$

If we let $\delta = 1/K^2$, then we have

$$\mathbb{P}\left(\left|\hat{\tau}_h(s) - \tau_h(s)\right| \leq 2\sqrt{\log(K)/k},\ \forall k \in [K]\right) \geq 1 - 2/K. \tag{B.33}$$

We find $T_h(s)$, the number of time the agent must play action $a_0(s)$ at state s and time-step h in an adaptive manner as follow. Let $T_h(s)$ be the first time that $\hat{\tau}_h(s) + 6\sqrt{\log(K)/T_h(s)} \leq \tau$. Thus, we have

$$\tau_h(s) + 4\sqrt{\log(K)/T_h(s)} \leq \tau \Rightarrow \frac{16\log(K)}{(\tau - \tau_h(s))^2} \leq T_h(s). \tag{B.34}$$

Note that in this case $4\sqrt{\log(K)/T_h(s)}$ is a conservative estimation for $\tau - \tau_h(s)$.

Now we show that it will not take much longer than $\frac{16\log(K)}{(\tau-\tau_h(s))^2}$ that this first time happens. Conversely, for any $k \geq \frac{64\log(K)}{(\tau-\tau_h(s))^2}$, we observe that

$$\hat{\tau}_h(s) + 6\sqrt{\log(K)/k} \leq \tau_h(s) + 8\sqrt{\log(K)/k} \leq \tau. \tag{B.35}$$

Therefore, we conclude that

$$\frac{16\log(K)}{(\tau - \tau_h(s))^2} \leq T_h(s) \leq \frac{64\log(K)}{(\tau - \tau_h(s))^2}, \tag{B.36}$$

and $4\sqrt{\log(K)/T_h(s)}$ is a conservative estimate for $\tau - \tau_h(s)$.

B.2 Randomized SLUCB-QVI Proofs

In this section, we prove the technical statements in Section 3.6. First, recall the definition of the following event that we repeatedly refer to throughout this section:

$$\mathcal{E}_3 := \Big\{ \Big| \langle \tilde{\mathbf{w}}_h^k, \boldsymbol{\phi}(s,a) \rangle - \tilde{Q}_h^\pi(s,a) + [\mathbb{P}_h \tilde{V}_{h+1}^\pi - \tilde{V}_{h+1}^k](s,a) \Big| \leq \beta \big\| \boldsymbol{\phi}(s,a) \big\|_{\left(\mathbf{A}_h^k\right)^{-1}} \\ , \forall (a,s,h,k) \in \mathcal{A} \times \mathcal{S} \times [H] \times [K] \Big\}. \tag{B.37}$$

In the following theorem, we state that $\mathcal{E}_3$, focusing on randomized policy selection, is a high probability event.

Theorem 14 (Thm. 2 in [2] and Lemma B.4 in [64])**.** *Define*

$$\tilde{V}_h^k(s) := \min \left\{ \max_{\theta \in \Gamma_h^k(s)} \mathbb{E}_{a \sim \theta} \left[\tilde{Q}_h^k(s,a) \right], H \right\} \tag{B.38}$$

and recall the definition of $\mathcal{E}_1$ in (B.1)*. Then, for any fixed policy π, under Assumptions 4, 5, 6, and 7, and the definition of β in Theorem 4, there exists an absolute constant $c_\beta > 0$, such that for any fixed $\delta \in (0, 0.5)$, with probability at least $1 - 2\delta$, the event $\tilde{\mathcal{E}} := \mathcal{E}_1 \cap \mathcal{E}_3$ holds.*

B.2.1 Proof of Lemma 4

First, similar to vectors $\{\mathbf{w}_h^\pi\}_{h \in [H]}$ in (B.6) for deterministic policy selection setting, we introduce vectors $\{\tilde{\mathbf{w}}_h^\pi\}_{h \in [H]}$ for any policy π:

$$\tilde{\mathbf{w}}_h^\pi := \boldsymbol{\theta}_h^* + \int_{\mathcal{S}} \tilde{V}_{h+1}^\pi(s') d\boldsymbol{\mu}(s'). \tag{B.39}$$

From the Bellman equation in (3.12) and the linearity of the MDP in Assumption 4, we have:

$$\tilde{Q}_h^\pi(s,a) := \left\langle \boldsymbol{\phi}(s,a), \tilde{\mathbf{w}}_h^\pi \right\rangle. \tag{B.40}$$

Now, similar to the proof of Lemma 3, we start proving this Lemma 4 by induction. First, we prove the base case at time-step $H+1$. The statement holds because $\tilde{V}_{H+1}^*(s) =$

$\tilde{V}_{H+1}^k(s) = 0$. Now, suppose the statement holds for time-step $h+1$. We prove it also holds for time-step h. For all $(s,h,k) \in \mathcal{S} \times [H] \times [K]$, let

$$\pi_k(s,h) := \arg\max_{\theta \in \Gamma_h^k(s)} \mathbb{E}_{a\sim\theta}\left[\tilde{Q}_h^k(s,a)\right] \quad \text{and} \quad \pi_*(s,h) := \arg\max_{\theta \in \Gamma_h^{\text{safe}}(s)} \mathbb{E}_{a\sim\theta}\left[\tilde{Q}_h^*(s,a)\right]. \tag{B.41}$$

We consider the following two cases:

1) If $\pi_*(s,h) \in \Gamma_h^k(s)$, we have

$$\begin{aligned}
\tilde{V}_h^k(s) = \min\left\{\max_{\theta\in\Gamma_h^k(s)} \mathbb{E}_{a\sim\theta}\left[\tilde{Q}_h^k(s,a)\right], H\right\} &\geq \min\left\{\mathbb{E}_{a\sim\pi_*(s,h)}\left[\tilde{Q}_h^k(s,a)\right], H\right\} \\
&\geq \min\left\{\mathbb{E}_{a\sim\pi_*(s,h)}\left[\tilde{Q}_h^*(s,a) + \mathbb{E}_{s'\sim\mathbb{P}(.|s,a)}\left[\tilde{V}_{h+1}^k(s') - \tilde{V}_{h+1}^*(s')\right]\right]\right. \\
&\left., H\right\} \qquad \text{(Conditioned on } \mathcal{E}_3 \text{ in (B.37))} \\
&\geq \min\left\{\mathbb{E}_{a\sim\pi_*(s,h)}\left[\tilde{Q}_h^*(s,a)\right], H\right\}, \qquad \text{(Induction assumption)} \\
&= \mathbb{E}_{a\sim\pi_*(s,h)}\left[\tilde{Q}_h^*(s,a)\right] = V_h^*(s). \qquad \text{(B.42)}
\end{aligned}$$

as desired.

2) Now, we recall the definition of $\Gamma_h^k(s)$ in (3.14) and focus on the other case when $\pi_*(s,h) \notin \Gamma_h^k(s)$, which means

$$\begin{aligned}
&\frac{\left\langle \Phi_0\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right), \tilde{\boldsymbol{\phi}}\left(s, a_0(s)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s, a_0(s)\right)\right\|_2}\tau_h(s) \\
&+ \left\langle \boldsymbol{\gamma}_{h,s}^k, \Phi_0^\perp\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\rangle + \beta\left\|\Phi_0^\perp\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}} > \tau. \qquad \text{(B.43)}
\end{aligned}$$

Let $\pi_0(s,h)$ be the policy that always selects $a_0(s)$ for all $(s,h) \in \mathcal{S} \times [H]$. Now, we observe that $\pi_0(s,h) \in \Gamma_h^k(s)$. Recall that $\tilde{\boldsymbol{\phi}}\left(s, a_0(s)\right) = \frac{\boldsymbol{\phi}\left(s,a_0(s)\right)}{\left\|\boldsymbol{\phi}\left(s,a_0(s)\right)\right\|_2}$ and note that $\Phi_0\left(s, \boldsymbol{\phi}\left(s, a_0(s)\right)\right) = \boldsymbol{\phi}\left(s, a_0(s)\right)$ and $\Phi_0^\perp\left(s, \boldsymbol{\phi}(s, a_0(s))\right) = \mathbf{0}$. Thus

$$\frac{\left\langle \Phi_0\left(s, \boldsymbol{\phi}^{\pi_0(s,h)}(s)\right), \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)\right\|_2}\tau_h(s) + \left\langle \boldsymbol{\gamma}_{h,s}^k, \Phi_0^{\perp}\left(s, \boldsymbol{\phi}^{\pi_0(s,h)}(s)\right)\right\rangle + \beta\left\|\Phi_0^{\perp}\left(s, \boldsymbol{\phi}^{\pi_0(s,h)}(s)\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}$$
$$= \frac{\left\langle \boldsymbol{\phi}\left(s, a_0\left(s\right)\right), \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)\right\|_2}\tau_h(s) + \left\langle \boldsymbol{\gamma}_{h,s}^k, \Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_0(s))\right)\right\rangle + \beta\left\|\Phi_0^{\perp}\left(s, \boldsymbol{\phi}(s, a_0(s))\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}$$
$$= \tau_h(s) < \tau, \tag{B.44}$$

which implies that $\pi_0(s,h) \in \Gamma_h^k(s)$. Now, let $\tilde{\pi}_k(s,h) := \alpha_h^k(s)\pi_*(s,h) + (1-\alpha_h^k(s))\pi_0(s,h)$, where

$$\alpha_h^k(s) := \left\{\max \alpha \in [0,1] : \alpha\pi_*(s,h) + (1-\alpha)\pi_0(s,h) \in \Gamma_h^k(s)\right\}. \tag{B.45}$$

Let $\boldsymbol{\phi}^{\theta}(s) := \mathbb{E}_{a\sim\theta}\boldsymbol{\phi}(s,a)$. We observe that

$$\begin{aligned}\boldsymbol{\phi}^{\tilde{\pi}_k(s,h)}(s) &= \alpha_h^k(s)\boldsymbol{\phi}^{\pi_*(s,h)}(s) + (1-\alpha_h^k(s))\boldsymbol{\phi}^{\pi_0(s,h)}(s)\\ &= \alpha_h^k(s)\boldsymbol{\phi}^{\pi_*(s,h)}(s) + (1-\alpha_h^k(s))\boldsymbol{\phi}\left(s, a_0\left(s\right)\right).\end{aligned} \tag{B.46}$$

Since $\tilde{\pi}_k(s,h) \in \Gamma_h^k(s)$ (see the definition of $\alpha_h^k(s)$ in (B.45)), for all $(s,k) \in \mathcal{S}\times[K]$, at time-step h, we have

$$\tilde{V}_h^k(s) = \min\left\{\max_{\theta\in\Gamma_h^k(s)} \mathbb{E}_{a\sim\theta}\left[\tilde{Q}_h^k(s,a)\right], H\right\} \geq \min\left\{\mathbb{E}_{a\sim\tilde{\pi}(s,h)}\left[\tilde{Q}_h^k(s,a)\right], H\right\} \tag{B.47}$$

$$= \min\left\{\mathbb{E}_{a\sim\tilde{\pi}_k(s,h)}\left[\langle\tilde{\mathbf{w}}_h^k, \boldsymbol{\phi}(s,a)\rangle + \kappa_h(s)\beta\|\boldsymbol{\phi}(s,a)\|_{\left(\mathbf{A}_h^k\right)^{-1}}\right], H\right\} \tag{B.48}$$

$$\geq \min\left\{\langle\tilde{\mathbf{w}}_h^k, \boldsymbol{\phi}^{\tilde{\pi}_k(s,h)}(s)\rangle + \kappa_h(s)\beta\left\|\boldsymbol{\phi}^{\tilde{\pi}_k(s,h)}(s)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}, H\right\} \tag{Jensen's Inequality}$$

$$= \min\left\{\left\langle\tilde{\mathbf{w}}_h^k - \tilde{\mathbf{w}}_h^*, \boldsymbol{\phi}^{\tilde{\pi}_k(s,h)}(s)\right\rangle + \left\langle\tilde{\mathbf{w}}_h^*, \boldsymbol{\phi}^{\tilde{\pi}_k(s,h)}(s)\right\rangle\right.$$
$$\left.+\kappa_h(s)\beta\left\|\boldsymbol{\phi}^{\tilde{\pi}_k(s,h)}(s)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}, H\right\}. \tag{B.49}$$

Conditioned on event $\mathcal{E}_3$ in (B.37) and by the induction assumption, we have

$$-\beta\left\|\boldsymbol{\phi}^{\tilde{\pi}_k(s,h)}(s)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}} \leq \left\langle\tilde{\mathbf{w}}_h^k - \tilde{\mathbf{w}}_h^*, \boldsymbol{\phi}^{\tilde{\pi}_k(s,h)}(s)\right\rangle + \mathbb{E}_{a\sim\tilde{\pi}_k(s,h)}\left[\mathbb{E}_{s'\sim\mathbb{P}(.|s,a)}\left[\tilde{V}_{h+1}^k(s') - \tilde{V}_{h+1}^*(s')\right]\right]$$
$$\leq \left\langle\tilde{\mathbf{w}}_h^k - \tilde{\mathbf{w}}_h^*, \boldsymbol{\phi}^{\tilde{\pi}_k(s,h)}(s)\right\rangle. \tag{B.50}$$

By combining (B.49) and (B.50), we conclude that

$$\begin{aligned}
\tilde{V}_h^k(s) &\geq \min\left\{\left\langle \tilde{\mathbf{w}}_h^*, \boldsymbol{\phi}^{\tilde{\pi}_k(s,h)}(s)\right\rangle + (\kappa_h(s)-1)\beta\left\|\boldsymbol{\phi}^{\tilde{\pi}_k(s,h)}(s)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}, H\right\} \\
&= \min\left\{\alpha_h^k(s)\left\langle \tilde{\mathbf{w}}_h^*, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right\rangle + (1-\alpha_h^k(s))\left\langle \tilde{\mathbf{w}}_h^*, \boldsymbol{\phi}^{\pi_0(s,h)}(s)\right\rangle + (\kappa_h(s)-1)\beta\left\|\boldsymbol{\phi}^{\tilde{\pi}_k(s,h)}(s)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}, H\right\} \\
&\geq \min\left\{\alpha_h^k(s)\left\langle \tilde{\mathbf{w}}_h^*, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right\rangle + (\kappa_h(s)-1)\beta\left\|\Phi_0^{\perp}\left(s, \boldsymbol{\phi}^{\tilde{\pi}_k(s,h)}(s)\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}, H\right\} \\
&= \min\left\{\alpha_h^k(s)\left(\left\langle \tilde{\mathbf{w}}_h^*, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right\rangle + (\kappa_h(s)-1)\beta\left\|\Phi_0^{\perp}\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}\right), H\right\}
\end{aligned} \tag{B.51}$$

where the third inequality holds because $\left\|\boldsymbol{\phi}^{\tilde{\pi}_k(s,h)}(s)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}} \geq \left\|\Phi_0^{\perp}\left(s, \boldsymbol{\phi}^{\tilde{\pi}_k(s,h)}(s)\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}$ (see Lemma 3 in [96] for a proof) and $(1-\alpha_h^k(s))\left\langle \tilde{\mathbf{w}}_h^*, \boldsymbol{\phi}\left(s, a_0(s)\right)\right\rangle \geq 0$ as the reward is always positive, i.e., $r_h(s,a) \in [0,1]$ for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$. The second equality follows from the fact that

$$\begin{aligned}
\Phi_0^{\perp}\left(s, \boldsymbol{\phi}^{\tilde{\pi}_k(s,h)}(s)\right) &= \alpha_h^k(s)\Phi_0^{\perp}\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right) + (1-\alpha_h^k(s))\Phi_0^{\perp}\left(s, \boldsymbol{\phi}^{\pi_0(s,h)}(s)\right) \\
&= \alpha_h^k(s)\Phi_0^{\perp}\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right).
\end{aligned} \tag{B.52}$$

Now, we show that $\alpha_h^k(s) \geq \frac{\tau-\tau_h(s)}{\tau-\tau_h(s)+2\beta\left\|\Phi_0^{\perp}\left(s,\boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}}$, which eventually leads to a proper value for $\kappa_h(s) > 1$ that guarantees for all $(s,h,k) \in \mathcal{S}\times[H]\times[K]$ it holds that $\tilde{V}_h^*(s) \leq \tilde{V}_h^k(s)$ conditioned on $\tilde{\mathcal{E}} = \mathcal{E}_1 \cap \mathcal{E}_3$. Definitions of $\alpha_h^k(s)$ in (B.45) and the estimated safe set $\Gamma_h^k(s)$ in (3.14) imply that for all $(s,h,k) \in \mathcal{S}\times[H]\times[K]$, we have

$$\begin{aligned}
&\frac{(1-\alpha_h^k(s))\left\langle \boldsymbol{\phi}\left(s,a_0(s)\right), \tilde{\boldsymbol{\phi}}\left(s,a_0(s)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s,a_0(s)\right)\right\|_2}\tau_h(s) + \alpha_h^k(s)\left[\frac{\left\langle \Phi_0\left(s,\boldsymbol{\phi}^{\pi_*(s,h)}(s)\right), \tilde{\boldsymbol{\phi}}\left(s,a_0(s)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s,a_0(s)\right)\right\|_2}\tau_h(s)\right. \\
&\left. + \left\langle \boldsymbol{\gamma}_{h,s}^k, \Phi_0^{\perp}\left(s,\boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\rangle + \beta\left\|\Phi_0^{\perp}\left(s,\boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}\right] = \tau.
\end{aligned} \tag{B.53}$$

Let $M = \frac{\left\langle \Phi_0\left(s,\boldsymbol{\phi}^{\pi_*(s,h)}(s)\right), \tilde{\boldsymbol{\phi}}\left(s,a_0(s)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s,a_0(s)\right)\right\|_2}\tau_h(s) + \left\langle \boldsymbol{\gamma}_{h,s}^k, \Phi_0^{\perp}\left(s,\boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\rangle + \beta\left\|\Phi_0^{\perp}\left(s,\boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}$. Note that due to (B.43), $M > \tau$, and recall that

$\tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right) = \frac{\boldsymbol{\phi}\left(s,a_0(s)\right)}{\left\|\boldsymbol{\phi}\left(s,a_0(s)\right)\right\|_2}$. Thus, (B.53) gives

$$0 < \alpha_h^k(s) = \frac{\tau - \tau_h(s)}{M - \tau_h(s)} < 1. \tag{B.54}$$

In order to lower bound $\alpha_h^k(s)$ (upper bound M), we first rewrite M as

$$\begin{aligned} M =& \frac{\left\langle \Phi_0\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right), \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)\right\|_2}\tau_h(s) + \left\langle \boldsymbol{\gamma}_h^*, \Phi_0^\perp\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\rangle \\ &+ \left\langle \boldsymbol{\gamma}_{h,s}^k - \boldsymbol{\gamma}_h^*, \Phi_0^\perp\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\rangle + \beta\left\|\Phi_0^\perp\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}, \end{aligned} \tag{B.55}$$

and show that

(a) $\frac{\left\langle \Phi_0\left(s,\boldsymbol{\phi}^{\pi_*(s,h)}(s)\right), \tilde{\boldsymbol{\phi}}\left(s,a_0(s)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s,a_0(s)\right)\right\|_2}\tau_h(s) + \left\langle \boldsymbol{\gamma}_h^*, \Phi_0^\perp\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\rangle \leq \tau$ because

$$\begin{aligned} &\frac{\left\langle \Phi_0\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right), \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)\right\rangle}{\left\|\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)\right\|_2}\tau_h(s) + \left\langle \boldsymbol{\gamma}_h^*, \Phi_0^\perp\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\rangle \\ &= \left\langle \boldsymbol{\gamma}_h^*, \left\langle \Phi_0\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right), \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)\right\rangle \tilde{\boldsymbol{\phi}}\left(s, a_0\left(s\right)\right)\right\rangle \\ &+ \left\langle \boldsymbol{\gamma}_h^*, \Phi_0^\perp\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\rangle \\ &= \left\langle \boldsymbol{\gamma}_h^*, \Phi_0\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\rangle + \left\langle \boldsymbol{\gamma}_h^*, \Phi_0^\perp\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\rangle \\ &= \left\langle \boldsymbol{\gamma}_h^*, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right\rangle \\ &\leq \tau. \end{aligned} \tag{B.56}$$

(b) $\left\langle \boldsymbol{\gamma}_{h,s}^k - \boldsymbol{\gamma}_h^*, \Phi_0^\perp\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\rangle \leq \beta\left\|\Phi_0^\perp\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}$, because conditioned on $\mathcal{E}_1$ in (B.1), we have

$$\begin{aligned} \left\langle \boldsymbol{\gamma}_{h,s}^k - \boldsymbol{\gamma}_h^*, \Phi_0^\perp\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\rangle &= \left\langle \boldsymbol{\gamma}_{h,s}^k - \Phi_0^\perp\left(s, \boldsymbol{\gamma}_h^*\right), \Phi_0^\perp\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\rangle \\ &\leq \beta\left\|\Phi_0^\perp\left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s)\right)\right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}. \end{aligned} \tag{B.57}$$

Now, we combine (B.55), (B.56) and (B.57) to conclude that

$$M \leq \tau + 2\beta \left\| \Phi_0^{\perp} \left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s) \right) \right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}} \quad \Rightarrow \quad \alpha_h^k(s) \geq \frac{\tau - \tau_h(s)}{\tau - \tau_h(s) + 2\beta \left\| \Phi_0^{\perp} \left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s) \right) \right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}}. \tag{B.58}$$

This lower bound on $\alpha_h^k(s)$ combined with (B.51) gives

$$\tilde{V}_h^k(s) \geq \min \left\{ \frac{\left(\tau - \tau_h(s)\right) \left(\tilde{V}_h^*(s) + (\kappa_h(s) - 1)\beta \left\| \Phi_0^{\perp} \left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s) \right) \right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}} \right)}{\tau - \tau_h(s) + 2\beta \left\| \Phi_0^{\perp} \left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s) \right) \right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}}, H \right\} \tag{B.59}$$

Let $M_1 = \beta \left\| \Phi_0^{\perp} \left(s, \boldsymbol{\phi}^{\pi_*(s,h)}(s) \right) \right\|_{\left(\mathbf{A}_{h,s}^k\right)^{-1}}$. Thus, $\tilde{V}_h^k(s) \geq \tilde{V}_h^*(s) = \min\left\{V_h^*(s), H\right\}$, if and only if

$$\left(\tau - \tau_h(s)\right) \left(V_h^*(s) + (\kappa_h(s) - 1)M_1 \right) \geq \left(\tau - \tau_h(s) + 2M_1\right) V_h^*(s), \tag{B.60}$$

which is true if and only if

$$\left(\tau - \tau_h(s)\right)(\kappa_h(s) - 1) \geq 2V_h^*(s) \iff \left(\tau - \tau_h(s)\right)(\kappa_h(s) - 1) \geq 2H \iff \kappa_h(s) \geq \frac{2H}{\tau - \tau_h(s)} + 1, \tag{B.61}$$

as desired.

B.2.2 Proof of Theorem 6

Conditioned on event $\mathcal{E}_3$, for any $(a, s, h, k) \in \mathcal{A} \times \mathcal{S} \times [H] \times [K]$, we have

$$\begin{aligned} \tilde{Q}_h^k(s,a) - \tilde{Q}_h^{\pi_k}(s,a) &= \left\langle \tilde{\mathbf{w}}_h^k, \boldsymbol{\phi}(s,a) \right\rangle + \kappa_h(s)\beta \left\| \boldsymbol{\phi}(s,a) \right\|_{\left(\mathbf{A}_h^k\right)^{-1}} - \tilde{Q}_h^{\pi_k}(s,a) \\ &\leq \mathbb{E}_{s' \sim \mathbb{P}(.|s,a)} \left[\tilde{V}_{h+1}^k(s') - \tilde{V}_{h+1}^{\pi_k}(s') \right] + \left(1 + \kappa_h(s)\right) \beta \left\| \boldsymbol{\phi}(s,a) \right\|_{\left(\mathbf{A}_h^k\right)^{-1}}. \end{aligned} \tag{B.62}$$

Let $\delta_h^k := \tilde{V}_h^k(s_h^k) - \tilde{V}_h^{\pi_k}(s_h^k)$ and $\zeta_{h+1}^k := \mathbb{E}_{a\sim\pi_k(s_h^k,h)}\left[\mathbb{E}_{s'\sim\mathbb{P}(.|s_h^k,a)}\left[\tilde{V}_{h+1}^k(s') - \tilde{V}_{h+1}^{\pi_k}(s')\right]\right] - \delta_{h+1}^k$. We can write

$$\begin{aligned}
\delta_h^k &= \tilde{V}_h^k(s_h^k) - \tilde{V}_h^{\pi_k}(s_h^k) \\
&= \min\left\{\max_{\theta\in\Gamma_h^k(s_h^k)} \mathbb{E}_{a\sim\theta}\left[\tilde{Q}_h^k(s_h^k,a)\right], H\right\} - \mathbb{E}_{a\sim\pi_k(s_h^k,h)}\left[\tilde{Q}_h^{\pi_k}(s_h^k,a)\right] \\
&\le \max_{\theta\in\Gamma_h^k(s_h^k)} \mathbb{E}_{a\sim\theta}\left[\tilde{Q}_h^k(s_h^k,a)\right] - \mathbb{E}_{a\sim\pi_k(s_h^k,h)}\left[\tilde{Q}_h^{\pi_k}(s_h^k,a)\right] \\
&= \mathbb{E}_{a\sim\pi_k(s_h^k,h)}\left[\tilde{Q}_h^k(s_h^k,a)\right] - \mathbb{E}_{a\sim\pi_k(s_h^k,h)}\left[\tilde{Q}_h^{\pi_k}(s_h^k,a)\right] \\
&\le \mathbb{E}_{a\sim\pi_k(s_h^k,h)}\left[\mathbb{E}_{s'\sim\mathbb{P}(.|s_h^k,a)}\left[\tilde{V}_{h+1}^k(s') - \tilde{V}_{h+1}^{\pi_k}(s')\right]\right] + \left(1+\kappa_h(s)\right)\beta\mathbb{E}_{a\sim\pi_k(s_h^k,h)}\left[\left\|\boldsymbol{\phi}(s_h^k,a)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(Eqn. (B.62))} \\
&= \delta_{h+1}^k + \zeta_{h+1}^k + \left(1+\kappa_h(s)\right)\beta\mathbb{E}_{a\sim\pi_k(s_h^k,h)}\left[\left\|\boldsymbol{\phi}(s_h^k,a)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}\right], \qquad \text{(B.63)}
\end{aligned}$$

Now, conditioning on event $\tilde{\mathcal{E}}$ defined in Theorem 14, we bound the cumulative regret as follows:

$$\begin{aligned}
R_K = \sum_{k=1}^K \tilde{V}_1^*(s_1^k) - \tilde{V}_1^{\pi_k}(s_1^k) &\le \sum_{k=1}^K \delta_1^k \qquad \text{(Lemma 4)} \\
&\le \sum_{k=1}^K\sum_{h=1}^H \zeta_h^k + \sum_{k=1}^K\sum_{h=1}^H \left(1+\kappa_h(s)\right)\beta\mathbb{E}_{a\sim\pi_k(s_h^k,h)}\left[\left\|\boldsymbol{\phi}(s_h^k,a)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}\right] \\
&\le \sum_{k=1}^K\sum_{h=1}^H \zeta_h^k + (1+\kappa)\,\beta\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{a\sim\pi_k(s_h^k,h)}\left[\left\|\boldsymbol{\phi}(s_h^k,a)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}\right]. \qquad \text{(B.64)}
\end{aligned}$$

We observe that $\{\zeta_h^k\}$ is a martingale difference sequence satisfying $|\zeta_h^k| \le 2H$. Thus, thanks to Azuma-Hoeffding inequality, we have

$$\mathbb{P}\left(\sum_{k=1}^K\sum_{h=1}^H \zeta_h^k \le 2H\sqrt{T\log(dT/\delta)}\right) \ge 1-\delta. \qquad \text{(B.65)}$$

In order to bound $\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{a\sim\pi_k(s_h^k,h)}\left[\left\|\boldsymbol{\phi}(s_h^k,a)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}\right]$, we define the martingle difference sequence $\iota_h^k := \mathbb{E}_{a\sim\pi_k(s_h^k,h)}\left[\left\|\boldsymbol{\phi}(s_h^k,a)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}\right] - \left\|\boldsymbol{\phi}(s_h^k,a_h^k)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}$, and note that for any

$(h,k) \in [H] \times [k]$, we have $|\iota_h^k| \leq 2/\sqrt{\lambda}$. Thus, thanks to Azuma-Hoeffding inequality, we have

$$\mathbb{P}\left(\sum_{k=1}^{K}\sum_{h=1}^{H} \iota_h^k \leq 2\sqrt{T\log(dT/\delta)/\lambda}\right) \geq 1-\delta. \tag{B.66}$$

Now, we are ready to bound $\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{a\sim\pi_k(s_h^k,h)}\left[\left\|\boldsymbol{\phi}(s_h^k,a)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}\right]$ as follows:

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{a\sim\pi_k(s_h^k,h)}\left[\left\|\boldsymbol{\phi}(s_h^k,a)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}\right] \leq 2\sqrt{T\log(dT/\delta)/\lambda} + \sum_{k=1}^{K}\sum_{h=1}^{H}\left\|\boldsymbol{\phi}(s_h^k,a_h^k)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}}. \tag{B.67}$$

In order to bound the second term, we have

$$\begin{aligned}
\sum_{k=1}^{K}\left\|\boldsymbol{\phi}(s_h^k,a_h^k)\right\|_{\left(\mathbf{A}_h^k\right)^{-1}} &\leq \sqrt{K\sum_{k=1}^{K}\left\|\boldsymbol{\phi}(s_h^k,a_h^k)\right\|^2_{\left(\mathbf{A}_h^k\right)^{-1}}} && \text{(Cauchy-Schwartz inequality)}\\
&\leq \sqrt{2K\log\left(\frac{\det\left(\mathbf{A}_h^K\right)}{\det\left(\mathbf{A}_h^1\right)}\right)} && \text{(B.68)}\\
&\leq \sqrt{2dK\log\left(1+\frac{K}{d\lambda}\right)}. && \text{(B.69)}
\end{aligned}$$

In inequality (B.68), we used the standard argument in regret analysis of linear bandits [2] (Lemma 11) as follows:

$$\sum_{t=1}^{n}\min\left(\|\mathbf{y}_t\|^2_{\mathbf{V}_t^{-1}},1\right) \leq 2\log\frac{\det\mathbf{V}_{n+1}}{\det\mathbf{V}_1} \quad \text{where} \quad \mathbf{V}_n = \mathbf{V}_1 + \sum_{t=1}^{n-1}\mathbf{y}_t\mathbf{y}_t^\top. \tag{B.70}$$

In inequality (B.69), we used Assumption 7 and the fact that $\det(\mathbf{A}) = \prod_{i=1}^{d}\lambda_i(\mathbf{A}) \leq (\ \mathrm{trace}(\mathbf{A})/d)^d$.

Combining (B.64), (B.65), (B.66), and (B.69), we have with probability at least $1-3\delta$

$$R_K \leq 2H\sqrt{T\log(dT/\delta)} + 2(1+\kappa)\beta\sqrt{2dHT\log\left(1+\frac{Td}{\delta}\right)/\lambda}. \tag{B.71}$$

B.3 Finite Star Convex Sets and Tractability of the Experiments

In this section, we show that if for all $s \in \mathcal{S}$, the sets $\mathcal{D}(s) = \{\boldsymbol{\phi}(s,a) : a \in \mathcal{A}\}$ are star convex and *finite* around $\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)$ (see Definition 1), then the optimization problem in Line 10 of SLUCB-QVI can be solved efficiently. Thanks to Definition 1, for each $s \in \mathcal{S}$, there exist finite number N of vectors $\boldsymbol{\phi}\left(s, a_i\left(s\right)\right)$ such that we can write $\mathcal{D}(s_h^k)$ as: $\mathcal{D}(s) := \cup_{i=1}^{N} \left[\boldsymbol{\phi}\left(s, a_0\left(s\right)\right), \boldsymbol{\phi}\left(s, a_i\left(s\right)\right)\right]$, where $\left[\boldsymbol{\phi}\left(s, a_0\left(s\right)\right), \boldsymbol{\phi}\left(s, a_i\left(s\right)\right)\right]$ is the line connecting $\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)$ to $\boldsymbol{\phi}\left(s, a_i\left(s\right)\right)$. Since $\boldsymbol{\phi}\left(s, a_0\left(s\right)\right) \in \mathcal{D}_h^k(s) := \left\{\boldsymbol{\phi}(s,a) : a \in \mathcal{A}_h^k(s)\right\}$, the set $\mathcal{D}_h^k(s)$ is also a finite star convex set around $\boldsymbol{\phi}\left(s, a_0\left(s\right)\right)$, and can be written as $\mathcal{D}_h^k(s) := \cup_{i=1}^{N} \left[\boldsymbol{\phi}\left(s, a_0\left(s\right)\right), \boldsymbol{\phi}\left(s, a_{i,h}^k\left(s\right)\right)\right]$, where $\boldsymbol{\phi}\left(s, a_{i,h}^k\left(s\right)\right) = \alpha_{i,h}^{s,k} \boldsymbol{\phi}\left(s, a_i\left(s\right)\right) + (1 - \alpha_{i,h}^{s,k}) \boldsymbol{\phi}\left(s, a_0\left(s\right)\right)$ and $\alpha_{i,h}^{s,k} = \max\left\{\alpha \in [0,1] : \alpha \boldsymbol{\phi}\left(s, a_i\left(s\right)\right) + (1-\alpha)\boldsymbol{\phi}\left(s, a_0\left(s\right)\right) \in \mathcal{D}_h^k(s)\right\}$, which can be solved by doing line search. The optimization problem at Line 10 of Algorithm 3 is equivalent to

$$\max_{\mathbf{x} \in \mathcal{D}_h^k(s_h^k)} \left\langle \mathbf{w}_h^k, \mathbf{x} \right\rangle + \kappa_h(s_h^k) \beta \|\mathbf{x}\|_{\left(\mathbf{A}_h^k\right)^{-1}}, \tag{B.72}$$

which can be executed by optimizing over each line $\left[\boldsymbol{\phi}\left(s_h^k, a_0\left(s_h^k\right)\right), \boldsymbol{\phi}\left(s_h^k, a_{i,h}^k\left(s_h^k\right)\right)\right]$ for all $i \in [N]$. Note that $\left\langle \mathbf{w}_h^k, \mathbf{x} \right\rangle + \kappa_h(s_h^k) \beta \|\mathbf{x}\|_{\left(\mathbf{A}_h^k\right)^{-1}}$ is a convex function in $\mathbf{x}$. Therefore, its maximum over the line $\left[\boldsymbol{\phi}\left(s_h^k, a_0\left(s_h^k\right)\right), \boldsymbol{\phi}\left(s_h^k, a_{i,h}^k\left(s_h^k\right)\right)\right]$ is achieved at either $\boldsymbol{\phi}\left(s_h^k, a_0\left(s_h^k\right)\right)$ or $\boldsymbol{\phi}\left(s_h^k, a_{i,h}^k\left(s_h^k\right)\right)$, which makes the optimization problem at line 10 of Algorithm 3 easy and tractable.

APPENDIX C

Proofs for Chapter 4

C.1 Analysis of Safe-DPVI

In this section, we first prove Lemma 5 and then prove the three points stated in Theorem 7.

C.1.1 Proof of Lemma 5

First, we summarize Lemma A.1 in [65] and Lemma 4.2 in [36] in Lemma 13.

Lemma 13. *Let π and π' be two arbitrary policies and let Q be any given Q-function such that $V_h(s) = \mathbb{E}_{a\sim\pi_h(.|s)}\left[Q_h(s,a)\right]$ for all $(s,h) \in \mathcal{S}\times[H]$. Then*

$$V_1(s) - V_1^{\pi'}(s) = \sum_{h=1}^{H} \mathbb{E}\left[\mathbb{E}_{a\sim\pi_h(.|s_h)}\left[Q_h(s_h,a)\right] - \mathbb{E}_{a\sim\pi'_h(.|s_h)}\left[Q_h(s_h,a)\right] \middle| s_1 = s, \pi'\right] + \sum_{h=1}^{H} \mathbb{E}\left[Q_h(s_h,a_h) - [\mathbb{B}_h V_{h+1}](s_h,a_h) \middle| s_1 = s, \pi'\right]. \tag{C.1}$$

Now, recall the definition of suboptimality gap $\Delta(\pi; s)$ in (4.5). We have

$$\Delta(\hat{\pi}; s) = \underbrace{V_1^*(s) - \hat{V}_1(s)}_{\text{Term I}} + \hat{V}_1(s) - V_1^{\hat{\pi}}(s). \tag{C.2}$$

Let $\pi = \pi' = \hat{\pi}$. Thus, applying Lemma 13, we have

$$\begin{aligned}
\hat{V}_1(s) - V_1^{\hat{\pi}}(s) &= \min\left\{\mathbb{E}_{a\sim\hat{\pi}_1(.|s)}\left[\hat{Q}_1(s,a)\right], H\right\} - V_1^{\hat{\pi}}(s) \\
&\leq \mathbb{E}_{a\sim\hat{\pi}_1(.|s)}\left[\hat{Q}_1(s,a)\right] - V_1^{\hat{\pi}}(s) \\
&= \sum_{h=1}^{H} \mathbb{E}\left[\hat{Q}_h(s_h,a_h) - [\mathbb{B}_h\bar{V}_{h+1}](s_h,a_h)\middle| s_1 = s, \hat{\pi}\right] \\
&\leq \sum_{h=1}^{H} \mathbb{E}\left[\hat{Q}_h(s_h,a_h) - [\mathbb{B}_h\hat{V}_{h+1}](s_h,a_h)\middle| s_1 = s, \hat{\pi}\right] \\
&= \sum_{h=1}^{H} \mathbb{E}\left[-\iota_h(s_h,a_h)\middle| s_1 = s, \hat{\pi}\right] = \text{ Term II},
\end{aligned}$$

which concludes Lemma 5.

C.1.2 Proof of Lemma 6

First note that

$$\begin{aligned}
[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - [\mathbb{B}_h\hat{V}_{h+1}](s,a) &= [\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - R_h(s,a) - [\mathbb{P}_h\hat{V}_{h+1}](s,a) \\
&= [\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - Q_h^{\pi}(s,a) + Q_h^{\pi}(s,a) - R_h(s,a) - [\mathbb{P}_h\hat{V}_{h+1}](s,a) \\
&= [\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - Q_h^{\pi}(s,a) + R_h(s,a) + [\mathbb{P}_hV_{h+1}^{\pi}](s,a) - R_h(s,a) \\
&\quad - [\mathbb{P}_h\hat{V}_{h+1}](s,a) \\
&= [\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - Q_h^{\pi}(s,a) - \left[\mathbb{P}_h\left(\hat{V}_{h+1} - V_{h+1}^{\pi}\right)\right](s,a).
\end{aligned}$$

Thus, if B' is a δ-Bellman uncertainty quantifier, then for any policy π and $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, with probability at least $1-\delta$, it holds that

$$\left|[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - Q_h^{\pi}(s,a) - \left[\mathbb{P}_h\left(\hat{V}_{h+1} - V_{h+1}^{\pi}\right)\right](s,a)\right| \leq B_h'(s,a). \tag{C.3}$$

Now, we start the formal proof of the lemma. We prove this lemma by induction. First, we prove the base case at time-step $H+1$. The statement holds for $H+1$ because $F_{H+1}(s) = 0 = V_{H+1}^*(s) = \hat{V}_{H+1}(s) = 0$. Now, suppose the statement holds for time-step $h+1$. We prove it also holds for time-step h. We consider the following two cases:

1) If $\pi_h^*(.|s) \in \hat{\Gamma}_h(s)$, we have

$$\begin{aligned}
\hat{V}_h(s) + F_h(s) &= \min\left\{\mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[\hat{Q}_h(s,a)\right], H\right\} + F_h(s) \\
&\geq \min\left\{\mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[\hat{Q}_h(s,a)\right], H\right\} + \sum_{h'=h}^{H} \alpha_{h'}\mathbb{E}\left[\bar{B}_{h'}(s_{h'},a_{h'})|s_h = s, \pi^*\right] \\
&\geq \min\left\{\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\hat{Q}_h(s,a)\right], H\right\} + \sum_{h'=h}^{H} \alpha_{h'}\mathbb{E}\left[\bar{B}_{h'}(s_{h'},a_{h'})|s_h = s, \pi^*\right] \\
&\geq \min\left\{\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[Q_h^*(s,a) + \left[\mathbb{P}_h\left(\hat{V}_{h+1} - V_{h+1}^*\right)\right](s,a) - 2B_h'(s,a)\right], H\right\} \\
&\quad + \sum_{h'=h}^{H} \alpha_{h'}\mathbb{E}\left[\bar{B}_{h'}(s_{h'},a_{h'})|s_h = s, \pi^*\right] \qquad \text{(Eqn. (C.3))} \\
&\geq \min\left\{\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[Q_h^*(s,a) + \alpha_h\bar{B}_h(s,a) - 2B_h(s,a)\right], H\right\} \\
&\qquad\qquad \text{(Induction assumption)} \\
&\geq \min\left\{\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[Q_h^*(s,a) + (\alpha_h - 2)\bar{B}_h(s,a)\right], H\right\} \\
&= \min\left\{V_h^*(s), H\right\} \qquad (\bigstar) \\
&= V_h^*(s).
\end{aligned}$$

$\bigstar$ is true because $\alpha_h \geq 2$.

2) Now, we focus on the other case when $\pi_h^*(.|s) \notin \hat{\Gamma}_h(s)$, which means

$$\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[u_h^c(s,a)\right] > \tau. \tag{C.4}$$

Let $\tilde{\pi}_h(.|s) := \gamma_h(s)\pi_h^*(.|s) + (1-\gamma_h(s))\pi_h^0(.|s)$, where

$$\gamma_h(s) := \left\{\max \gamma \in [0,1] : \gamma\pi_h^*(.|s) + (1-\gamma)\pi_h^0(.|s) \in \hat{\Gamma}_h(s)\right\}. \tag{C.5}$$

Now, we show that $\gamma_h(s) \geq \frac{\tau - \tau_h(s)}{\tau - \tau_h(s) + 2\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\bar{B}_h(s,a)\right]}$, which eventually leads to a proper value for α_h that guarantees for all $s \in \mathcal{S}$, with probability at least $1-2\delta$, it holds that $\hat{V}_h(s) + F_h(s) \geq V_h^*(s)$. Definitions of $\gamma_h(s)$ in (C.5) and the estimated safe set $\hat{\Gamma}_h(s)$ in (4.6)

imply that

$$\begin{aligned}\mathbb{E}_{a\sim\tilde{\pi}_h(.|s)}\left[u_h^c(s,a)\right] &= \gamma_h(s)\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[u_h^c(s,a)\right] + (1-\gamma_h(s))\mathbb{E}_{a\sim\pi_h^0(.|s)}\left[u_h^c(s,a)\right] \\ &= \gamma_h(s)\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[u_h^c(s,a)\right] + (1-\gamma_h(s))\tau_h(s) \\ &\leq \tau. \end{aligned} \tag{C.6}$$

Thus

$$0 < \gamma_h(s) = \frac{\tau - \tau_h(s)}{\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[u_h^c(s,a)\right] - \tau_h(s)} < 1. \tag{C.7}$$

Recall the definition of $\Gamma_h^{\text{safe}}(s)$ in (4.2) and note that $\pi_h^*(.|s) \in \Gamma_h^{\text{safe}}(s)$. Due to the definition δ-safety uncertainty quantifier B, for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, with probability at least $1-\delta$, it holds that

$$\begin{aligned}\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[u_h^c(s,a)\right] &\leq \mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\hat{C}_h(s,a) + B_h(s,a)\right] \\ &\leq \mathbb{E}_{a\sim\pi_h^*(.|s)}\left[C_h(s,a) + 2B_h(s,a)\right] \\ &\leq \tau + 2\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[B_h(s,a)\right] && (\pi_h^*(.|s) \in \Gamma_h^{\text{safe}}(s)) \\ &\leq \tau + 2\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\bar{B}_h(s,a)\right]. \end{aligned} \tag{C.8}$$

Combining (C.7) and (C.8), we conclude that

$$\gamma_h(s) \geq \frac{\tau - \tau_h(s)}{\tau - \tau_h(s) + 2\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\bar{B}_h(s,a)\right]}. \tag{C.9}$$

We have

$$
\begin{aligned}
\hat{V}_h(s) + F_h(s) &= \min\left\{\mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[\hat{Q}_h(s,a)\right], H\right\} + F_h(s) \\
&= \min\left\{\mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[\left\{[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a)\right\}^+\right], H\right\} + F_h(s) \\
&\geq \min\left\{\mathbb{E}_{a\sim\tilde{\pi}_h(.|s)}\left[\left\{[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a)\right\}^+\right], H\right\} + F_h(s) \\
&\geq \min\left\{\mathbb{E}_{a\sim\tilde{\pi}_h(.|s)}\left[[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a)\right], H\right\} + F_h(s) \\
&= \min\Bigg\{\gamma_h(s)\mathbb{E}_{a\sim\pi^*_h(.|s)}\left[[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a)\right] \\
&\quad +(1-\gamma_h(s))\mathbb{E}_{a\sim\pi^0_h(.|s)}\left[[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a)\right], H\Bigg\} + F_h(s) \\
&\geq \min\Bigg\{\gamma_h(s)\left(\mathbb{E}_{a\sim\pi^*_h(.|s)}\left[[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a)\right] + F_h(s)\right) \\
&\quad +(1-\gamma_h(s))\left(\mathbb{E}_{a\sim\pi^0_h(.|s)}\left[[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a)\right] + F_h(s)\right), H\Bigg\} \\
&\geq \min\left\{\gamma_h(s)\left(\mathbb{E}_{a\sim\pi^*_h(.|s)}\left[[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a)\right] + F_h(s)\right), H\right\} \qquad (\bigstar) \\
&\geq \min\left\{\gamma_h(s)\mathbb{E}_{a\sim\pi^*_h(.|s)}\left[Q^*_h(s,a) + (\alpha_h - 2)\bar{B}_h(s,a)\right], H\right\}. \qquad (\bigstar\bigstar)
\end{aligned}
$$

$\bigstar$ is true because $(1-\gamma_h(s)) \geq 0$ and

$$
\begin{aligned}
&\mathbb{E}_{a\sim\pi^0_h(.|s)}\left[[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a)\right] + F_h(s) \\
&\geq \mathbb{E}_{a\sim\pi^0_h(.|s)}\left[Q^0_h(s,a) + \left[\mathbb{P}_h\left(\hat{V}_{h+1} - V^0_{h+1}\right)\right](s,a) - 2B'_h(s,a)\right] + F_h(s) && \text{(Equation (C.3))} \\
&\geq \mathbb{E}_{a\sim\pi^0_h(.|s)}\left[Q^0_h(s,a) + \left[\mathbb{P}_h\left(\hat{V}_{h+1} - V^0_{h+1}\right)\right](s,a) - 2B'_h(s,a)\right] + \sum_{h'=h}^{H}\alpha_{h'}\mathbb{E}\left[\bar{B}_{h'}(s_{h'},a_{h'})|s_h = s,\pi^0\right] \\
& && \text{(Definition of } F_h(s) \text{ in Lemma 6)} \\
&\geq \mathbb{E}_{a\sim\pi^0_h(.|s)}\left[Q^0_h(s,a) + \alpha_h\bar{B}_h(s,a) - 2B'_h(s,a)\right] && \text{(Induction assumption)} \\
&\geq \mathbb{E}_{a\sim\pi^0_h(.|s)}\left[Q^0_h(s,a) + (\alpha_h - 2)\bar{B}_h(s,a)\right] \\
&\geq \mathbb{E}_{a\sim\pi^0_h(.|s)}\left[Q^0_h(s,a)\right] && (\alpha_h \geq 2) \\
&\geq 0.
\end{aligned}
$$

★★ is true because

$$
\begin{aligned}
&\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a)\right] + F_h(s) \\
&\geq \mathbb{E}_{a\sim\pi_h^*(.|s)}\left[Q_h^*(s,a) + \left[\mathbb{P}_h\left(\hat{V}_{h+1} - V_{h+1}^*\right)\right](s,a) - 2B'_h(s,a)\right] + F_h(s) \quad \text{(Equation (C.3))} \\
&\geq \mathbb{E}_{a\sim\pi_h^*(.|s)}\left[Q_h^*(s,a) + \left[\mathbb{P}_h\left(\hat{V}_{h+1} - V_{h+1}^*\right)\right](s,a) - 2B'_h(s,a)\right] \\
&+ \sum_{h'=h}^{H}\alpha_{h'}\mathbb{E}\left[\bar{B}_{h'}(s_{h'},a_{h'})|s_h = s,\pi^*\right] \qquad \text{(Definition of } F_h(s) \text{ in Lemma 6)} \\
&\geq \mathbb{E}_{a\sim\pi_h^*(.|s)}\left[Q_h^*(s,a) + \alpha_h\bar{B}_h(s,a) - 2B'_h(s,a)\right] \qquad \text{(Induction assumption)} \\
&\geq \mathbb{E}_{a\sim\pi_h^*(.|s)}\left[Q_h^*(s,a) + (\alpha_h - 2)\bar{B}_h(s,a)\right].
\end{aligned}
$$

Now, we continue from ★★ and observe that $\hat{V}_h(s) + F_h(s) \geq V_h^*(s)$ if and only if

$$
\begin{aligned}
&\gamma_h(s)\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[Q_h^*(s,a) + (\alpha_h - 2)\bar{B}_h(s,a)\right] \geq V_h^*(s) \\
&\overset{\text{(C.9)}}{\iff} \frac{\left(\tau - \tau_h(s)\right)V_h^*(s) + \left(\tau - \tau_h(s)\right)\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[(\alpha_h - 2)\bar{B}_h(s,a)\right]}{\tau - \tau_h(s) + 2\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\bar{B}_h(s,a)\right]} \geq V_h^*(s) \\
&\iff (\alpha_h - 2)\left(\tau - \tau_h(s)\right)\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\bar{B}_h(s,a)\right] \geq 2\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\bar{B}_h(s,a)\right]V_h^*(s) \\
&\overset{H\geq V_h^*(s)}{\iff} (\alpha_h - 2)\left(\tau - \tau_h(s)\right)\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\bar{B}_h(s,a)\right] \geq 2\mathbb{E}_{a\sim\pi_h^*(.|s)}\left[\bar{B}_h(s,a)\right]H \\
&\iff \alpha_h \geq 2 + \frac{2H}{\tau - \tau_h(s)}
\end{aligned}
$$

as desired.

C.1.3 Proof of Theorem 7

Proof of point 1 of Theorem 7 Recall the definition of $\hat{\Gamma}_h(s)$ in (4.6). Since B is a δ-safety uncertainty quantifier, thus for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, $u_h^c(s,a)$ is an upper bound on $C_h(s,a)$ with probability at least $1-\delta$. Thus,

1. $\pi^0 \in \hat{\Pi}$ or equivalently $\pi_h^0(.|s) \in \hat{\Gamma}_h(s)$, for all $(s,h) \in \mathcal{S} \times [H]$ because

$$\begin{aligned}
\mathbb{E}_{a\sim\pi_h^0(.|s)}\left[u_h^c(s,a)\right] &= \mathbb{E}_{a\sim\pi_h^0(.|s)}\left[\hat{C}_h(s,a) + B_h(s,a)\right] \\
&\leq \mathbb{E}_{a\sim\pi_h^0(.|s)}\left[C_h(s,a) + 2B_h(s,a)\right] \\
&= \tau_h(s) + 2\mathbb{E}_{a\sim\pi_h^0(.|s)}\left[B_h(s,a)\right] \\
&\leq \tau. \qquad \text{(C.10)}
\end{aligned}$$

2. $\hat{\pi}_h(.|s) \in \left\{\theta(.|s) \in \Delta_{\mathcal{A}} : \mathbb{E}_{a\sim\theta(.|s)}\left[u_h^c(s,a)\right] \leq \tau\right\}$, which implies that with probability at least $1-\delta$, it holds that

$$\mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[C_h(s,a)\right] \leq \mathbb{E}_{a\sim\hat{\pi}_h(.|s)}\left[u_h^c(s,a)\right] \leq \tau, \qquad \text{(C.11)}$$

This concludes point 1 of Theorem 7.

Proof of point 2 of Theorem 7 Note that if $[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a) < 0$, then $\hat{Q}_h(s,a) = 0$ and therefore $-\iota_h(s,a) = -[\mathbb{B}_h\hat{V}_{h+1}](s,a) \leq 0$. Now, suppose $[\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a) \geq 0$. Since B' is a δ-Bellman uncertainty quantifier, we have

$$\begin{aligned}
-\iota_h(s,a) &= \hat{Q}_h(s,a) - [\mathbb{B}_h\hat{V}_{h+1}](s,a) \\
&= [\hat{\mathbb{B}}_h\hat{V}_{h+1}](s,a) - B'_h(s,a) - [\mathbb{B}_h\hat{V}_{h+1}](s,a) \\
&\leq 0.
\end{aligned}$$

This concludes that for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, with probability at least $1-\delta$, it holds that $-\iota_h(s,a) \leq 0$, and therefore

$$\text{Term II} = \sum_{h=1}^{H} \mathbb{E}\left[-\iota_h(s_h,a_h)\middle| s_1 = s, \hat{\pi}\right] \leq 0. \qquad \text{(C.12)}$$

Now, we are ready to use Lemma 6 and (C.12) to complete the proof of point 2 as follows

$$V_h^*(s) - \hat{V}_h(s) \leq F_1(s) = \max\left\{\sum_{h=1}^{H} \alpha_h \mathbb{E}\left[\bar{B}_h(s_h, a_h) | s_1 = s, \pi^*\right], \sum_{h=1}^{H} \alpha_h \mathbb{E}\left[\bar{B}_h(s_h, a_h) | s_1 = s, \pi^0\right]\right\}, \tag{C.13}$$

as desired.

C.2 Analysis of Safe-DPVI: Linear MDP

In this section, we prove the technical statements in Section 4.5.

We make use of Theorem 7, which is stated for general MDPs, to prove Theorem 8 in two steps: 1) We first state Lemma 14, in which we specify B and B' such that they are δ-safety uncertainty quantifier and δ-Bellman uncertainty quantifier as in Definition 2 for the corresponding to the linear MDP in Definition 3; 2) Next, we lower bound $\lambda_{\min}(\mathbf{\Lambda}_h)$ for each $h \in [H]$ in Lemma 15, which is followed by a high probability upper bound on $\bar{B}_h(s, a)$ for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

Lemma 14 (Theorem 2 in [2] and Lemma 5.2 in [65])**.** *Let the underlying MDP of Safe-DPVI be a linear MDP as in Definition 3. Under Assumptions 10 and 11, if we set* $B_h(s, a) = \beta \left\|\boldsymbol{\phi}(s, a)\right\|_{\mathbf{\Lambda}_h^{-1}}$ *and* $B'_h(s, a) = \beta' \left\|\boldsymbol{\phi}(s, a)\right\|_{\mathbf{\Lambda}_h^{-1}}$*, where* $\beta = \sigma\sqrt{d \log\left(\frac{2 + \frac{2K}{\lambda}}{\delta}\right)} + \sqrt{\lambda d}$ *and* $\beta' = cdH\sqrt{\log(\frac{dK}{\delta})}$ *for an absolute constant* $c > 0$*, then* B *and* B' *are* δ*-safety uncertainty quantifier and* δ*-Bellman uncertainty quantifier as in Definition 2.*

Lemma 15. *Let* $\delta \in (0, 1)$ *and Assumption 12 holds. If* $K \geq \frac{8}{\bar{c}} \log(\frac{dH}{\delta})$*, then* $\mathbb{P}\left(\lambda_{min}(\mathbf{\Lambda}_h) \geq \lambda + \frac{\bar{c}K}{2},\ \forall h \in [H]\right) \geq 1 - \delta$*.*

Proof. In order to bound the minimum eigenvalue of the Gram matrix $\mathbf{\Lambda}_h$, we use the Matrix Chernoff Inequality [125, Thm. 5.1.1].

Theorem 15 (Matrix Chernoff Inequality, [125])**.** *Consider a finite sequence* $\{\mathbf{X}_k\}$ *of independent, random, symmetric matrices in* $\mathbb{R}^{d \times d}$*. Assume that* $\lambda_{min}(\mathbf{X}_k) \geq 0$ *and* $\lambda_{max}(\mathbf{X}_k) \leq L$ *for each index* k*. Introduce the random matrix* $\mathbf{Y} = \sum_k \mathbf{X}_k$*. Let* μ_{min} *denote the minimum*

eigenvalue of the expectation $\mathbb{E}[\mathbf{Y}]$*,*

$$\mu_{min} = \lambda_{min}\left(\mathbb{E}[\mathbf{Y}]\right) = \lambda_{min}\left(\sum_k E[\mathbf{X}_k]\right).$$

Then, for any $\epsilon \in (0,1)$*, it holds,*

$$\mathbb{P}\left(\lambda_{min}(\mathbf{Y}) \leq \epsilon\mu_{min}\right) \leq d \cdot \exp\left(-(1-\epsilon)^2\frac{\mu_{min}}{2L}\right).$$

Now, let $\mathbf{X}_k = \boldsymbol{\phi}(s_h^k, a_h^k)\boldsymbol{\phi}(s_h^k, a_h^k)^\top$, such that each $\mathbf{X}_k$ is a symmetric matrix with $\lambda_{\min}(\mathbf{X}_k) \geq 0$ and $\lambda_{\max}(\mathbf{X}_k) \leq 1$ (see Assumption 11). In this notation, $\boldsymbol{\Lambda}_h = \lambda I + \sum_{k=1}^K \mathbf{X}_k$. In order to apply Theorem 15, we compute

$$\mu_{\min} := \lambda_{\min}\left(\sum_{k=1}^K \mathbb{E}[\mathbf{X}_k]\right) = \lambda_{\min}\left(\sum_{k=1}^K \mathbb{E}[\boldsymbol{\phi}(s_h^k, a_h^k)\boldsymbol{\phi}(s_h^k, a_h^k)^\top]\right) = \lambda_{\min}\left(K\Sigma_h\right) \geq \bar{c}K,$$

where the last inequity follows from Assumption 12. Thus, the theorem implies the following for any $\epsilon \in [0,1)$:

$$\mathbb{P}\left[\lambda_{\min}(\sum_{k=1}^K \mathbf{X}_k) \leq \epsilon\bar{c}K\right] \leq d \cdot \exp\left(-(1-\epsilon)^2\frac{\bar{c}K}{2}\right). \tag{C.14}$$

To complete the proof of the lemma, simply choose $\epsilon = 0.5$ (say) and $K \geq \frac{8}{\bar{c}}\log(\frac{dH}{\delta})$ in (C.14). This gives $\mathbb{P}\left(\lambda_{\min}(\boldsymbol{\Lambda}_h) \geq \lambda + \frac{\bar{c}K}{2},\ \forall h \in [H]\right) \geq 1-\delta$, as desired.

□

C.2.1 Proof of Theorem 8

As a direct conclusion of Lemma 15, we upper bound $\bar{B}_h(s,a)$ for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$. In particular, Assumption 11 and Lemma 15 imply that for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ with probability at least $1-\delta$, it holds that

$$\left\|\boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Lambda}_h^{-1}} \leq \left\|\boldsymbol{\phi}(s,a)\right\|_2 \sqrt{\frac{1}{\lambda_{\min}\left(\boldsymbol{\Lambda}_h\right)}} \leq \sqrt{\frac{2}{2\lambda + \bar{c}K}}.$$

Now that we have established B and B' and obtained an upper bound on $\left\|\boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Lambda}_h^{-1}}$, we are able to exploit the results stated in Theorem 7 to conclude that if $K \geq$

$\frac{8\beta^2}{\bar{c}\left(\tau - \max_{(s,h)\in\mathcal{S}\times[H]} \tau_h(s)\right)^2}$, which implies that

$$\begin{aligned}
\mathbb{E}_{a\sim\pi_h^0(.|s)}\left[B_h(s,a)\right] &= \mathbb{E}_{a\sim\pi_h^0(.|s)}\left[\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\boldsymbol{\Lambda}_h^{-1}}\right) \\
&\leq \beta\sqrt{\frac{2}{2\lambda+\bar{c}K}} \\
&\leq \frac{\tau - \max_{(s,h)\in\mathcal{S}\times[H]} \tau_h(s)}{2} \\
&\leq \frac{\tau - \tau_h(s)}{2} \quad \forall (s,h)\in\mathcal{S}\times[H],
\end{aligned}$$

then

$$\mathbb{P}\left(\Delta(\hat{\pi};s) \leq \frac{\sqrt{2}\bar{\beta}\sum_{h=1}^{H}\alpha_h}{\sqrt{2\lambda+\bar{c}K}},\ \forall s\in\mathcal{S} \quad \text{and} \quad \hat{\pi}\in\Pi^{\text{safe}}\right) \geq 1-3\delta. \tag{C.15}$$

C.3 Unknown $\tau_h(s)$

In this section, we relax Assumption 9, and instead assume that we only have the knowledge of a safe policy π^0, and remove the assumption on the knowledge about the costs $\tau_h(s)$.

In this case, we compute a conservative estimation of the gap $\tau - \tau_h(s)$ in an adaptive manner. We show that the agent needs N samples of each tuple $(s, a, C_h(s, a_0(s)) + \epsilon_h)$ in the dataset that are collected by executing policy π^0 in order to be able to construct this conservative estimators of the gap $\tau - \tau_h(s)$, and thereafter rely on these conservative estimates in the computation of estimated safe set of policies (discussed shortly). We show that if $\frac{16\log(K)}{(\tau-\tau_h(s))^2} \leq N \leq \frac{64\log(K)}{(\tau-\tau_h(s))^2}$, then the agent is able to construct these conservative estimates.

Let k be the number of times policy π^0 has been executed in the dataset, and $\hat{\tau}_h(s)$ be the empirical mean estimator of $\tau_h(s)$. Then, for any $\delta\in(0,1)$, we have

$$\mathbb{P}\left(\tau_h(s) \leq \hat{\tau}_h(s) + \sqrt{2\log(1/\delta)/k}\right) \geq 1-\delta. \tag{C.16}$$

If we let $\delta = 1/K^2$, then we have

$$\mathbb{P}\left(\left|\hat{\tau}_h(s) - \tau_h(s)\right| \leq 2\sqrt{\log(K)/k},\ \forall k\in[K]\right) \geq 1-2/K. \tag{C.17}$$

We start from the first sample of $(s, a, C_h(s,a) + \epsilon_h)$ and continue to update the empirical mean $\hat{\tau}_h(s)$. Let N be the first time that $\hat{\tau}_h(s) + 6\sqrt{\log(K)/N} \leq \tau$. Thus, we have

$$\tau_h(s) + 4\sqrt{\log(K)/N} \leq \tau \Rightarrow \frac{16\log(K)}{(\tau - \tau_h(s))^2} \leq N. \tag{C.18}$$

Note that in this case $4\sqrt{\log(K)/N}$ is a conservative estimation for $\tau - \tau_h(s)$. Thus, we have

$$\tau_h(s) + 4\sqrt{\log(K)/N} \leq \tau \Rightarrow \frac{16\log(K)}{(\tau - \tau_h(s))^2} \leq N. \tag{C.19}$$

Now we show that it will not take much more number of this tuple than $\frac{16\log(K)}{(\tau-\tau_h(s))^2}$ that this first time happens. Conversely, for any $N \geq \frac{64\log(K)}{(\tau-\tau_h(s))^2}$, we observe that

$$\hat{\tau}_h(s) + 6\sqrt{\log(K)/N} \leq \tau_h(s) + 8\sqrt{\log(K)/N} \leq \tau. \tag{C.20}$$

Therefore, we conclude that

$$\frac{16\log(K)}{(\tau - \tau_h(s))^2} \leq N \leq \frac{64\log(K)}{(\tau - \tau_h(s))^2}, \tag{C.21}$$

and $4\sqrt{\log(K)/N}$ is a conservative estimator for $\tau - \tau_h(s)$.

APPENDIX D

Proofs for Chapter 5

D.1 Proofs of Section 5.3

To prove Theorem 9, we will use the high probability event $\mathcal{E}_2$ defined in Lemma 17 to prove the UCB nature of Lifelong-LSVI in Lemma 18, which is the key to controlling the regret. We first state the following lemma that will be used in the proof of Lemma 17.

Lemma 16. *Under the setting of Theorem 9, let c_β be the constant in the definition of β. Then, for a fixed w, there is an absolute constant c_0 independent of c_β, such that for all $(h,k) \in [H] \times [K]$, with probability at least $1-\delta$ it holds that*

$$\left\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau \cdot \left(V_{h+1}^k(s_{h+1}^\tau, w) - \mathbb{P}_h[V_{h+1}^k(.,w)](s_h^\tau, a_h^\tau) \right) \right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} \leq c_0 H \left(d + \sqrt{d'}\right) \sqrt{\log((c_\beta+1)dd'T/\delta)},$$

where c_0 and c_β are two independent absolute constants.

Proof. We note that $\|\boldsymbol{\eta}_h\|_2 \leq \sqrt{d'}$ (Assumption 13), $\left\|\theta_h^k(w)\right\|_2 \leq H\sqrt{d}$ (Lemma 32), and $\left\| \left(\boldsymbol{\Lambda}_h^k\right)^{-1} \right\| \leq \frac{1}{\lambda}$. Thus, Lemmas 33 and 35 together imply that for all $(h,k) \in [H] \times [K]$, with probability at least $1-\delta$ it holds that

$$\begin{aligned}
&\left\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau \left(V_{h+1}^k(s_{h+1}^\tau, w) - \mathbb{P}_h[V_{h+1}^k(.,w)](s_h^\tau, a_h^\tau) \right) \right\|^2_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} \\
&\leq 4H^2 \left(\frac{d}{2} \log\left(\frac{k+\lambda}{\lambda} \right) + d' \log(1 + 4d'/\epsilon) + d \log(1 + 4Hd/\epsilon) + d^2 \log\left(\frac{1 + 8B^2\sqrt{d}}{\lambda \epsilon^2} \right) + \log\left(\frac{1}{\delta} \right) \right) \\
&+ \frac{8k^2\epsilon^2}{\lambda}.
\end{aligned}$$

If we let $\epsilon = \frac{dH}{k}$ and $\beta = c_\beta(d+\sqrt{d'})H\sqrt{\log(dT/\delta)}$, then, there exists an absolute constant $C > 0$ that is independent of c_β such that

$$\left\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}_h^\tau\left(V_{h+1}^k(s_{h+1}^\tau,w)-\mathbb{P}_h[V_{h+1}^k(.,w)](s_h^\tau,a_h^\tau)\right)\right\|^2_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} \leq C(d'+d^2)H^2\log\left((c_\beta+1)dd'T/\delta\right).$$

□

Lemma 17. *Let the setting of Theorem 9 holds. The event*

$$\mathcal{E}_2(w) \coloneqq \left\{\left\|\boldsymbol{\theta}_h^k(w)-\tilde{\boldsymbol{\theta}}_h^k(w)\right\|_{\boldsymbol{\Lambda}_h^k} \leq \beta, \forall(h,k)\in[H]\times[K]\right\}. \tag{D.1}$$

holds with probability at least $1-\delta$ for a fixed w.

Proof.

$$\begin{aligned}
\boldsymbol{\theta}_h^k(w)-\tilde{\boldsymbol{\theta}}_h^k(w) &= \boldsymbol{\theta}_h^k(w) - \left(\boldsymbol{\Lambda}_h^k\right)^{-1}\sum_{\tau=1}^{k-1}\boldsymbol{\phi}_h^\tau V_{h+1}^k(s_{h+1}^\tau,w)\\
&= \left(\boldsymbol{\Lambda}_h^k\right)^{-1}\left(\boldsymbol{\Lambda}_h^k\boldsymbol{\theta}_h^k(w) - \sum_{\tau=1}^{k-1}\boldsymbol{\phi}_h^\tau V_{h+1}^k(s_{h+1}^\tau,w)\right)\\
&= \underbrace{\lambda\left(\boldsymbol{\Lambda}_h^k\right)^{-1}\boldsymbol{\theta}_h^k(w)}_{\mathbf{q}_1} \underbrace{- \left(\boldsymbol{\Lambda}_h^k\right)^{-1}\left(\sum_{\tau=1}^{k-1}\boldsymbol{\phi}_h^\tau\left(V_{h+1}^k(s_{h+1}^\tau,w)-\mathbb{P}_h[V_{h+1}^k(.,w)](s_h^\tau,a_h^\tau)\right)\right)}_{\mathbf{q}_2}.
\end{aligned}$$

Thus, in order to upper bound $\left\|\boldsymbol{\theta}_h^k(w)-\tilde{\boldsymbol{\theta}}_h^k(w)\right\|_{\boldsymbol{\Lambda}_h^k}$, we bound $\|\mathbf{q}_1\|_{\boldsymbol{\Lambda}_h^k}$ and $\|\mathbf{q}_2\|_{\boldsymbol{\Lambda}_h^k}$ separately.

From Lemma 32, we have

$$\|\mathbf{q}_1\|_{\boldsymbol{\Lambda}_h^k} = \lambda\left\|\boldsymbol{\theta}_h^k(w)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} \leq \sqrt{\lambda}\left\|\boldsymbol{\theta}_h^k(w)\right\|_2 \leq H\sqrt{\lambda d}. \tag{D.2}$$

Thanks to Lemma 16, for all (w,h,k), with probability at least $1-\delta$, it holds that

$$\begin{aligned}
\|\mathbf{q}_2\|_{\boldsymbol{\Lambda}_h^k} &\leq \left\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}_h^\tau\left(V_{h+1}^k(s_{h+1}^\tau,w)-\mathbb{P}_h[V_{h+1}^k(.,w)](s_h^\tau,a_h^\tau)\right)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}}\\
&\leq c_0 H\left(d+\sqrt{d'}\right)\sqrt{\log((c_\beta+1)dd'T/\delta)},
\end{aligned} \tag{D.3}$$

where c_0 and c_β are two independent absolute constants.

Combining (D.2) and (D.3), for all (w, h, k), with probability at least $1 - \delta$, it holds that

$$\left\|\boldsymbol{\theta}_h^k(w) - \tilde{\boldsymbol{\theta}}_h^k(w)\right\|_{\boldsymbol{\Lambda}_h^k} \leq cH\left(d + \sqrt{d'}\right)\sqrt{\lambda \log(dd'T/\delta)}$$

for some absolute constant $c > 0$.

□

Lemma 18. *Let $\widetilde{\mathcal{W}} = \{w^1, w^2, \ldots, w^K\}$. Under the setting of Theorem 9 and conditioned on events $\{\mathcal{E}_2(w)\}_{w\in\widetilde{\mathcal{W}}}$ defined in* (D.1)*, and with Q_h^k computed as in* (5.6)*, it holds that $Q_h^k(s, a, w) \geq Q_h^*(s, a, w)$ for all $(s, a, w, h, k) \in \mathcal{S} \times \mathcal{A} \times \widetilde{\mathcal{W}} \times [H] \times [K]$.*

Proof. We first note that conditioned on events $\{\mathcal{E}_2(w)\}_{w\in\widetilde{\mathcal{W}}}$, for all $(s, a, w, h, k) \in \mathcal{S} \times \mathcal{A} \times \widetilde{\mathcal{W}} \times [H] \times [K]$, it holds that

$$\begin{aligned}
&\left|r_h(s, a, w) + \left\langle \tilde{\boldsymbol{\theta}}_h^k(w), \boldsymbol{\phi}(s, a)\right\rangle - Q_h^\pi(s, a, w) - \mathbb{P}_h\left[V_{h+1}^k(., w) - V_{h+1}^\pi(., w)\right](s, a)\right| \\
&= \left|r_h(s, a, w) + \left\langle \tilde{\boldsymbol{\theta}}_h^k(w), \boldsymbol{\phi}(s, a)\right\rangle - r_h(s, a, w) - \mathbb{P}_h\left[V_{h+1}^k(., w)\right](s, a)\right| \\
&= \left|\left\langle \tilde{\boldsymbol{\theta}}_h^k(w), \boldsymbol{\phi}(s, a)\right\rangle - \mathbb{P}_h\left[V_{h+1}^k(., w)\right](s, a)\right| \\
&= \left|\left\langle \tilde{\boldsymbol{\theta}}_h^k(w) - \boldsymbol{\theta}_h^k(w), \boldsymbol{\phi}(s, a)\right\rangle\right| \\
&\leq \left\|\tilde{\boldsymbol{\theta}}_h^k(w) - \boldsymbol{\theta}_h^k(w)\right\|_{\boldsymbol{\Lambda}_h^k} \left\|\boldsymbol{\phi}(s, a)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} \\
&\leq \beta\left\|\boldsymbol{\phi}(s, a)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}}, \qquad \text{(Lemma 17)}
\end{aligned}$$

for any policy π.

Now, we prove the lemma by induction. The statement holds for H because $Q_{H+1}^k(., ., .) = Q_{H+1}^*(., ., .) = 0$ and thus conditioned on events $\{\mathcal{E}_2(w)\}_{w\in\widetilde{\mathcal{W}}}$, defined in (D.1), for all $(s, a, w, k) \in \mathcal{S} \times \mathcal{A} \times \widetilde{\mathcal{W}} \times [K]$, we have

$$\left|r_H(s, a, w) + \left\langle \tilde{\boldsymbol{\theta}}_H^k(w), \boldsymbol{\phi}(s, a)\right\rangle - Q_H^*(s, a, w)\right| \leq \beta\left\|\boldsymbol{\phi}(s, a)\right\|_{\left(\boldsymbol{\Lambda}_H^k\right)^{-1}}.$$

Therefore, conditioned on events $\{\mathcal{E}_2(w)\}_{w\in\widetilde{\mathcal{W}}}$, for all $(s,a,w,k)\in\mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[K]$, we have

$$Q_H^*(s,a,w)\le r_H(s,a,w)+\left\langle\tilde{\boldsymbol{\theta}}_H^k(w),\boldsymbol{\phi}(s,a)\right\rangle+\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{(\boldsymbol{\Lambda}_H^k)^{-1}}=Q_H^k(s,a,w).$$

Now, suppose the statement holds at time-step $h+1$ and consider time-step h. Conditioned on events $\{\mathcal{E}_2(w)\}_{w\in\widetilde{\mathcal{W}}}$, for all $(s,a,w,h,k)\in\mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[H]\times[K]$, we have

$$\begin{aligned}0&\le r_h(s,a,w)+\left\langle\tilde{\boldsymbol{\theta}}_h^k(w),\boldsymbol{\phi}(s,a)\right\rangle-Q_h^*(s,a,w)-\mathbb{P}_h\left[V_{h+1}^k(.,w)-V_{h+1}^*(.,w)\right](s,a)+\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}}\\&\le r_h(s,a,w)+\left\langle\tilde{\boldsymbol{\theta}}_h^k(w),\boldsymbol{\phi}(s,a)\right\rangle-Q_h^*(s,a,w)+\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}}.\qquad\text{(Induction assumption)}\end{aligned}$$

Therefore, conditioned on events $\{\mathcal{E}_2(w)\}_{w\in\widetilde{\mathcal{W}}}$, for all $(s,a,w,h,k)\in\mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[H]\times[K]$, we have

$$Q_h^*(s,a,w)\le r_h(s,a,w)+\left\langle\tilde{\boldsymbol{\theta}}_h^k(w),\boldsymbol{\phi}(s,a)\right\rangle+\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}}=Q_h^k(s,a,w).$$

This completes the proof.

□

D.1.1 Proof of Theorem 9

Let $\delta_h^k=V_h^k(s_h^k,w^k)-V_h^{\pi^k}(s_h^k,w^k)$ and $\xi_{h+1}^k=\mathbb{E}\left[\delta_{h+1}^k|s_h^k,a_h^k\right]-\delta_{h+1}^k$. Conditioned on events $\{\mathcal{E}_2(w)\}_{w\in\widetilde{\mathcal{W}}}$, for all $(s,a,w,h,k)\in\mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[H]\times[K]$, we have

$$\begin{aligned}Q_h^k(s,a,w)-Q_h^{\pi^k}(s,a,w)&=r_h(s,a,w)+\left\langle\boldsymbol{\theta}_h^k(w),\boldsymbol{\phi}(s,a)\right\rangle-Q_h^{\pi^k}(s,a,w)+\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}}\\&\le\mathbb{P}_h\left[V_{h+1}^k(.,w)-V_{h+1}^{\pi^k}(.,w)\right](s,a)+2\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}}.\qquad\text{(D.4)}\end{aligned}$$

Note that $\delta_h^k\le Q_h^k(s_h^k,a_h^k,w^k)-Q_h^{\pi^k}(s_h^k,a_h^k,w^k)$. Thus, combining (D.4), Lemma 17, and a union bound over $\widetilde{\mathcal{W}}$, we conclude that for all $(h,k)\in[H]\times[K]$, with probability at least $1-\delta$, it holds that

$$\delta_h^k\le\xi_{h+1}^k+\delta_{h+1}^k+2\beta\left\|\boldsymbol{\phi}(s_h^k,a_h^k)\right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}}.$$

Now, we complete the regret analysis

$$\begin{aligned}
R_K &= \sum_{k=1}^{K} V_1^*(s_1^k, w^k) - V_1^{\pi^k}(s_1^k, w^k) \\
&\leq \sum_{k=1}^{K} V_1^k(s_1^k, w^k) - V_1^{\pi^k}(s_1^k, w^k) && \text{(Lemma 18)} \\
&= \sum_{k=1}^{K} \delta_1^k \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} \xi_h^k + 2\beta \sum_{k=1}^{K}\sum_{h=1}^{H} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} \\
&\leq 2H\sqrt{T\log(dT/\delta)} + 2H\beta\sqrt{2dK\log(1+K/\lambda)} \\
&\leq \tilde{\mathcal{O}}\left(\sqrt{\lambda(d^3+dd')H^3T}\right).
\end{aligned}$$

The third inequality is true because of the following: we observe that $\{\xi_h^k\}$ is a martingale difference sequence satisfying $|\xi_h^k| \leq 2H$. Thus, thanks to Azuma-Hoeffding inequality, we have

$$\mathbb{P}\left(\sum_{k=1}^{K}\sum_{h=1}^{H} \xi_h^k \leq 2H\sqrt{T\log(dT/\delta)}\right) \geq 1-\delta. \tag{D.5}$$

In order to bound $\sum_{k=1}^{K}\sum_{h=1}^{H}\left\|\boldsymbol{\phi}_h^k\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}}$, note that for any $h \in [H]$, we have

$$\begin{aligned}
\sum_{k=1}^{K}\left\|\boldsymbol{\phi}_h^k\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} &\leq \sqrt{K\sum_{k=1}^{K}\left\|\boldsymbol{\phi}_h^k\right\|^2_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}}} && \text{(Cauchy-Schwartz inequality)} \\
&\leq \sqrt{2K\log\left(\frac{\det\left(\boldsymbol{\Lambda}_h^K\right)}{\det\left(\boldsymbol{\Lambda}_h^1\right)}\right)} && \text{(D.6)} \\
&\leq \sqrt{2dK\log\left(1+\frac{K}{d\lambda}\right)}. && \text{(D.7)}
\end{aligned}$$

In inequality (D.6), we used the standard argument in regret analysis of linear bandits [2, Lemma 11] as follows:

$$\sum_{t=1}^{n}\min\left(\|\mathbf{y}_t\|^2_{\mathbf{V}_t^{-1}}, 1\right) \leq 2\log\frac{\det\mathbf{V}_{n+1}}{\det\mathbf{V}_1} \quad \text{where} \quad \mathbf{V}_n = \mathbf{V}_1 + \sum_{t=1}^{n-1}\mathbf{y}_t\mathbf{y}_t^\top. \tag{D.8}$$

In inequality (D.7), we used Assumption 13 and the fact that $\det(\mathbf{A}) = \prod_{i=1}^{d} \lambda_i(\mathbf{A}) \leq (\text{trace}(\mathbf{A})/d)^d$.

D.2 Proofs of Section 5.4

We start by introducing the high probability event $\mathcal{E}_1$, which is the foundation of our analysis in the following lemma.

Lemma 19. *Follow the setting of Theorem 10. The event*

$$\mathcal{E}_1(w) \coloneqq \left\{ \left\| \boldsymbol{\theta}_h^k(w) - \tilde{\boldsymbol{\theta}}_h^k(w) \right\|_{\boldsymbol{\Lambda}_h^k} \leq \beta, \forall (h,k) \in [H] \times [K] \right\}. \tag{D.9}$$

holds with probability at least $1-\delta$ for a fixed w.

D.2.1 Proof of Lemma 19

First, we state the following lemma that will be used in the proof of Lemma 19.

Lemma 20. *Under the setting of Lemma 19, let c_β be a constant in the definition of β. Then, for a fixed w, there is an absolute constant c_0 independent of c_β, such that for all $(h,k) \in [H] \times [K]$, with probability at least $1-\delta$ it holds that*

$$\left\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau \cdot \left(V_{h+1}^k(s_{h+1}^\tau, w) - \mathbb{P}_h[V_{h+1}^k(.,w)](s_h^\tau, a_h^\tau) \right) \right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}} \leq c_0 H \left(d + \sqrt{md} \right) \sqrt{\log((c_\beta+1)mdT/\delta)},$$

where c_0 and c_β are two independent absolute constants.

Proof. We note that $\left\| \boldsymbol{\eta}_h + \hat{\boldsymbol{\xi}}_h^k \right\|_2 \leq (1+H)\sqrt{md}$ and $\left\| \left(\boldsymbol{\Lambda}_h^k \right)^{-1} \right\| \leq \frac{1}{\lambda}$. Thus, Lemmas 33 and 36 together imply that for all $(h,k) \in [H] \times [K]$, with probability at least $1-\delta$ it holds that

$$\begin{aligned}
&\left\| \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau \left(V_{h+1}^k(s_{h+1}^\tau, w) - \mathbb{P}_h[V_{h+1}^k(.,w)](s_h^\tau, a_h^\tau) \right) \right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}}^2 \\
&\leq 4H^2 \left(\frac{d}{2} \log\left(\frac{k+\lambda}{\lambda} \right) + md \log(1 + 8H\sqrt{md}/\epsilon) + d^2 \log\left(\frac{1 + 32L^2\beta^2\sqrt{d}}{\lambda \epsilon^2} \right) + \log\left(\frac{1}{\delta} \right) \right) + \frac{8k^2\epsilon^2}{\lambda}.
\end{aligned}$$

If we let $\epsilon = \frac{dH}{k}$ and $\beta = c_\beta(d + \sqrt{md})H\sqrt{\log(dT/\delta)}$, then, there exists an absolute constant $C > 0$ that is independent of c_β such that

$$\left\|\sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau \left(V_{h+1}^k(s_{h+1}^\tau, w) - \mathbb{P}_h[V_{h+1}^k(.,w)](s_h^\tau, a_h^\tau)\right)\right\|^2_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} \leq C(md + d^2)H^2 \log\left((c_\beta + 1)mdT/\delta\right).$$

□

Now, we begin the formal proof of Lemma 19:

$$\begin{aligned}
\boldsymbol{\theta}_h^k(w) - \tilde{\boldsymbol{\theta}}_h^k(w) &= \boldsymbol{\theta}_h^k(w) - \left(\boldsymbol{\Lambda}_h^k\right)^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau V_{h+1}^k(s_{h+1}^\tau, w) \\
&= \left(\boldsymbol{\Lambda}_h^k\right)^{-1} \left(\boldsymbol{\Lambda}_h^k \boldsymbol{\theta}_h^k(w) - \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau V_{h+1}^k(s_{h+1}^\tau, w)\right) \\
&= \underbrace{\lambda \left(\boldsymbol{\Lambda}_h^k\right)^{-1} \boldsymbol{\theta}_h^k(w)}_{\mathbf{q}_1} \underbrace{- \left(\boldsymbol{\Lambda}_h^k\right)^{-1} \left(\sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau \left(V_{h+1}^k(s_{h+1}^\tau, w) - \mathbb{P}_h[V_{h+1}^k(.,w)](s_h^\tau, a_h^\tau)\right)\right)}_{\mathbf{q}_2}.
\end{aligned}$$

Thus, in order to upper bound $\left\|\boldsymbol{\theta}_h^k(w) - \tilde{\boldsymbol{\theta}}_h^k(w)\right\|_{\boldsymbol{\Lambda}_h^k}$, we bound $\|\mathbf{q}_1\|_{\boldsymbol{\Lambda}_h^k}$ and $\|\mathbf{q}_2\|_{\boldsymbol{\Lambda}_h^k}$ separately.

From Lemma 32, we have

$$\|\mathbf{q}_1\|_{\boldsymbol{\Lambda}_h^k} = \lambda \left\|\boldsymbol{\theta}_h^k(w)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} \leq \sqrt{\lambda} \left\|\boldsymbol{\theta}_h^k(w)\right\|_2 \leq H\sqrt{\lambda d}. \tag{D.10}$$

Thanks to Lemma 20, for all (w, h, k), with probability at least $1 - \delta$, it holds that

$$\begin{aligned}
\|\mathbf{q}_2\|_{\boldsymbol{\Lambda}_h^k} &\leq \left\|\sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau \left(V_{h+1}^k(s_{h+1}^\tau, w) - \mathbb{P}_h[V_{h+1}^k(.,w)](s_h^\tau, a_h^\tau)\right)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} \\
&\leq c_0 H \left(d + \sqrt{md}\right) \sqrt{\log((c_\beta + 1)mdT/\delta)},
\end{aligned} \tag{D.11}$$

where c_0 and c_β are two independent absolute constants.

Combining (D.10) and (D.11), for all $(h, k) \in [H] \times [K]$, with probability at least $1 - \delta$, it holds that

$$\left\|\boldsymbol{\theta}_h^k(w) - \tilde{\boldsymbol{\theta}}_h^k(w)\right\|_{\boldsymbol{\Lambda}_h^k} \leq cH\left(d + \sqrt{md}\right)\sqrt{\lambda \log(mdT/\delta)}$$

for some absolute constant $c > 0$.

D.2.2 Proof of Lemma 7

Thanks to Assumption 14 and conditioned on events $\{\mathcal{E}_1(w)\}_{w\in\widetilde{\mathcal{W}}}$, one set of solution for (5.8) is $\left\{\boldsymbol{\theta}_h^k\left(w^{(j)}\right)\right\}_{j\in[n]}$ and $\boldsymbol{\xi}_h^{V_{h+1}^k}$ with corresponding zero optimal objective value. Therefore, it holds that

$$\left\langle \hat{\boldsymbol{\theta}}_h^{k(j)}, \boldsymbol{\phi}(s,a)\right\rangle = \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}\left(s,a,w^{(j)}\right)\right\rangle, \quad \forall (j,(s,a)) \in [n]\times\mathcal{D}. \tag{D.12}$$

Let $\left(s^{(i)}, a^{(i)}\right)$ be the i-th element of $\mathcal{D}$ and $\{c_i'(s,a)\}_{i\in[d]}$ be the coefficients such that

$$\boldsymbol{\phi}(s,a) = \sum_{i\in[d]} c_i'(s,a)\boldsymbol{\phi}\left(s^{(i)}, a^{(i)}\right).$$

For any triple $(s,a,j) \in \mathcal{S}\times\mathcal{A}\times[n]$, we have

$$\begin{aligned}
\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}\left(s,a,w^{(j)}\right)\right\rangle &= \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\phi}(s,a)\otimes\boldsymbol{\rho}\left(w^{(j)}\right)\right\rangle \\
&= \left\langle \hat{\boldsymbol{\xi}}_h^k, \sum_{i\in[d]} c_i'(s,a)\boldsymbol{\phi}\left(s^{(i)},a^{(i)}\right)\otimes\boldsymbol{\rho}\left(w^{(j)}\right)\right\rangle \\
&= \sum_{i\in[d]} c_i'(s,a)\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}\left(s^{(i)},a^{(i)},w^{(j)}\right)\right\rangle && \text{(Assumption 15)} \\
&= \sum_{i\in[d]} c_i'(s,a)\left\langle \hat{\boldsymbol{\theta}}_h^{k(j)}, \boldsymbol{\phi}\left(s^{(i)},a^{(i)}\right)\right\rangle && \text{(Eqn. (D.12))} \\
&= \left\langle \hat{\boldsymbol{\theta}}_h^{k(j)}, \boldsymbol{\phi}(s,a)\right\rangle. && \text{(D.13)}
\end{aligned}$$

For any $(s,a,w) \in \mathcal{S} \times \mathcal{A} \times \mathcal{W}$, it holds that

$$\begin{aligned}
\mathbb{P}_h \left[V_{h+1}^k(.,w) \right](s,a) &= \left\langle \boldsymbol{\theta}_h^k(w), \boldsymbol{\phi}(s,a) \right\rangle && \text{(Eqn. (5.4))} \\
&= \left\langle \boldsymbol{\xi}_h^{V_{h+1}^k}, \boldsymbol{\psi}(s,a,w) \right\rangle && \text{(Assumption 14)} \\
&= \sum_{j \in [n]} c_j(w) \left\langle \boldsymbol{\xi}_h^{V_{h+1}^k}, \boldsymbol{\psi}\left(s,a,w^{(j)}\right) \right\rangle && \text{(Assumption 15)} \\
&= \sum_{j \in [n]} c_j(w) \mathbb{P}_h \left[V_{h+1}^k\left(.,w^{(j)}\right) \right](s,a) && \text{(Assumption 14)} \\
&= \sum_{j \in [n]} c_j(w) \left\langle \boldsymbol{\theta}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a) \right\rangle. && \text{(D.14)}
\end{aligned}$$

Finally, conditioned on events $\{\mathcal{E}_1(w)\}_{w \in \widetilde{\mathcal{W}}}$, for all $(s,a,w,h,k) \in \mathcal{S} \times \mathcal{A} \times \widetilde{\mathcal{W}} \times [H] \times [K]$, it holds that

$$\begin{aligned}
&\left| \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle - \mathbb{P}_h \left[V_{h+1}^k(.,w) \right](s,a) \right| \\
&= \left| \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle - \left\langle \boldsymbol{\theta}_h^k(w), \boldsymbol{\phi}(s,a) \right\rangle \right| \\
&= \left| \sum_{j \in [n]} c_j(w) \left(\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}\left(s,a,w^{(j)}\right) \right\rangle - \left\langle \boldsymbol{\theta}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a) \right\rangle \right) \right| \\
&\qquad \text{(Assumption 15 and Eqn. (D.14))} \\
&\leq \left| \sum_{j \in [n]} c_j(w) \left(\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}\left(s,a,w^{(j)}\right) \right\rangle - \left\langle \hat{\boldsymbol{\theta}}_h^{k(j)}, \boldsymbol{\phi}(s,a) \right\rangle \right) \right| \\
&+ \left| \sum_{j \in [n]} c_j(w) \left\langle \hat{\boldsymbol{\theta}}_h^{k(j)} - \tilde{\boldsymbol{\theta}}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a) \right\rangle \right| + \left| \sum_{j \in [n]} c_j(w) \left\langle \tilde{\boldsymbol{\theta}}_h^k\left(w^{(j)}\right) - \boldsymbol{\theta}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a) \right\rangle \right| \\
&= \left| \sum_{j \in [n]} c_j(w) \left\langle \hat{\boldsymbol{\theta}}_h^{k(j)} - \tilde{\boldsymbol{\theta}}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a) \right\rangle \right| + \left| \sum_{j \in [n]} c_j(w) \left\langle \tilde{\boldsymbol{\theta}}_h^k\left(w^{(j)}\right) - \boldsymbol{\theta}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a) \right\rangle \right| \\
&\qquad \text{(Eqn. (D.13))} \\
&\leq 2L\beta \left\| \boldsymbol{\phi}(s,a) \right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}}. && \text{(Lemma 19)}
\end{aligned}$$

D.2.3 Proof of Optimistic Nature of UCBlvd

Lemma 21. *Let $\widetilde{\mathcal{W}} = \{w^\tau : \tau \in [K]\} \cup \{w^{(j)} : j \in [n]\}$. Under the setting of Theorem 10 and conditioned on events $\{\mathcal{E}_1(w)\}_{w\in\widetilde{\mathcal{W}}}$ defined in (5.9), and with Q_h^k computed as in (5.7), it holds that $Q_h^k(s,a,w) \geq Q_h^*(s,a,w)$ for all $(s,a,w,h,k) \in \mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[H]\times[K]$.*

Proof. We first note that conditioned on events $\{\mathcal{E}_1(w)\}_{w\in\widetilde{\mathcal{W}}}$, for all $(s,a,w,h,k) \in \mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[H]\times[K]$, it holds that

$$\begin{aligned}
&\left|r_h(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle - Q_h^\pi(s,a,w) - \mathbb{P}_h\left[V_{h+1}^k(.,w) - V_{h+1}^\pi(.,w)\right](s,a)\right| \\
&= \left|r_h(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle - r_h(s,a,w) - \mathbb{P}_h\left[V_{h+1}^k(.,w)\right](s,a)\right| \\
&= \left|\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle - \mathbb{P}_h\left[V_{h+1}^k(.,w)\right](s,a)\right| \\
&\leq 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}}, \qquad \text{(Lemma 7)}
\end{aligned}$$

for any policy π.

Now, we prove the lemma by induction. The statement holds for H because $Q_{H+1}^k(.,.,.) = Q_{H+1}^*(.,.,.) = 0$ and thus conditioned events $\{\mathcal{E}_1(w)\}_{w\in\widetilde{\mathcal{W}}}$, defined in (5.9), for all $(s,a,w,k) \in \mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[K]$, we have

$$\left|r_H(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_H^k, \boldsymbol{\psi}(s,a,w)\right\rangle - Q_H^*(s,a,w)\right| \leq 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{(\boldsymbol{\Lambda}_H^k)^{-1}}.$$

Therefore, conditioned on events $\{\mathcal{E}_1(w)\}_{w\in\widetilde{\mathcal{W}}}$, for all $(s,a,w,k) \in \mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[K]$, we have

$$\begin{aligned}
Q_H^*(s,a,w) &\leq r_H(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_H^k, \boldsymbol{\psi}(s,a,w)\right\rangle + 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{(\boldsymbol{\Lambda}_H^k)^{-1}} \\
&= \left\{r_H(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_H^k, \boldsymbol{\psi}(s,a,w)\right\rangle + 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{(\boldsymbol{\Lambda}_H^k)^{-1}}\right\}^+ \\
&= Q_H^k(s,a,w),
\end{aligned}$$

where the first equality follows from the fact that $Q_H^*(s,a,w) \geq 0$. Now, suppose the statement holds at time-step $h+1$ and consider time-step h. Conditioned on events $\{\mathcal{E}_1(w)\}_{w\in\widetilde{\mathcal{W}}}$, for

all $(s,a,w,h,k) \in \mathcal{S} \times \mathcal{A} \times \widetilde{\mathcal{W}} \times [H] \times [K]$, we have

$$0 \leq r_h(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle - Q_h^*(s,a,w) - \mathbb{P}_h \left[V_{h+1}^k(.,w) - V_{h+1}^*(.,w) \right](s,a) + 2L\beta \left\| \boldsymbol{\phi}(s,a) \right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}}$$
$$\leq r_h(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle - Q_h^*(s,a,w) + 2L\beta \left\| \boldsymbol{\phi}(s,a) \right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}}. \qquad \text{(Induction assumption)}$$

Therefore, conditioned on events $\{\mathcal{E}_1(w)\}_{w \in \widetilde{\mathcal{W}}}$, for all $(s,a,w,h,k) \in \mathcal{S} \times \mathcal{A} \times \widetilde{\mathcal{W}} \times [H] \times [K]$, we have

$$\begin{aligned} Q_h^*(s,a,w) &\leq r_h(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle + 2L\beta \left\| \boldsymbol{\phi}(s,a) \right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}} \\ &= \left\{ r_h(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle + 2L\beta \left\| \boldsymbol{\phi}(s,a) \right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}} \right\}^+ \\ &= Q_h^k(s,a,w), \end{aligned}$$

where the first equality follows from the fact that $Q_h^*(s,a,w) \geq 0$. This completes the proof. □

D.2.4 Proof of Theorem 10

First, we bound the number of times Algorithm 7 updates $\hat{\boldsymbol{\xi}}_h^k$, i.e., number of planning calls. Let P be the total number of updates and k_p be the episode at which, the agent did replanning for the p-th time. Note that $\det \boldsymbol{\Lambda}_h^1 = \lambda^d$ and $\det \boldsymbol{\Lambda}_h^K \leq \text{trace}(\boldsymbol{\Lambda}_h^K/d)^d \leq \left(\lambda + \frac{K}{d}\right)^d$, and consequently:

$$\frac{\det \boldsymbol{\Lambda}_h^K}{\det \boldsymbol{\Lambda}_h^1} = \prod_{p=1}^{P} \frac{\det \boldsymbol{\Lambda}_h^{k_p}}{\det \boldsymbol{\Lambda}_h^{k_{p-1}}} \leq \left(1 + \frac{K}{d\lambda}\right)^d,$$

and therefore

$$\prod_{h=1}^{H} \frac{\det \boldsymbol{\Lambda}_h^K}{\det \boldsymbol{\Lambda}_h^1} = \prod_{h=1}^{H} \prod_{p=1}^{P} \frac{\det \boldsymbol{\Lambda}_h^{k_p}}{\det \boldsymbol{\Lambda}_h^{k_{p-1}}} \leq \left(1 + \frac{K}{d\lambda}\right)^{dH}. \tag{D.15}$$

Since $1 \leq \frac{\det \boldsymbol{\Lambda}_h^{k_p}}{\det \boldsymbol{\Lambda}_h^{k_{p-1}}}$ for all $p \in [P]$, we can deduce from (D.15) that

$$\exists h \in [H] \quad \text{such that} \quad e < \frac{\det \boldsymbol{\Lambda}_h^k}{\det \boldsymbol{\Lambda}_h^{\tilde{k}}}$$

happens for at most $dH \log\left(1+\frac{K}{d\lambda}\right)$ number of episodes $k \in [K]$. This concludes that the number of planing calls in UCBlvd is $dH \log\left(1+\frac{K}{d\lambda}\right)$.

Now, we prove the regret bound. Let $\delta_h^k = V_h^{\tilde{k}}(s_h^k, w^k) - V_h^{\pi^k}(s_h^k, w^k)$ and $\xi_{h+1}^k = \mathbb{E}\left[\delta_{h+1}^k | s_h^k, a_h^k\right] - \delta_{h+1}^k$. Conditioned on events $\{\mathcal{E}_1(w)\}_{w\in\widetilde{\mathcal{W}}}$, for all $(s,a,w,h,k) \in \mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[H]\times[K]$, we have

$$\begin{aligned} Q_h^{\tilde{k}}(s,a,w) - Q_h^{\pi^k}(s,a,w) &= r_h(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_h^{\tilde{k}}, \boldsymbol{\psi}(s,a,w)\right\rangle - Q_h^{\pi^k}(s,a,w) + 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{(\boldsymbol{\Lambda}_h^{\tilde{k}})^{-1}} \\ &\leq \mathbb{P}_h\left[V_{h+1}^{\tilde{k}}(.,w) - V_{h+1}^{\pi^k}(.,w)\right](s,a) + 4L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{(\boldsymbol{\Lambda}_h^{\tilde{k}})^{-1}}. \end{aligned} \tag{D.16}$$

Note that $\delta_h^k \leq Q_h^{\tilde{k}}(s_h^k, a_h^k, w^k) - Q_h^{\pi^k}(s_h^k, a_h^k, w^k)$. Thus, combining (D.16), Lemma 19, and a union bound over $\widetilde{\mathcal{W}}$, we conclude that for all $(h,k)\in[H]\times[K]$, with probability at least $1-\delta$, it holds that gives

$$\delta_h^k \leq \xi_{h+1}^k + \delta_{h+1}^k + 4L\beta\left\|\boldsymbol{\phi}(s_h^k, a_h^k)\right\|_{(\boldsymbol{\Lambda}_h^{\tilde{k}})^{-1}}.$$

Note that for any positive semi-definite matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ such that $\mathbf{A} = \mathbf{B} + \mathbf{C}$, we have:

$$\det(\mathbf{A}) \geq \det(\mathbf{B}),\ \det(\mathbf{A}) \geq \det(\mathbf{C}), \tag{D.17}$$

and for any $\mathbf{x} \neq 0$ ([2, Lemm. 12]):

$$\frac{\|\mathbf{x}\|_{\mathbf{A}}^2}{\|\mathbf{x}\|_{\mathbf{B}}^2} \leq \frac{\det(\mathbf{A})}{\det(\mathbf{B})} \quad \text{and} \quad \frac{\|\mathbf{x}\|_{\mathbf{B}^{-1}}^2}{\|\mathbf{x}\|_{\mathbf{A}^{-1}}^2} \leq \frac{\det(\mathbf{A})}{\det(\mathbf{B})}. \tag{D.18}$$

Now, we complete the regret analysis following similar steps as those of Theorem 9's

proof:

$$
\begin{aligned}
R_K &= \sum_{k=1}^{K} V_1^*(s_1^k, w^k) - V_1^{\pi^k}(s_1^k, w^k) \\
&\le \sum_{k=1}^{K} V_1^{\tilde{k}}(s_1^k, w^k) - V_1^{\pi^k}(s_1^k, w^k) && \text{(Lemma 21)} \\
&= \sum_{k=1}^{K} \delta_1^k \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H} \xi_h^k + 4L\beta \sum_{k=1}^{K}\sum_{h=1}^{H} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\left(\boldsymbol{\Lambda}_h^{\tilde{k}}\right)^{-1}} \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H} \xi_h^k + 4L\beta \sum_{k=1}^{K}\sum_{h=1}^{H} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\left(\boldsymbol{\Lambda}_h^{k}\right)^{-1}} \sqrt{\frac{\det \boldsymbol{\Lambda}_h^k}{\det \boldsymbol{\Lambda}_h^{\tilde{k}}}} && \text{(Eqn. (D.18))} \\
&\le 2H\sqrt{T\log(dT/\delta)} + 8HL\beta\sqrt{2dK\log(1+K/\lambda)} \\
&\le \tilde{\mathcal{O}}\left(L\sqrt{\lambda(d^3+md^2)H^3T}\right).
\end{aligned}
$$

D.2.5 Discussion on the Time Complexity of UCBlvd and Lifelong-LSVI

In what follows, we clarify on how the time complexity of UCBlvd compares to that of Lifelong LSVI. When we compute $\left(\boldsymbol{\Lambda}_h^k\right)^{-1}$ by the Sherman-Morrison formula, the computational complexity of Lifelong-LSVI is dominated by Line 5 in computing $\max_{a\in\mathcal{A}} Q_{h+1}^k(s_{h+1}^\tau, a)$ for all $\tau \in [k]$. This takes $\mathcal{O}(d^2|\mathcal{A}|K)$ per step, which gives a total runtime $\mathcal{O}(d^2|\mathcal{A}|HK^2)$. In UCBlvd, every planning call takes $\tilde{\mathcal{O}}(md^2|\mathcal{A}|K + m^3d^3)$, where the second term is the time-complexity of the convex QCQP with $m+1$ constraints and $2md$ variables. This gives a total runtime of $\tilde{\mathcal{O}}(H^2(md^3|\mathcal{A}|K + m^3d^4))$. Therefore, UCBlvd enjoys a smaller time complexity by a factor of K compared to that of Lifelong-LSVI, which is a significant reduction in practical scenarios where $K >> d' = md$.

D.3 Details of Remark 2: UCBlvd with unknown rewards

In order for our analysis to go through, we need a slightly different completeness assumption as below:

Assumption 19. *Given feature maps $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\boldsymbol{\psi} : \mathcal{S} \times \mathcal{A} \times \mathcal{W} \to \mathbb{R}^{d'}$, consider function class*

$$\mathcal{F} = \Bigg\{ f : f(s,w) = \min\Bigg\{ \max_{a \in \mathcal{A}} \Big\{ \langle \boldsymbol{\nu}, \boldsymbol{\psi}(s,a,w) \rangle + \beta \big\| \boldsymbol{\phi}(s,a) \big\|_{\boldsymbol{\Lambda}^{-1}} + \tilde{\beta} \big\| \boldsymbol{\psi}(s,a,w) \big\|_{\tilde{\boldsymbol{\Lambda}}^{-1}} \Big\}^+, H \Bigg\}$$
$$, \boldsymbol{\nu} \in \mathbb{R}^{d'}, \boldsymbol{\Lambda} \in \mathbf{S}^d_{++}, \tilde{\boldsymbol{\Lambda}} \in \mathbf{S}^{d'}_{++}, \beta \geq 0, \tilde{\beta} \geq 0 \Big\}.$$

Then for any $f \in \mathcal{F}$, and $h \in [H]$, there exists a vector $\boldsymbol{\xi}^f_h \in \mathbb{R}^{d'}$ with $\left\| \boldsymbol{\xi}^f_h \right\| \leq H\sqrt{d'}$ such that

$$\mathbb{P}_h \left[f(.,w) \right](s,a) = \langle \boldsymbol{\xi}^f_h, \boldsymbol{\psi}(s,a,w) \rangle.$$

D.3.1 Overview

Algorithm 9 UCBlvd with Unknown Rewards

Set: $Q^k_{H+1}(.,.,.) = 0,\ \forall k \in [K],\ \tilde{k} = 1$
for episodes $k = 1, \ldots, K$ **do**
  Observe the initial state s^k_1 and the task context w^k.
  if $\exists h \in [H]$ such that $\frac{\det \boldsymbol{\Lambda}^k_h}{\det \boldsymbol{\Lambda}^{\tilde{k}}_h} > e$ or $\frac{\det \tilde{\boldsymbol{\Lambda}}^k_h}{\det \tilde{\boldsymbol{\Lambda}}^{\tilde{k}}_h} > e$ **then**
    $\tilde{k} = k$
    for time-steps $h = H, \ldots, 1$ **do**
      Compute $\hat{\boldsymbol{\xi}}^k_h$ as in (D.21).
    end for
  end if
  for time-steps $h = 1, \ldots, H$ **do**
    Compute $Q^{\tilde{k}}_h(s^k_h, a, w^k)$ for all $a \in \mathcal{A}$ as in (D.19).
    Play $a^k_h = \arg\max_{a \in \mathcal{A}} Q^{\tilde{k}}_h(s^k_h, a, w^k)$ and observe s^k_{h+1} and r^k_h.
  end for
end for

Let $\boldsymbol{\psi}^\tau_h = \boldsymbol{\psi}(s^\tau_h, a^\tau_h, w^\tau)$. UCBlvd with unknown rewards works with the following action-value functions:

$$Q^k_h(s,a,w) = \left\{ \left\langle \tilde{\boldsymbol{\eta}}^k_h + \hat{\boldsymbol{\xi}}^k_h, \boldsymbol{\psi}(s,a,w) \right\rangle + \beta \big\| \boldsymbol{\phi}(s,a) \big\|_{(\boldsymbol{\Lambda}^k_h)^{-1}} + \tilde{\beta} \big\| \boldsymbol{\psi}(s,a,w) \big\|_{(\tilde{\boldsymbol{\Lambda}}^k_h)^{-1}} \right\}^+, \quad \text{(D.19)}$$

where

$$\tilde{\boldsymbol{\eta}}_h^k = \left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\psi}_h^\tau . r_h^\tau \quad \text{and} \quad \tilde{\boldsymbol{\Lambda}}_h^k = \lambda \mathbf{I}_{md} + \sum_{\tau=1}^{k-1} \boldsymbol{\psi}_h^\tau \boldsymbol{\psi}_h^{\tau\top}, \tag{D.20}$$

and

$$\begin{aligned} \hat{\boldsymbol{\xi}}_h^k, \left\{\hat{\boldsymbol{\theta}}_h^{k(j)}\right\}_{j\in[n]} = \underset{\boldsymbol{\xi}, \left\{\boldsymbol{\theta}^{(j)}\right\}_{j\in[n]}}{\arg\min} & \sum_{j\in[n]} \sum_{(s,a)\in\mathcal{D}} \left(\left\langle \boldsymbol{\theta}^{(j)}, \boldsymbol{\phi}(s,a)\right\rangle - \left\langle \boldsymbol{\xi}, \boldsymbol{\psi}\left(s,a,w^{(j)}\right)\right\rangle\right)^2 \\ \text{s.t.} & \left\|\boldsymbol{\theta}^{(j)} - \tilde{\boldsymbol{\theta}}_h^k\left(w^{(j)}\right)\right\|_{\boldsymbol{\Lambda}_h^k} \le \beta, \ \forall j \in [n] \quad \text{and} \quad \|\boldsymbol{\xi}\|_2 \le H\sqrt{md}, \end{aligned} \tag{D.21}$$

$\mathcal{D} = \left\{(s,a) : \boldsymbol{\phi}(s,a) \text{ are } d \text{ linearly independent vectors.}\right\}$, and $\tilde{\boldsymbol{\theta}}_h^k(w)$ and $\boldsymbol{\Lambda}_h^k$ are defined in (5.5).

We note that compared to (5.7), action-value function defined in (D.19) involves an extra term $\left\langle \tilde{\boldsymbol{\eta}}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle + \tilde{\beta}\left\|\boldsymbol{\psi}(s,a,w)\right\|_{(\tilde{\boldsymbol{\Lambda}}_h^k)^{-1}}$. This term is in fact an upper bound on $r_h(s,a,w)$. Specifically, from Theorem 2 in [2], we know that for $\tilde{\beta} = \sqrt{\lambda md}$, it holds that

$$\left\|\boldsymbol{\eta}_h - \tilde{\boldsymbol{\eta}}_h^k\right\|_{\tilde{\boldsymbol{\Lambda}}_h^k} \le \tilde{\beta}, \ \forall (h,k) \in [H]\times[K]. \tag{D.22}$$

Theorem 16. *Let $T = KH$. Under Assumptions 13, 15, and 19, the number of planning calls in Algorithm 9 is at most $dH\log\left(1+\frac{K}{d\lambda}\right) + mdH\log\left(1+\frac{K}{md\lambda}\right)$, and there exists an absolute constant $c>0$ such that for any fixed $\delta \in (0, 0.5)$, if we set $\lambda = 1$, $\beta = cH\,(md)\sqrt{\log(mdT/\delta)}$ and $\tilde{\beta} = \sqrt{md}$ in Algorithm 9, then with probability at least $1-2\delta$, it holds that*

$$\begin{aligned} R_K &\le 2H\sqrt{T\log(dT/\delta)} + 4H\sqrt{K}\left(L\beta\sqrt{2d\log(1+K/\lambda)} + \tilde{\beta}\sqrt{2md\log(1+K/\lambda)}\right) \\ &\le \tilde{\mathcal{O}}\left(L\sqrt{m^2d^3H^3T}\right). \end{aligned}$$

D.3.2 Necessary Analysis for the Proof of Theorem 16

Lemma 22. *Let c_β be a constant in the definition of β. Then, under Assumptions 13, 15, and 19, for a fixed w, there is an absolute constant c_0 independent of c_β, such that for all $(h,k) \in [H]\times[K]$, with probability at least $1-\delta$ it holds that*

$$\left\|\sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau \cdot \left(V_{h+1}^k(s_{h+1}^\tau, w) - \mathbb{P}_h[V_{h+1}^k(.,w)](s_h^\tau, a_h^\tau)\right)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} \le c_0 mdH\sqrt{\log((c_\beta+1)mdT/\delta)},$$

where c_0 and c_β are two independent absolute constants.

Proof. We note that $\left\|\tilde{\boldsymbol{\eta}}_h^k + \hat{\boldsymbol{\xi}}_h^k\right\|_2 \leq H\sqrt{md} + K/\lambda$ and $\left\|\left(\boldsymbol{\Lambda}_h^k\right)^{-1}\right\| \leq \frac{1}{\lambda}$ and $\left\|\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}\right\| \leq \frac{1}{\lambda}$. Thus, Lemmas 33 and 37 together imply that for all $(h,k) \in [H] \times [K]$, with probability at least $1-\delta$ it holds that

$$\begin{aligned}
&\left\|\sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau \left(V_{h+1}^k(s_{h+1}^\tau, w) - \mathbb{P}_h[V_{h+1}^k(.,w)](s_h^\tau, a_h^\tau)\right)\right\|^2_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} \\
&\leq 4H^2 \left(\frac{d}{2}\log\left(\frac{k+\lambda}{\lambda}\right) + md\log(1+8H\sqrt{md}/\epsilon) + d^2\log\left(\frac{1+32L^2\beta^2\sqrt{d}}{\lambda\epsilon^2}\right)\right. \\
&\left. + m^2d^2\log\left(\frac{1+8\tilde{\beta}^2\sqrt{md}}{\lambda\epsilon^2}\right) + \log\left(\frac{1}{\delta}\right)\right) + \frac{8k^2\epsilon^2}{\lambda}.
\end{aligned}$$

If we let $\epsilon = \frac{dH}{k}$ and $\beta = c_\beta(md)H\sqrt{\log(mdT/\delta)}$, then, there exists an absolute constant $C > 0$ that is independent of c_β such that

$$\left\|\sum_{\tau=1}^{k-1} \boldsymbol{\phi}_h^\tau \left(V_{h+1}^k(s_{h+1}^\tau, w) - \mathbb{P}_h[V_{h+1}^k(.,w)](s_h^\tau, a_h^\tau)\right)\right\|^2_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} \leq C(m^2d^2)H^2\log\left((c_\beta+1)mdT/\delta\right).$$

□

Lemma 23. *Under Assumptions 13, 15, and 19, if we let $\beta = cmdH\sqrt{\lambda\log(mdT/\delta)}$ with an absolute constant $c > 0$, then the event*

$$\mathcal{E}_3(w) \coloneqq \left\{\left\|\boldsymbol{\theta}_h^k(w) - \tilde{\boldsymbol{\theta}}_h^k(w)\right\|_{\boldsymbol{\Lambda}_h^k} \leq \beta,\ \forall (h,k) \in [H] \times [K]\right\}. \tag{D.23}$$

holds with probability at least $1-\delta$ for a fixed w.

Proof. The proof follows the same steps as those of Lemma 19, except that it uses Lemma 22 instead of Lemma 20 due to different structure of action-value functions Q_h^k in this section. □

Lemma 24. *Let $\widetilde{\mathcal{W}} = \{w^\tau : \tau \in [K]\} \cup \{w^{(j)} : j \in [n]\}$. Under the setting of Theorem 16 and conditioned on events $\{\mathcal{E}_3(w)\}_{w\in\widetilde{\mathcal{W}}}$ defined in* (D.23)*, for all $(s,a,w,h,k) \in \mathcal{S} \times \mathcal{A} \times \widetilde{\mathcal{W}} \times [H] \times [K]$, it holds that*

$$\left|\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle - \mathbb{P}_h\left[V_{h+1}^k(.,w)\right](s,a)\right| \leq 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}}.$$

Proof. The proof follows the exact same steps as those of Lemma 7's proof. □

Lemma 25. *Let* $\widetilde{\mathcal{W}} = \{w^\tau : \tau \in [K]\} \cup \{w^{(j)} : j \in [n]\}$. *Under the setting of Theorem 16 and conditioned on events* $\{\mathcal{E}_3(w)\}_{w\in\widetilde{\mathcal{W}}}$ *defined in* (D.23)*, and with* Q_h^k *computed as in* (D.19)*, it holds that* $Q_h^k(s,a,w) \geq Q_h^*(s,a,w)$ *for all* $(s,a,w,h,k) \in \mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[H]\times[K]$.

Proof. We first note that conditioned on events $\{\mathcal{E}_3(w)\}_{w\in\widetilde{\mathcal{W}}}$, for all $(s,a,w,h,k) \in \mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[H]\times[K]$, it holds that

$$
\begin{aligned}
&\left|\left\langle \tilde{\boldsymbol{\eta}}_h^k + \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle - Q_h^\pi(s,a,w) - \mathbb{P}_h\left[V_{h+1}^k(.,w) - V_{h+1}^\pi(.,w)\right](s,a)\right| \\
&= \left|\left\langle \tilde{\boldsymbol{\eta}}_h^k + \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle - r_h(s,a,w) - \mathbb{P}_h\left[V_{h+1}^k(.,w)\right](s,a)\right| \\
&\leq \left|\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle - \mathbb{P}_h\left[V_{h+1}^k(.,w)\right](s,a)\right| + \tilde{\beta}\left\|\boldsymbol{\psi}(s,a,w)\right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}} \qquad \text{(Eqn. (D.22))} \\
&\leq 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} + \tilde{\beta}\left\|\boldsymbol{\psi}(s,a,w)\right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}}, \qquad \text{(Lemma 24)}
\end{aligned}
$$

for any policy π.

Now, we prove the lemma by induction. The statement holds for H because $Q_{H+1}^k(.,.,.) = Q_{H+1}^*(.,.,.) = 0$ and thus conditioned events $\{\mathcal{E}_3(w)\}_{w\in\widetilde{\mathcal{W}}}$, defined in (D.23), for all $(s,a,w,k) \in \mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[K]$, we have

$$
\left|\left\langle \tilde{\boldsymbol{\eta}}_H^k + \hat{\boldsymbol{\xi}}_H^k, \boldsymbol{\psi}(s,a,w)\right\rangle - Q_H^*(s,a,w)\right| \leq 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\boldsymbol{\Lambda}_H^k\right)^{-1}} + \tilde{\beta}\left\|\boldsymbol{\psi}(s,a,w)\right\|_{\left(\tilde{\boldsymbol{\Lambda}}_H^k\right)^{-1}}. \tag{D.24}
$$

Therefore, conditioned on events $\{\mathcal{E}_3(w)\}_{w\in\widetilde{\mathcal{W}}}$, for all $(s,a,w,k) \in \mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[K]$, we have

$$
\begin{aligned}
Q_H^*(s,a,w) &\leq \left\langle \tilde{\boldsymbol{\eta}}_H^k + \hat{\boldsymbol{\xi}}_H^k, \boldsymbol{\psi}(s,a,w)\right\rangle + 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\boldsymbol{\Lambda}_H^k\right)^{-1}} + \tilde{\beta}\left\|\boldsymbol{\psi}(s,a,w)\right\|_{\left(\tilde{\boldsymbol{\Lambda}}_H^k\right)^{-1}} \\
&= \left\{\left\langle \tilde{\boldsymbol{\eta}}_H^k + \hat{\boldsymbol{\xi}}_H^k, \boldsymbol{\psi}(s,a,w)\right\rangle + 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\boldsymbol{\Lambda}_H^k\right)^{-1}} + \tilde{\beta}\left\|\boldsymbol{\psi}(s,a,w)\right\|_{\left(\tilde{\boldsymbol{\Lambda}}_H^k\right)^{-1}}\right\}^+ \\
&= Q_H^k(s,a,w),
\end{aligned}
$$

where the first equality follows from the fact that $Q^*_H(s,a,w) \geq 0$. Now, suppose the statement holds at time-step $h+1$ and consider time-step h. Conditioned on events $\{\mathcal{E}_3(w)\}_{w\in\widetilde{\mathcal{W}}}$, for all $(s,a,w,h,k) \in \mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[H]\times[K]$, we have

$$\begin{aligned}
0 &\leq \left\langle \tilde{\boldsymbol{\eta}}_h^k + \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle - Q^*_h(s,a,w) - \mathbb{P}_h\left[V^k_{h+1}(.,w) - V^*_{h+1}(.,w)\right](s,a) \\
&+ 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} + \tilde{\beta}\left\|\boldsymbol{\psi}(s,a,w)\right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}} \\
&\leq \left\langle \tilde{\boldsymbol{\eta}}_h^k + \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle - Q^*_h(s,a,w) + 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} + \tilde{\beta}\left\|\boldsymbol{\psi}(s,a,w)\right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}}.
\end{aligned}$$

(Induction assumption)

Therefore, conditioned on events $\{\mathcal{E}_3(w)\}_{w\in\widetilde{\mathcal{W}}}$, for all $(s,a,w,h,k) \in \mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[H]\times[K]$, we have

$$\begin{aligned}
Q^*_h(s,a,w) &\leq \left\langle \tilde{\boldsymbol{\eta}}_h^k + \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle + 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} + \tilde{\beta}\left\|\boldsymbol{\psi}(s,a,w)\right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}} \\
&= \left\{\left\langle \tilde{\boldsymbol{\eta}}_h^k + \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle + 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} + \tilde{\beta}\left\|\boldsymbol{\psi}(s,a,w)\right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}}\right\}^+ \\
&= Q^k_h(s,a,w),
\end{aligned}$$

where the first equality follows from the fact that $Q^*_h(s,a,w) \geq 0$. This completes the proof. □

D.3.3 Proof of Theorem 16

First, we bound the number of times Algorithm 9 updates $\hat{\boldsymbol{\xi}}_h^k$, i.e., number of planning calls. Let P be the total number of policy updates and k_p be the episode at, the agent did replanning for the p-th time. Note that $\det \boldsymbol{\Lambda}_h^1 = \lambda^d$ and $\det \boldsymbol{\Lambda}_h^K \leq \operatorname{trace}(\boldsymbol{\Lambda}_h^K/d)^d \leq \left(\lambda + \frac{K}{d}\right)^d$, and consequently:

$$\frac{\det \boldsymbol{\Lambda}_h^K}{\det \boldsymbol{\Lambda}_h^1} = \prod_{p=1}^{P} \frac{\det \boldsymbol{\Lambda}_h^{k_p}}{\det \boldsymbol{\Lambda}_h^{k_{p-1}}} \leq \left(1 + \frac{K}{d\lambda}\right)^d,$$

and therefore

$$\prod_{h=1}^{H} \frac{\det \boldsymbol{\Lambda}_h^K}{\det \boldsymbol{\Lambda}_h^1} = \prod_{h=1}^{H}\prod_{p=1}^{P} \frac{\det \boldsymbol{\Lambda}_h^{k_p}}{\det \boldsymbol{\Lambda}_h^{k_{p-1}}} \leq \left(1 + \frac{K}{d\lambda}\right)^{dH}. \tag{D.25}$$

We similarly have

$$\prod_{h=1}^{H} \frac{\det \tilde{\mathbf{\Lambda}}_h^{K}}{\det \tilde{\mathbf{\Lambda}}_h^{1}} = \prod_{h=1}^{H} \prod_{p=1}^{P} \frac{\det \tilde{\mathbf{\Lambda}}_h^{k_p}}{\det \tilde{\mathbf{\Lambda}}_h^{k_{p-1}}} \leq \left(1 + \frac{K}{md\lambda}\right)^{mdH}. \tag{D.26}$$

Since $1 \leq \frac{\det \mathbf{\Lambda}_h^{k_p}}{\det \mathbf{\Lambda}_h^{k_{p-1}}}$ for all $p \in [P]$, we can deduce from (D.25) and (D.26) that

$$\exists h \in [H] \quad \text{such that} \quad e < \frac{\det \mathbf{\Lambda}_h^{k}}{\det \mathbf{\Lambda}_h^{\tilde{k}}} \quad \text{or} \quad e < \frac{\det \tilde{\mathbf{\Lambda}}_h^{k}}{\det \tilde{\mathbf{\Lambda}}_h^{\tilde{k}}} \tag{D.27}$$

happens for at most $dH \log\left(1 + \frac{K}{d\lambda}\right) + mdH \log\left(1 + \frac{K}{md\lambda}\right)$ number of episodes $k \in [K]$. This concludes that number of planning calls in Algorithm 9 is at most $dH \log\left(1 + \frac{K}{d\lambda}\right) + mdH \log\left(1 + \frac{K}{md\lambda}\right)$.

Now, we prove the regret bound. Let $\delta_h^k = V_h^{\tilde{k}}(s_h^k, w^k) - V_h^{\pi^k}(s_h^k, w^k)$ and $\xi_{h+1}^k = \mathbb{E}\left[\delta_{h+1}^k | s_h^k, a_h^k\right] - \delta_{h+1}^k$. Conditioned on events $\{\mathcal{E}_3(w)\}_{w \in \widetilde{\mathcal{W}}}$, for all $(s, a, w, h, k) \in \mathcal{S} \times \mathcal{A} \times \widetilde{\mathcal{W}} \times [H] \times [K]$, we have

$$\begin{aligned}
&Q_h^{\tilde{k}}(s, a, w) - Q_h^{\pi^k}(s, a, w) \\
&= \left\langle \tilde{\boldsymbol{\eta}}_h^{\tilde{k}} + \hat{\boldsymbol{\xi}}_h^{\tilde{k}}, \boldsymbol{\psi}(s, a, w) \right\rangle - Q_h^{\pi^k}(s, a, w) + 2L\beta \left\| \boldsymbol{\phi}(s, a) \right\|_{(\mathbf{\Lambda}_h^{\tilde{k}})^{-1}} + \tilde{\beta} \left\| \boldsymbol{\psi}(s, a, w) \right\|_{(\tilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1}} \\
&\leq \mathbb{P}_h \left[V_{h+1}^{\tilde{k}}(., w) - V_{h+1}^{\pi^k}(., w) \right](s, a) + 4L\beta \left\| \boldsymbol{\phi}(s, a) \right\|_{(\mathbf{\Lambda}_h^{\tilde{k}})^{-1}} + 2\tilde{\beta} \left\| \boldsymbol{\psi}(s, a, w) \right\|_{(\tilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1}}. \quad \text{(D.28)}
\end{aligned}$$

Note that $\delta_h^k \leq Q_h^{\tilde{k}}(s_h^k, a_h^k, w^k) - Q_h^{\pi^k}(s_h^k, a_h^k, w^k)$. Thus, combining (D.28), Lemma 23, and a union bound over $\widetilde{\mathcal{W}}$, we conclude that for all $(h, k) \in [H] \times [K]$, with probability at least $1 - \delta$, it holds that gives

$$\delta_h^k \leq \xi_{h+1}^k + \delta_{h+1}^k + 4L\beta \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{(\mathbf{\Lambda}_h^{\tilde{k}})^{-1}} + 2\tilde{\beta} \left\| \boldsymbol{\psi}(s_h^k, a_h^k, w^k) \right\|_{(\tilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1}}.$$

Now, we complete the regret analysis following similar steps as those of Theorem 9's

proof:

$$
\begin{aligned}
R_K &= \sum_{k=1}^{K} V_1^*(s_1^k, w^k) - V_1^{\pi^k}(s_1^k, w^k) \\
&\le \sum_{k=1}^{K} V_1^{\tilde{k}}(s_1^k, w^k) - V_1^{\pi^k}(s_1^k, w^k) && \text{(Lemma 25)} \\
&= \sum_{k=1}^{K} \delta_1^k \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H} \xi_h^k + 4L\beta \sum_{k=1}^{K}\sum_{h=1}^{H} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\left(\boldsymbol{\Lambda}_h^{\tilde{k}}\right)^{-1}} + 2\tilde{\beta} \sum_{k=1}^{K}\sum_{h=1}^{H} \left\| \boldsymbol{\psi}(s_h^k, a_h^k, w^k) \right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^{\tilde{k}}\right)^{-1}} \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H} \xi_h^k + 4L\beta \sum_{k=1}^{K}\sum_{h=1}^{H} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}} \sqrt{\frac{\det \boldsymbol{\Lambda}_h^k}{\det \boldsymbol{\Lambda}_h^{\tilde{k}}}} + 2\tilde{\beta} \sum_{k=1}^{K}\sum_{h=1}^{H} \left\| \boldsymbol{\psi}(s_h^k, a_h^k, w^k) \right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}} \sqrt{\frac{\det \tilde{\boldsymbol{\Lambda}}_h^k}{\det \tilde{\boldsymbol{\Lambda}}_h^{\tilde{k}}}} \\
& && \text{(Eqn. (D.18))} \\
&\le 2H\sqrt{T\log(dT/\delta)} + 4H\sqrt{K}\left(L\beta\sqrt{2d\log(1+K/\lambda)} + \tilde{\beta}\sqrt{2md\log(1+K/\lambda)}\right) \\
&\le \tilde{\mathcal{O}}\left(L\sqrt{\lambda m^2 d^3 H^3 T}\right).
\end{aligned}
$$

D.4 Details of Remark 3: Relaxation of Assumption 15

In this section, we replace Assumption 15 with the following assumption:

Assumption 20. *There is a known set $\{w^{(1)}, w^{(2)}, \ldots, w^{(n)}\}$ of $n \le d'$ tasks such that $\boldsymbol{\psi}(s,a,w) \in Span\left(\left\{\boldsymbol{\psi}(s,a,w^{(j)})\right\}_{j\in[n]}\right)$ for all $(s,a,w) \in \mathcal{S}\times\mathcal{A}\times\mathcal{W}$. This implies that for any $(s,a,w) \in \mathcal{S}\times\mathcal{A}\times\mathcal{W}$, there exist coefficients $\{c_j(s,a,w)\}_{j\in[n]}$ such that*

$$
\boldsymbol{\psi}(s,a,w) = \sum_{j\in[n]} c_j(s,a,w)\boldsymbol{\psi}\left(s,a,w^{(j)}\right). \tag{D.29}
$$

Moreover, $\sum_{j\in[n]}\left|c_j(s,a,w)\right| \le L$ for all $(s,a,w) \in \mathcal{S}\times\mathcal{A}\times\mathcal{W}$.

Define the concatenated mapping $\tilde{\boldsymbol{\psi}} : \mathcal{S}\times\mathcal{A}\times\mathcal{W} \to \mathbb{R}^{d+d'}$ such that $\tilde{\boldsymbol{\psi}}(s,a,w) = \left[\boldsymbol{\phi}(s,a)^\top, \boldsymbol{\psi}(s,a,w)^\top\right]^\top$. For any $w \in \mathcal{W}$, define $\mathcal{D}(w) = \Big\{(s,a) : \tilde{\boldsymbol{\psi}}(s,a,w) \text{ are } d+d' \text{ linearly independent vectors.}\Big\}$. Given Assumption 20, we mod-

ify the planning step of UCBlvd to the following:

$$\hat{\boldsymbol{\xi}}_h^k, \left\{\hat{\boldsymbol{\theta}}_h^{k(j)}\right\}_{j\in[n]} = \underset{\boldsymbol{\xi}, \left\{\boldsymbol{\theta}^{(j)}\right\}_{j\in[n]}}{\arg\min} \sum_{j\in[n]} \sum_{(s,a)\in\mathcal{D}\left(w^{(j)}\right)} \left(\left\langle \boldsymbol{\theta}^{(j)}, \boldsymbol{\phi}(s,a)\right\rangle - \left\langle \boldsymbol{\xi}, \boldsymbol{\psi}\left(s,a,w^{(j)}\right)\right\rangle\right)^2 \tag{D.30}$$

$$\text{s.t.} \left\|\boldsymbol{\theta}^{(j)} - \tilde{\boldsymbol{\theta}}_h^k\left(w^{(j)}\right)\right\|_{\boldsymbol{\Lambda}_h^k} \leq \beta,\ \forall j \in [n] \quad \text{and} \quad \|\boldsymbol{\xi}\|_2 \leq H\sqrt{d'}.$$

The only change we make in Algorithm 7 is in Line 11, in which $\hat{\boldsymbol{\xi}}_h^k$ is now computed as defined in (D.30). We present this modification in Algorithm 10 for completeness.

Theorem 17. *Let $T = KH$. Under Assumptions 13, 14, and 20, the number or planning calls in Algorithm 10 is at most $dH\log\left(1+\frac{K}{d\lambda}\right)$ and there exists an absolute constant $c>0$ such that for any fixed $\delta \in (0, 0.5)$, if we set $\lambda = 1$ and $\beta = cH\left(d+\sqrt{d'}\right)\sqrt{\lambda\log(dd'T/\delta)}$ in Algorithm 10, then with probability at least $1-2\delta$, it holds that*

$$R_K \leq 2H\sqrt{T\log(dT/\delta)} + 8HL\beta\sqrt{2dK\log(K)} \leq \tilde{\mathcal{O}}\left(L\sqrt{(d^3+dd')H^3T}\right). \tag{D.31}$$

Proof of Theorem 17 follows exactly the same steps as those of Theorem 10. The only difference is the proof of Lemma 7, which we clarify in the proof of following lemma.

Lemma 26. *Let $\widetilde{\mathcal{W}} = \{w^\tau : \tau \in [K]\} \cup \{w^{(j)} : j \in [n]\}$. Under Assumptions 13, 14, and 20, if we let $\beta = cH\left(d+\sqrt{d'}\right)\sqrt{\lambda\log(dd'T/\delta)}$ with an absolute constant $c>0$, then for all $(s,a,w,h,k) \in \mathcal{S}\times\mathcal{A}\times\mathcal{W}\times[H]\times[K]$ with probability at least $1-\delta$, it holds that*

$$\left|\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle - \mathbb{P}_h\left[V_{h+1}^k(.,w)\right](s,a)\right| \leq 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}}.$$

Proof. We let $\tilde{\boldsymbol{\psi}}_i(w) = \left[\boldsymbol{\phi}_i^\top, \boldsymbol{\psi}_i(w)^\top\right]^\top$ be the i-th element of $\tilde{\mathcal{D}}(w) = \left\{\tilde{\boldsymbol{\psi}}(s,a,w) : (s,a) \in \mathcal{D}(w)\right\}$ and for any triple $(s,a,w) \in \mathcal{S}\times\mathcal{A}\times\mathcal{W}$, we let $\{c_i'(s,a,w)\}_{i\in[d+d']}$ be the coefficients such that

$$\tilde{\boldsymbol{\psi}}(s,a,w) = \sum_{i\in[d+d']} c_i'(s,a,w)\tilde{\boldsymbol{\psi}}_i(w),$$

Algorithm 10 Modified UCBlvd

Set: $Q^k_{H+1}(.,.,.) = 0,\ \forall k \in [K],\ \tilde{k} = 1$
for episodes $k = 1, \ldots, K$ **do**
  Observe the initial state s^k_1 and the task context w^k.
  if $\exists h \in [H]$ such that $\frac{\det \mathbf{\Lambda}^k_h}{\det \mathbf{\Lambda}^{\tilde{k}}_h} > e$ **then**
    $\tilde{k} = k$
    for time-steps $h = H, \ldots, 1$ **do**
      Compute $\hat{\boldsymbol{\xi}}^k_h$ as in (D.30).
    end for
  end if
  for time-steps $h = 1, \ldots, H$ **do**
    Compute $Q^{\tilde{k}}_h(s^k_h, a, w^k)$ for all $a \in \mathcal{A}$ as in (5.7).
    Play $a^k_h = \arg\max_{a \in \mathcal{A}} Q^{\tilde{k}}_h(s^k_h, a, w^k)$ and observe s^k_{h+1} and r^k_h.
  end for
end for

which implies that

$$\boldsymbol{\phi}(s,a) = \sum_{i\in[d+d']} c_i'(s,a,w)\boldsymbol{\phi}_i \quad \text{and} \quad \boldsymbol{\psi}(s,a,w) = \sum_{i\in[d+d']} c_i'(s,a,w)\boldsymbol{\psi}_i(w). \tag{D.32}$$

Thanks to Assumption 14 and conditioned on events $\{\mathcal{E}_1(w)\}_{w\in\widetilde{\mathcal{W}}}$, one set of solution for (D.30) is $\left\{\boldsymbol{\theta}_h^k\left(w^{(j)}\right)\right\}_{j\in[n]}$ and $\boldsymbol{\xi}_h^{V_{h+1}^k}$ with corresponding zero optimal objective value. Therefore, it holds that

$$\left\langle \hat{\boldsymbol{\theta}}_h^{k(j)}, \boldsymbol{\phi}_i \right\rangle = \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}_i\left(w^{(j)}\right)\right\rangle, \quad \forall (i,j) \in [d+d'] \times [n]. \tag{D.33}$$

Moreover, for any triple $(s,a,j) \in \mathcal{S}\times\mathcal{A}\times[n]$, we have

$$\begin{aligned}
\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}\left(s,a,w^{(j)}\right)\right\rangle &= \sum_{i\in[d+d']} c_i'\left(s,a,w^{(j)}\right)\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}_i\left(w^{(j)}\right)\right\rangle && \text{(Eqn. (D.32))}\\
&= \sum_{i\in[d+d']} c_i'\left(s,a,w^{(j)}\right)\left\langle \hat{\boldsymbol{\theta}}_h^{k(j)}, \boldsymbol{\phi}_i\right\rangle && \text{(Eqn. (D.33))}\\
&= \left\langle \hat{\boldsymbol{\theta}}_h^{k(j)}, \boldsymbol{\phi}(s,a)\right\rangle. && \text{(D.34)}
\end{aligned}$$

For any $(s,a,w) \in \mathcal{S}\times\mathcal{A}\times\mathcal{W}$, it holds that

$$\begin{aligned}
\mathbb{P}_h\left[V_{h+1}^k(.,w)\right](s,a) &= \left\langle \boldsymbol{\theta}_h^k(w), \boldsymbol{\phi}(s,a)\right\rangle && \text{(Eqn. (5.4))}\\
&= \left\langle \boldsymbol{\xi}_h^{V_{h+1}^k}, \boldsymbol{\psi}(s,a,w)\right\rangle && \text{(Assumption 14)}\\
&= \sum_{j\in[n]} c_j(s,a,w)\left\langle \boldsymbol{\xi}_h^{V_{h+1}^k}, \boldsymbol{\psi}\left(s,a,w^{(j)}\right)\right\rangle && \text{(Eqn. (D.29))}\\
&= \sum_{j\in[n]} c_j(s,a,w)\mathbb{P}_h\left[V_{h+1}^k\left(.,w^{(j)}\right)\right](s,a)\rangle && \text{(Assumption 14)}\\
&= \sum_{j\in[n]} c_j(s,a,w)\left\langle \boldsymbol{\theta}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a)\right\rangle. && \text{(D.35)}
\end{aligned}$$

Finally, conditioned on events $\{\mathcal{E}_1(w)\}_{w\in\widetilde{\mathcal{W}}}$, for all $(s,a,w,h,k) \in \mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[H]\times[K]$,

it holds that

$$
\begin{aligned}
&\left|\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle - \mathbb{P}_h\left[V_{h+1}^k(.,w)\right](s,a)\right| && \text{(D.36)}\\
&= \left|\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle - \left\langle \boldsymbol{\theta}_h^k(w), \boldsymbol{\phi}(s,a)\right\rangle\right| \\
&= \left|\sum_{j\in[n]} c_j(s,a,w)\left(\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}\left(s,a,w^{(j)}\right)\right\rangle - \left\langle \boldsymbol{\theta}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a)\right\rangle\right)\right| && \text{(Eqns. (D.29) and (D.14))}\\
&\le \left|\sum_{j\in[n]} c_j(s,a,w)\left(\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}\left(s,a,w^{(j)}\right)\right\rangle - \left\langle \hat{\boldsymbol{\theta}}_h^{k(j)}, \boldsymbol{\phi}(s,a)\right\rangle\right)\right| \\
&+ \left|\sum_{j\in[n]} c_j(s,a,w)\left\langle \hat{\boldsymbol{\theta}}_h^{k(j)} - \tilde{\boldsymbol{\theta}}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a)\right\rangle\right| + \left|\sum_{j\in[n]} c_j(s,a,w)\left\langle \tilde{\boldsymbol{\theta}}_h^k\left(w^{(j)}\right) - \boldsymbol{\theta}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a)\right\rangle\right| \\
&= \left|\sum_{j\in[n]} c_j(s,a,w)\left\langle \hat{\boldsymbol{\theta}}_h^{k(j)} - \tilde{\boldsymbol{\theta}}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a)\right\rangle\right| + \left|\sum_{j\in[n]} c_j(s,a,w)\left\langle \tilde{\boldsymbol{\theta}}_h^k\left(w^{(j)}\right) - \boldsymbol{\theta}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a)\right\rangle\right| \\
& && \text{(Eqn. (D.13))}\\
&\le 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}}. && \text{(Lemma 19)}
\end{aligned}
$$

□

D.5 Details of Remark 4

In this section, we only rely on the following two assumptions:

Assumption 21. *Given a feature map $\boldsymbol{\psi}: \mathcal{S}\times\mathcal{A}\times\mathcal{W}\to\mathbb{R}^{d'}$, consider function class*

$$
\mathcal{F} = \left\{ f : f(s,w) = \min\left\{\max_{a\in\mathcal{A}}\left\{\langle \boldsymbol{\nu}, \boldsymbol{\psi}(s,a,w)\rangle + \beta\left\|\boldsymbol{\psi}(s,a,w)\right\|_{\boldsymbol{\Lambda}^{-1}}\right\}^+, H\right\} \boldsymbol{\nu}\in\mathbb{R}^{d'}, \beta\ge 0, \boldsymbol{\Lambda}\in\mathbf{S}_{++}^{d'}\right\}. \tag{D.37}
$$

Then for any $f\in\mathcal{F}$ and $h\in[H]$, there exists a vector $\boldsymbol{\nu}_h^f\in\mathbb{R}^{d'}$ with $\left\|\boldsymbol{\nu}_h^f\right\|_2 \le H\sqrt{d'}$ such that

$$
\mathbb{P}_h\left[f(.,w)\right](s,a) = \langle \boldsymbol{\psi}(s,a,w), \boldsymbol{\nu}_h^f\rangle. \tag{D.38}
$$

Moreover, for every $h\in[H]$, there exists a vector $\boldsymbol{\eta}_h$ such that $r_h(s,a,w) = \left\langle \boldsymbol{\eta}_h, \boldsymbol{\psi}(s,a,w)\right\rangle$.

Assumption 22. *Without loss of generality, $\left\|\boldsymbol{\psi}(s,a,w)\right\|_2 \le 1$ for all $(s,a,w)\in\mathcal{S}\times\mathcal{A}\times\mathcal{W}$, and $\left\|\boldsymbol{\eta}_h\right\|_2 \le \sqrt{d'}$ for all $h\in[H]$.*

D.5.1 Overview

Let $\boldsymbol{\psi}_h^\tau = \boldsymbol{\psi}(s_h^\tau, a_h^\tau, w^\tau)$. Standard Lifelong-LSVI with computation sharing works with the following action-value functions:

$$Q_h^k(s,a,w) = \left\{ r_h(s,a,w) + \left\langle \tilde{\boldsymbol{\nu}}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle + \beta \left\| \boldsymbol{\psi}(s,a,w) \right\|_{(\tilde{\boldsymbol{\Lambda}}_h^k)^{-1}} \right\}^+, \tag{D.39}$$

where

$$\tilde{\boldsymbol{\nu}}_h^k = \left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\psi}_h^\tau . \min\left\{ \max_{a\in\mathcal{A}} Q_{h+1}^k(s_{h+1}^\tau, a, w^\tau), H \right\} \quad \text{and} \quad \tilde{\boldsymbol{\Lambda}}_h^k = \lambda \mathbf{I}_{d'} + \sum_{\tau=1}^{k-1} \boldsymbol{\psi}_h^\tau \boldsymbol{\psi}_h^{\tau\top}. \tag{D.40}$$

Algorithm 11 Standard Lifelong-LSVI with Computation Sharing

Set: $Q_{H+1}^k(.,.,.) = 0,\ \forall k \in [K],\ \tilde{k} = 1$
for episodes $k = 1, \ldots, K$ **do**
  Observe the initial state s_1^k and the task context w^k.
  if $\exists h \in [H]$ such that $\frac{\det \tilde{\boldsymbol{\Lambda}}_h^k}{\det \tilde{\boldsymbol{\Lambda}}_h^{\tilde{k}}} > e$ **then**
    $\tilde{k} = k$
    for time-steps $h = H, \ldots, 1$ **do**
      Compute $\tilde{\boldsymbol{\nu}}_h^{\tilde{k}}$ as in (D.40).
    end for
  end if
  for time-steps $h = 1, \ldots, H$ **do**
    Compute $Q_h^{\tilde{k}}(s_h^k, a, w^k)$ for all $a \in \mathcal{A}$ as in (D.39).
    Play $a_h^k = \arg\max_{a\in\mathcal{A}} Q_h^{\tilde{k}}(s_h^k, a, w^k)$ and observe s_{h+1}^k and r_h^k.
  end for
end for

Theorem 18. *Let $T = KH$. Under Assumptions 21 and 22, the number of planning calls in 11 is at most $d'H \log\left(1 + \frac{K}{d'\lambda}\right)$ and there exists an absolute constant $c > 0$ such that for any fixed $\delta \in (0, 0.5)$, if we set $\lambda = 1$ and $\beta = cd'H\sqrt{\log(d'T/\delta)}$ in Algorithm 11, then with*

probability at least $1-2\delta$*, it holds that*

$$R_K \leq 2H\sqrt{T\log(d'T/\delta)} + 4H\beta\sqrt{2d'K\log(K)} \leq \tilde{\mathcal{O}}\left(\sqrt{d'^3H^3T}\right).$$

D.5.2 Necessary Analysis for the Proof of Theorem 18

Thanks to Assumption 21, we have

$$\mathbb{P}_h\left[V_{h+1}^k(.,w)\right](s,a) = \left\langle \boldsymbol{\nu}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle, \tag{D.41}$$

where $\boldsymbol{\nu}_h^k = \boldsymbol{\nu}_h^{V_{h+1}^k}$.

Lemma 27. *Let* c_β *be a constant in the definition of* β*. Then, under Assumption 22, there is an absolute constant* c_0 *independent of* c_β*, such that for all* $(h,k) \in [H]\times[K]$*, with probability at least* $1-\delta$ *it holds that*

$$\left\|\sum_{\tau=1}^{k-1}\boldsymbol{\psi}_h^\tau.\left(V_{h+1}^k(s_{h+1}^\tau,w^\tau) - \mathbb{P}_h[V_{h+1}^k(.,w^\tau)](s_h^\tau,a_h^\tau)\right)\right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}} \leq c_0 d'H\sqrt{\log((c_\beta+1)d'T/\delta)},$$

where c_0 *and* c_β *are two independent absolute constants.*

Proof. We note that $\left\|\boldsymbol{\eta}_h + \tilde{\boldsymbol{\nu}}_h^k\right\|_2 \leq (1+H)\sqrt{d'}$ and $\left\|\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}\right\| \leq \frac{1}{\lambda}$. Thus, Lemmas 33 and 38 together imply that for all $(h,k)\in[H]\times[K]$, with probability at least $1-\delta$ it holds that

$$\begin{aligned}
&\left\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}_h^\tau\left(V_{h+1}^k(s_{h+1}^\tau,w^\tau) - \mathbb{P}_h[V_{h+1}^k(.,w^\tau)](s_h^\tau,a_h^\tau)\right)\right\|^2_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}} \\
&\leq 4H^2\left(\frac{d'}{2}\log\left(\frac{k+\lambda}{\lambda}\right) + d'\log(1+8H\sqrt{d'}/\epsilon) + d'^2\log\left(\frac{1+32L^2\beta^2\sqrt{d'}}{\lambda\epsilon^2}\right) + \log\left(\frac{1}{\delta}\right)\right) + \frac{8k^2\epsilon^2}{\lambda}.
\end{aligned}$$

If we let $\epsilon = \frac{dH}{k}$ and $\beta = c_\beta(d'+\sqrt{d'})H\sqrt{\log(dT/\delta)}$, then, there exists an absolute constant $C>0$ that is independent of c_β such that

$$\left\|\sum_{\tau=1}^{k-1}\boldsymbol{\phi}_h^\tau\left(V_{h+1}^k(s_{h+1}^\tau,w^\tau) - \mathbb{P}_h[V_{h+1}^k(.,w^\tau)](s_h^\tau,a_h^\tau)\right)\right\|^2_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}} \leq C(d'+d'^2)H^2\log\left((c_\beta+1)d'T/\delta\right).$$

□

Lemma 28. *Under Assumptions 21 and 22, if we let* $\beta = cd'H\sqrt{\lambda \log(d'T/\delta)}$ *with an absolute constant* $c > 0$*, then the event*

$$\mathcal{E}_4 := \left\{ \left\| \boldsymbol{\nu}_h^k - \tilde{\boldsymbol{\nu}}_h^k \right\|_{\tilde{\boldsymbol{\Lambda}}_h^k} \leq \beta,\ \forall (h,k) \in [H] \times [K] \right\}. \tag{D.42}$$

holds with probability at least $1 - \delta$*.*

Proof.

$$\begin{aligned}
\boldsymbol{\nu}_h^k - \tilde{\boldsymbol{\nu}}_h^k &= \boldsymbol{\nu}_h^k - \left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\psi}_h^\tau V_{h+1}^k(s_{h+1}^\tau, w^\tau) \\
&= \left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1} \left(\tilde{\boldsymbol{\Lambda}}_h^k \boldsymbol{\nu}_h^k - \sum_{\tau=1}^{k-1} \boldsymbol{\psi}_h^\tau V_{h+1}^k(s_{h+1}^\tau, w^\tau) \right) \\
&= \underbrace{\lambda \left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1} \boldsymbol{\nu}_h^k}_{\mathbf{q}_1} \underbrace{- \left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1} \left(\sum_{\tau=1}^{k-1} \boldsymbol{\psi}_h^\tau \left(V_{h+1}^k(s_{h+1}^\tau, w^\tau) - \mathbb{P}_h[V_{h+1}^k(., w^\tau)](s_h^\tau, a_h^\tau) \right) \right)}_{\mathbf{q}_2}.
\end{aligned}$$

(Eqn. (D.41))

Thus, in order to upper bound $\left\| \boldsymbol{\nu}_h^k - \tilde{\boldsymbol{\nu}}_h^k(w) \right\|_{\tilde{\boldsymbol{\Lambda}}_h^k}$, we bound $\|\mathbf{q}_1\|_{\tilde{\boldsymbol{\Lambda}}_h^k}$ and $\|\mathbf{q}_2\|_{\tilde{\boldsymbol{\Lambda}}_h^k}$ separately.

From Assumption 22, we have

$$\|\mathbf{q}_1\|_{\boldsymbol{\Lambda}_h^k} = \lambda \left\| \boldsymbol{\nu}_h^k \right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}} \leq \sqrt{\lambda} \left\| \boldsymbol{\nu}_h^k \right\|_2 \leq H\sqrt{\lambda d'}. \tag{D.43}$$

Thanks to Lemma 27, for all $(h,k) \in [H] \times [K]$, with probability at least $1 - \delta$, it holds that

$$\begin{aligned}
\|\mathbf{q}_2\|_{\tilde{\boldsymbol{\Lambda}}_h^k} &\leq \left\| \sum_{\tau=1}^{k-1} \boldsymbol{\psi}_h^\tau \left(V_{h+1}^k(s_{h+1}^\tau, w^\tau) - \mathbb{P}_h[V_{h+1}^k(., w^\tau)](s_h^\tau, a_h^\tau) \right) \right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} \\
&\leq c_0 d' H \sqrt{\log((c_\beta + 1)d'T/\delta)},
\end{aligned} \tag{D.44}$$

where c_0 and c_β are two independent absolute constants.

Combining (D.43) and (D.44), for all $(h,k) \in [H] \times [K]$, with probability at least $1 - \delta$, it holds that

$$\left\| \boldsymbol{\nu}_h^k - \tilde{\boldsymbol{\nu}}_h^k \right\|_{\tilde{\boldsymbol{\Lambda}}_h^k} \leq cd'H\sqrt{\lambda \log(d'T/\delta)}$$

for some absolute constant $c > 0$. □

Lemma 29. *Let the setting of Lemma 28 holds. Conditioned on events $\mathcal{E}_4$ defined in* (D.42)*, and with Q_h^k computed as in* (D.39)*, it holds that $Q_h^k(s,a,w) \geq Q_h^*(s,a,w)$ for all $(s,a,w,h,k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{W} \times [H] \times [K]$.*

Proof. We first note that conditioned on the event $\mathcal{E}_4$, for all $(s,a,w,h,k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{W} \times [H] \times [K]$, it holds that

$$
\begin{aligned}
&\left| r_h(s,a,w) + \left\langle \tilde{\boldsymbol{\nu}}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle - Q_h^\pi(s,a,w) - \mathbb{P}_h \left[V_{h+1}^k(.,w) - V_{h+1}^\pi(.,w) \right](s,a) \right| \\
&= \left| r_h(s,a,w) + \left\langle \tilde{\boldsymbol{\nu}}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle - r_h(s,a,w) - \mathbb{P}_h \left[V_{h+1}^k(.,w) \right](s,a) \right| \\
&= \left| \left\langle \tilde{\boldsymbol{\nu}}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle - \mathbb{P}_h \left[V_{h+1}^k(.,w) \right](s,a) \right| \\
&= \left| \left\langle \tilde{\boldsymbol{\nu}}_h^k - \boldsymbol{\nu}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle \right| \\
&\leq \left\| \tilde{\boldsymbol{\nu}}_h^k - \boldsymbol{\nu}_h^k \right\|_{\tilde{\boldsymbol{\Lambda}}_h^k} \left\| \boldsymbol{\psi}(s,a,w) \right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}} \\
&\leq \beta \left\| \boldsymbol{\psi}(s,a,w) \right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}}, && \text{(Lemma 28)}
\end{aligned}
$$

for any policy π.

Now, we prove the lemma by induction. The statement holds for H because $Q_{H+1}^k(.,.,.) = Q_{H+1}^*(.,.,.) = 0$ and thus conditioned on the event $\mathcal{E}_4$, defined in (D.42), for all $(s,a,w,k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{W} \times [K]$, we have

$$
\left| r_h(s,a,w) + \left\langle \boldsymbol{\nu}_H^k, \boldsymbol{\psi}(s,a,w) \right\rangle - Q_H^*(s,a,w) \right| \leq \beta \left\| \boldsymbol{\psi}(s,a,w) \right\|_{\left(\tilde{\boldsymbol{\Lambda}}_H^k\right)^{-1}}.
$$

Therefore, conditioned on the event $\mathcal{E}_4$, for all $(s,a,w,k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{W} \times [K]$, we have

$$
\begin{aligned}
Q_H^*(s,a,w) &\leq r_H(s,a,w) + \left\langle \boldsymbol{\nu}_H^k, \boldsymbol{\psi}(s,a,w) \right\rangle + \beta \left\| \boldsymbol{\psi}(s,a,w) \right\|_{(\tilde{\boldsymbol{\Lambda}}_H^k)^{-1}} \\
&= \left\{ r_H(s,a,w) + \left\langle \boldsymbol{\nu}_H^k, \boldsymbol{\psi}(s,a,w) \right\rangle + \beta \left\| \boldsymbol{\psi}(s,a,w) \right\|_{(\tilde{\boldsymbol{\Lambda}}_H^k)^{-1}} \right\}^+ \\
&= Q_H^k(s,a,w),
\end{aligned}
$$

where the first equality follows from the fact that $Q_H^*(s,a,w) \geq 0$. Now, suppose the statement holds at time-step $h+1$ and consider time-step h. Conditioned on events $\mathcal{E}_4$, for

all $(s,a,w,h,k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{W} \times [H] \times [K]$, we have

$$
\begin{aligned}
0 &\le r_h(s,a,w) + \left\langle \boldsymbol{\nu}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle - Q_h^*(s,a,w) - \mathbb{P}_h \left[V_{h+1}^k(.,w) - V_{h+1}^*(.,w) \right](s,a) \\
&+ \beta \left\| \boldsymbol{\psi}(s,a,w) \right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}} \\
&\le r_h(s,a,w) + \left\langle \boldsymbol{\nu}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle - Q_h^*(s,a,w) + \beta \left\| \boldsymbol{\psi}(s,a,w) \right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}}.
\end{aligned}
$$

(Induction assumption)

Therefore, conditioned on events $\mathcal{E}_4$, for all $(s,a,w,h,k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{W} \times [H] \times [K]$, we have

$$
\begin{aligned}
Q_h^*(s,a,w) &\le r_h(s,a,w) + \left\langle \boldsymbol{\nu}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle + \beta \left\| \boldsymbol{\psi}(s,a,w) \right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}} \\
&= \left\{ r_h(s,a,w) + \left\langle \boldsymbol{\nu}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle + \beta \left\| \boldsymbol{\psi}(s,a,w) \right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}} \right\}^+ \\
&= Q_h^k(s,a,w),
\end{aligned}
$$

where the first equality follows from the fact that $Q_H^*(s,a,w) \ge 0$. This completes the proof.

□

D.5.3 Proof of Theorem 18

First, we bound the number of times Algorithm 11 updates $\tilde{\boldsymbol{\nu}}_h^k$. Let P be the total number of updates and k_p be the episode at which, the agent did replanning for the p-th time. Note that $\det \tilde{\boldsymbol{\Lambda}}_h^1 = \lambda^{d'}$ and $\det \tilde{\boldsymbol{\Lambda}}_h^K \le \mathrm{trace}(\tilde{\boldsymbol{\Lambda}}_h^K / d')^{d'} \le \left(\lambda + \frac{K}{d'}\right)^{d'}$, and consequently:

$$
\frac{\det \tilde{\boldsymbol{\Lambda}}_h^K}{\det \tilde{\boldsymbol{\Lambda}}_h^1} = \prod_{p=1}^{P} \frac{\det \tilde{\boldsymbol{\Lambda}}_h^{k_p}}{\det \tilde{\boldsymbol{\Lambda}}_h^{k_{p-1}}} \le \left(1 + \frac{K}{d'\lambda}\right)^{d'},
$$

and therefore

$$
\prod_{h=1}^{H} \frac{\det \tilde{\boldsymbol{\Lambda}}_h^K}{\det \tilde{\boldsymbol{\Lambda}}_h^1} = \prod_{h=1}^{H} \prod_{p=1}^{P} \frac{\det \tilde{\boldsymbol{\Lambda}}_h^{k_p}}{\det \tilde{\boldsymbol{\Lambda}}_h^{k_{p-1}}} \le \left(1 + \frac{K}{d'\lambda}\right)^{d'H}. \tag{D.45}
$$

Since $1 \le \frac{\det \tilde{\boldsymbol{\Lambda}}_h^{k_p}}{\det \tilde{\boldsymbol{\Lambda}}_h^{k_{p-1}}}$ for all $p \in [P]$, we can deduce from (D.45) that

$$
\exists h \in [H] \quad \text{such that} \quad e < \frac{\det \tilde{\boldsymbol{\Lambda}}_h^k}{\det \tilde{\boldsymbol{\Lambda}}_h^{\tilde{k}}}
$$

happens for at most $d'H\log\left(1+\frac{K}{d'\lambda}\right)$ number of episodes $k\in[K]$. This concludes that number of planning calls in Algorithm 11 is at most $d'H\log\left(1+\frac{K}{d'\lambda}\right)$.

Now, we prove the regret bound. Let $\delta_h^k = V_h^{\tilde{k}}(s_h^k,w^k) - V_h^{\pi^k}(s_h^k,w^k)$ and $\xi_{h+1}^k = \mathbb{E}\left[\delta_{h+1}^k|s_h^k,a_h^k\right]-\delta_{h+1}^k$. Conditioned on $\mathcal{E}_4$, for all $(s,a,w,h,k)\in\mathcal{S}\times\mathcal{A}\times\mathcal{W}\times[H]\times[K]$, we have

$$\begin{aligned}Q_h^{\tilde{k}}(s,a,w) - Q_h^{\pi^k}(s,a,w) &= r_h(s,a,w) + \left\langle \boldsymbol{\theta}_h^{\tilde{k}}, \boldsymbol{\psi}(s,a,w)\right\rangle - Q_h^{\pi^k}(s,a,w) + \beta\left\|\boldsymbol{\psi}(s,a,w)\right\|_{(\tilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1}}\\ &\le \mathbb{P}_h\left[V_{h+1}^{\tilde{k}}(.,w) - V_{h+1}^{\pi^k}(.,w)\right](s,a) + 2\beta\left\|\boldsymbol{\psi}(s,a,w)\right\|_{(\tilde{\boldsymbol{\Lambda}}_h^{v})^{-1}}.\end{aligned} \tag{D.46}$$

Note that $\delta_h^{\tilde{k}} \le Q_h^k(s_h^k,a_h^k,w^k) - Q_h^{\pi^k}(s_h^k,a_h^k,w^k)$. Thus, (D.46) and Lemma 28 imply that for all $(h,k)\in[H]\times[K]$, it holds that

$$\delta_h^k \le \xi_{h+1}^k + \delta_{h+1}^k + 2\beta\left\|\boldsymbol{\psi}(s_h^k,a_h^k,w^k)\right\|_{(\tilde{\boldsymbol{\Lambda}}_h^k)^{-1}}.$$

Now, we complete the regret analysis following similar steps as those of Theorem 9's proof:

$$\begin{aligned}R_K &= \sum_{k=1}^K V_1^*(s_1^k,w^k) - V_1^{\pi^k}(s_1^k,w^k)\\ &\le \sum_{k=1}^K V_1^{\tilde{k}}(s_1^k,w^k) - V_1^{\pi^k}(s_1^k,w^k) && \text{(Lemma 29)}\\ &= \sum_{k=1}^K \delta_1^k\\ &\le \sum_{k=1}^K\sum_{h=1}^H \xi_h^k + 2\beta\sum_{k=1}^K\sum_{h=1}^H\left\|\boldsymbol{\psi}(s_h^k,a_h^k,w^k)\right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^{\tilde{k}}\right)^{-1}}\\ &\le \sum_{k=1}^K\sum_{h=1}^H \xi_h^k + 2\beta\sum_{k=1}^K\sum_{h=1}^H\left\|\boldsymbol{\psi}(s_h^k,a_h^k,w^k)\right\|_{\left(\tilde{\boldsymbol{\Lambda}}_h^{k}\right)^{-1}}\sqrt{\frac{\det\tilde{\boldsymbol{\Lambda}}_h^k}{\det\tilde{\boldsymbol{\Lambda}}_h^{\tilde{k}}}} && \text{(Eqn. (D.18))}\\ &\le 2H\sqrt{T\log(d'T/\delta)} + 4H\beta\sqrt{2\lambda d'K\log(1+K/\lambda)}\\ &\le \tilde{\mathcal{O}}\left(\sqrt{\lambda d'^3H^3T}\right).\end{aligned}$$

D.6 Details of Remark 5: A misspecified setting

We first present a definition for an approximate completeness model.

Assumption 23 (ζ-Approximate Completeness)**.** *Given feature maps $\boldsymbol{\phi}: \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and $\boldsymbol{\psi}: \mathcal{S} \times \mathcal{A} \times \mathcal{W} \to \mathbb{R}^{d'}$ in Assumption 13, consider the function class*

$$\mathcal{F} = \left\{ f : f(s,w) = \min\left\{ \max_{a \in \mathcal{A}} \left\{ \langle \boldsymbol{\nu}, \boldsymbol{\psi}(s,a,w) \rangle + \beta \left\| \boldsymbol{\phi}(s,a) \right\|_{\boldsymbol{\Lambda}^{-1}} \right\}^+, H \right\}, \boldsymbol{\nu} \in \mathbb{R}^{d'}, \boldsymbol{\Lambda} \in \mathbf{S}_{++}^d, \beta \geq 0 \right\}.$$

For any $f \in \mathcal{F}$ and $h \in [H]$, there exists a vector $\boldsymbol{\xi}_h^f \in \mathbb{R}^{d'}$ with $\left\| \boldsymbol{\xi}_h^f \right\| \leq H\sqrt{d'}$ such that for all $(s,a,w) \in \mathcal{S} \times \mathcal{A} \times \mathcal{W}$

$$\left| \mathbb{P}_h \left[f(.,w) \right](s,a) - \langle \boldsymbol{\xi}_h^f, \boldsymbol{\psi}(s,a,w) \rangle \right| \leq \zeta.$$

Theorem 19. *Let $T = KH$. Under Assumptions 13, 23, and 15, the number of planning calls in Algorithm 7 is at most $dH \log(1 + \frac{K}{d\lambda})$, and there exists an absolute constant $c > 0$ such that for any fixed $\delta \in (0, 0.5)$, if we set $\lambda = 1$ and $\beta = cH(d + \sqrt{md})\sqrt{\log(mdT/\delta)}$ in Algorithm 7, then with probability at least $1 - 2\delta$, it holds that*

$$R_K \leq \tilde{\mathcal{O}} \left(\sqrt{md} T\zeta + \sqrt{(d^3 + md^2)H^3 T} \right).$$

D.6.1 Necessary Analysis for the Proof of Theorem 19

Let $\left(s^{(i)}, a^{(i)} \right)$ be the i-th element of $\mathcal{D}$ and $\{c_i'(s,a)\}_{i \in [d]}$ be the coefficients such that

$$\boldsymbol{\phi}(s,a) = \sum_{i \in [d]} c_i'(s,a) \boldsymbol{\phi}\left(s^{(i)}, a^{(i)} \right).$$

Then, $L_{\boldsymbol{\phi}}$ is a positive constant such that $\sum_{i \in [d]} \left| c_i'(s,a) \right| \leq L_{\boldsymbol{\phi}}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Lemma 30. *Let $\widetilde{\mathcal{W}} = \{w^\tau : \tau \in [K]\} \cup \{w^{(j)} : j \in [n]\}$. Under the setting of Theorem 19 and conditioned on events $\{\mathcal{E}_1(w)\}_{w \in \widetilde{\mathcal{W}}}$ defined in (5.9), for all $(s,a,w,h,k) \in \mathcal{S} \times \mathcal{A} \times \widetilde{\mathcal{W}} \times [H] \times [K]$, it holds that*

$$\left| \langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w) \rangle - \mathbb{P}_h[V_{h+1}^k(.,w)](s,a) \right| \leq (2L + L_{\boldsymbol{\phi}}\sqrt{md})\zeta + 2L\beta \left\| \boldsymbol{\phi}(s,a) \right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}}.$$

Proof. Thanks to Assumption 23 and conditioned on events $\{\mathcal{E}_1(w)\}_{w\in\widetilde{\mathcal{W}}}$, one set of feasible parameters for (5.8) is $\left\{\boldsymbol{\theta}_h^k\left(w^{(j)}\right)\right\}_{j\in[n]}$ and $\boldsymbol{\xi}_h^{V_{h+1}^k}$ such that

$$\left|\left\langle\hat{\boldsymbol{\theta}}_h^{k(j)},\boldsymbol{\phi}(s,a)\right\rangle-\left\langle\hat{\boldsymbol{\xi}}_h^k,\boldsymbol{\psi}\left(s,a,w^{(j)}\right)\right\rangle\right|\leq\zeta\sqrt{md},\quad\forall(j,(s,a))\in[n]\times\mathcal{D}. \tag{D.47}$$

For any triple $(s,a,j)\in\mathcal{S}\times\mathcal{A}\times[n]$, we have

$$\begin{aligned}
\left\langle\hat{\boldsymbol{\xi}}_h^k,\boldsymbol{\psi}\left(s,a,w^{(j)}\right)\right\rangle&=\left\langle\hat{\boldsymbol{\xi}}_h^k,\boldsymbol{\phi}(s,a)\otimes\boldsymbol{\rho}\left(w^{(j)}\right)\right\rangle\\
&=\left\langle\hat{\boldsymbol{\xi}}_h^k,\sum_{i\in[d]}c_i'(s,a)\boldsymbol{\phi}\left(s^{(i)},a^{(i)}\right)\otimes\boldsymbol{\rho}\left(w^{(j)}\right)\right\rangle\\
&=\sum_{i\in[d]}c_i'(s,a)\left\langle\hat{\boldsymbol{\xi}}_h^k,\boldsymbol{\psi}\left(s^{(i)},a^{(i)},w^{(j)}\right)\right\rangle&\text{(Assumption 15)}\\
&\leq\sqrt{md}\zeta\sum_{i\in[d]}c_i'(s,a)+\sum_{i\in[d]}c_i'(s,a)\left\langle\hat{\boldsymbol{\theta}}_h^{k(j)},\boldsymbol{\phi}\left(s^{(i)},a^{(i)}\right)\right\rangle&\text{(Eqn. (D.47))}\\
&\leq L_{\boldsymbol{\phi}}\sqrt{md}\zeta+\left\langle\hat{\boldsymbol{\theta}}_h^{k(j)},\boldsymbol{\phi}(s,a)\right\rangle.
\end{aligned}$$

Similarly, it holds that $\left\langle\hat{\boldsymbol{\xi}}_h^k,\boldsymbol{\psi}\left(s,a,w^{(j)}\right)\right\rangle\geq-L_{\boldsymbol{\phi}}\sqrt{md}\zeta+\left\langle\hat{\boldsymbol{\theta}}_h^{k(j)},\boldsymbol{\phi}(s,a)\right\rangle$. Therefore, for any $(s,a,j)\in\mathcal{S}\times\mathcal{A}\times[n]$, it holds that

$$\left|\left\langle\hat{\boldsymbol{\xi}}_h^k,\boldsymbol{\psi}\left(s,a,w^{(j)}\right)\right\rangle-\left\langle\hat{\boldsymbol{\theta}}_h^{k(j)},\boldsymbol{\phi}(s,a)\right\rangle\right|\leq L_{\boldsymbol{\phi}}\sqrt{md}\zeta. \tag{D.48}$$

For any $(s,a,w) \in \mathcal{S} \times \mathcal{A} \times \mathcal{W}$, it holds that

$$\begin{aligned}
\mathbb{P}_h \left[V_{h+1}^k(.,w) \right](s,a) &= \left\langle \boldsymbol{\theta}_h^k(w), \boldsymbol{\phi}(s,a) \right\rangle && \text{(Eqn. (5.4))} \\
&\leq \zeta + \left\langle \boldsymbol{\xi}_h^{V_{h+1}^k}, \boldsymbol{\psi}(s,a,w) \right\rangle && \text{(Assumption 23)} \\
&= \zeta + \sum_{j \in [n]} c_j(w) \left\langle \boldsymbol{\xi}_h^{V_{h+1}^k}, \boldsymbol{\psi}\left(s,a,w^{(j)}\right) \right\rangle && \text{(Assumption 15)} \\
&\leq \zeta \left(1 + \sum_{j \in [n]} c_j(w) \right) + \sum_{j \in [n]} c_j(w) \mathbb{P}_h \left[V_{h+1}^k \left(., w^{(j)}\right) \right](s,a) \\
& && \text{(Assumption 23)} \\
&\leq 2L\zeta + \sum_{j \in [n]} c_j(w) \left\langle \boldsymbol{\theta}_h^k \left(w^{(j)}\right), \boldsymbol{\phi}(s,a) \right\rangle. && \text{(Assumption 15)}
\end{aligned}$$

Similarly, it holds that $\mathbb{P}_h \left[V_{h+1}^k(.,w) \right](s,a) \geq -2L\zeta + \sum_{j \in [n]} c_j(w) \left\langle \boldsymbol{\theta}_h^k \left(w^{(j)}\right), \boldsymbol{\phi}(s,a) \right\rangle$. Therefore, for any $(s,a,w) \in \mathcal{S} \times \mathcal{A} \times \mathcal{W}$, it holds that

$$\left| \mathbb{P}_h \left[V_{h+1}^k(.,w) \right](s,a) - \sum_{j \in [n]} c_j(w) \left\langle \boldsymbol{\theta}_h^k \left(w^{(j)}\right), \boldsymbol{\phi}(s,a) \right\rangle \right| \leq 2L\zeta. \tag{D.49}$$

Finally, conditioned on events $\{\mathcal{E}_1(w)\}_{w \in \widetilde{\mathcal{W}}}$, for all $(s,a,w,h,k) \in \mathcal{S} \times \mathcal{A} \times \widetilde{\mathcal{W}} \times [H] \times [K]$,

it holds that

$$
\begin{aligned}
&\left|\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w)\right\rangle - \mathbb{P}_h\left[V_{h+1}^k(.,w)\right](s,a)\right| \\
&\leq 2L\zeta + \left|\sum_{j\in[n]} c_j(w)\left(\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}\left(s,a,w^{(j)}\right)\right\rangle - \left\langle \boldsymbol{\theta}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a)\right\rangle\right)\right| \\
&\qquad\qquad \text{(Assumption 15 and Eqn. (D.49))} \\
&\leq 2L\zeta + \left|\sum_{j\in[n]} c_j(w)\left(\left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}\left(s,a,w^{(j)}\right)\right\rangle - \left\langle \hat{\boldsymbol{\theta}}_h^{k(j)}, \boldsymbol{\phi}(s,a)\right\rangle\right)\right| \\
&+ \left|\sum_{j\in[n]} c_j(w)\left\langle \hat{\boldsymbol{\theta}}_h^{k(j)} - \tilde{\boldsymbol{\theta}}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a)\right\rangle\right| + \left|\sum_{j\in[n]} c_j(w)\left\langle \tilde{\boldsymbol{\theta}}_h^k\left(w^{(j)}\right) - \boldsymbol{\theta}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a)\right\rangle\right| \\
&\leq (2L + L_{\boldsymbol{\phi}}\sqrt{md})\zeta + \left|\sum_{j\in[n]} c_j(w)\left\langle \hat{\boldsymbol{\theta}}_h^{k(j)} - \tilde{\boldsymbol{\theta}}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a)\right\rangle\right| \\
&+ \left|\sum_{j\in[n]} c_j(w)\left\langle \tilde{\boldsymbol{\theta}}_h^k\left(w^{(j)}\right) - \boldsymbol{\theta}_h^k\left(w^{(j)}\right), \boldsymbol{\phi}(s,a)\right\rangle\right| \qquad \text{(Eqn. (D.48))} \\
&\leq (2L + L_{\boldsymbol{\phi}}\sqrt{md})\zeta + 2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}}. \qquad \text{(Lemma 19)}
\end{aligned}
$$

□

As the final step in the regret analysis, we state the following lemma which uses Lemma 30 to prove the optimistic nature of UCBlvd. Then following the standard analysis of single-task LSVI-UCB we derive the regret bound for misspecified settings.

Lemma 31. *Let $\widetilde{\mathcal{W}} = \{w^\tau : \tau \in [K]\} \cup \{w^{(j)} : j \in [n]\}$. Under the setting of Theorem 19 and conditioned on events $\{\mathcal{E}_1(w)\}_{w\in\widetilde{\mathcal{W}}}$ defined in (5.9), and with Q_h^k computed as in (5.7), it holds that $(2L + L_{\boldsymbol{\phi}}\sqrt{md})(H-h+1)\zeta + Q_h^k(s,a,w) \geq Q_h^*(s,a,w)$ for all $(s,a,w,h,k) \in \mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[H]\times[K]$.*

Proof. We first note that conditioned on events $\{\mathcal{E}_1(w)\}_{w\in\widetilde{\mathcal{W}}}$, for all $(s,a,w,h,k) \in \mathcal{S}\times\mathcal{A}\times$

$\widetilde{\mathcal{W}} \times [H] \times [K]$, it holds that

$$\begin{aligned}
&\left| r_h(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle - Q_h^\pi(s,a,w) - \mathbb{P}_h \left[V_{h+1}^k(.,w) - V_{h+1}^\pi(.,w) \right](s,a) \right| \\
&= \left| r_h(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle - r_h(s,a,w) - \mathbb{P}_h \left[V_{h+1}^k(.,w) \right](s,a) \right| \\
&= \left| \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle - \mathbb{P}_h \left[V_{h+1}^k(.,w) \right](s,a) \right| \\
&\leq (2L + L_{\boldsymbol{\phi}}\sqrt{md})\zeta + 2L\beta \left\| \boldsymbol{\phi}(s,a) \right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}}, \qquad \text{(Lemma 30)}
\end{aligned}$$

for any policy π.

Now, we prove the lemma by induction. The statement holds for H because $Q_{H+1}^k(.,.,.) = Q_{H+1}^*(.,.,.) = 0$ and thus conditioned events $\{\mathcal{E}_1(w)\}_{w \in \widetilde{\mathcal{W}}}$, defined in (5.9), for all $(s,a,w,k) \in \mathcal{S} \times \mathcal{A} \times \widetilde{\mathcal{W}} \times [K]$, we have

$$\left| r_H(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_H^k, \boldsymbol{\psi}(s,a,w) \right\rangle - Q_H^*(s,a,w) \right| \leq (2L + L_{\boldsymbol{\phi}}\sqrt{md})\zeta + 2L\beta \left\| \boldsymbol{\phi}(s,a) \right\|_{\left(\boldsymbol{\Lambda}_H^k\right)^{-1}}.$$

Therefore, conditioned on events $\{\mathcal{E}_1(w)\}_{w \in \widetilde{\mathcal{W}}}$, for all $(s,a,w,k) \in \mathcal{S} \times \mathcal{A} \times \widetilde{\mathcal{W}} \times [K]$, we have

$$\begin{aligned}
Q_H^*(s,a,w) &\leq r_H(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_H^k, \boldsymbol{\psi}(s,a,w) \right\rangle + 2L\beta \left\| \boldsymbol{\phi}(s,a) \right\|_{(\boldsymbol{\Lambda}_H^k)^{-1}} + (2L + L_{\boldsymbol{\phi}}\sqrt{md})\zeta \\
&= \left\{ r_H(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_H^k, \boldsymbol{\psi}(s,a,w) \right\rangle + 2L\beta \left\| \boldsymbol{\phi}(s,a) \right\|_{(\boldsymbol{\Lambda}_H^k)^{-1}} \right\}^+ + (2L + L_{\boldsymbol{\phi}}\sqrt{md})\zeta \\
&= Q_H^k(s,a,w) + (2L + L_{\boldsymbol{\phi}}\sqrt{md})\zeta,
\end{aligned}$$

where the first equality follows from the fact that $Q_H^*(s,a,w) \geq 0$. Now, suppose the statement holds at time-step $h+1$ and consider time-step h. Conditioned on events $\{\mathcal{E}_1(w)\}_{w \in \widetilde{\mathcal{W}}}$, for all $(s,a,w,h,k) \in \mathcal{S} \times \mathcal{A} \times \widetilde{\mathcal{W}} \times [H] \times [K]$, we have

$$\begin{aligned}
0 \leq\ & r_h(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle - Q_h^*(s,a,w) - \mathbb{P}_h \left[V_{h+1}^k(.,w) - V_{h+1}^*(.,w) \right](s,a) \\
&+ (2L + L_{\boldsymbol{\phi}}\sqrt{md})\zeta + 2L\beta \left\| \boldsymbol{\phi}(s,a) \right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} \\
&\leq r_h(s,a,w) + \left\langle \hat{\boldsymbol{\xi}}_h^k, \boldsymbol{\psi}(s,a,w) \right\rangle - Q_h^*(s,a,w) + (2L + L_{\boldsymbol{\phi}}\sqrt{md})(H-h+1)\zeta \\
&+ 2L\beta \left\| \boldsymbol{\phi}(s,a) \right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}}. \qquad \text{(Induction assumption)}
\end{aligned}$$

Therefore, conditioned on events $\{\mathcal{E}_1(w)\}_{w\in\widetilde{\mathcal{W}}}$, for all $(s,a,w,h,k)\in\mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[H]\times[K]$, we have

$$\begin{aligned}
Q_h^*(s,a,w) &\le r_h(s,a,w)+\left\langle\hat{\boldsymbol{\xi}}_h^k,\boldsymbol{\psi}(s,a,w)\right\rangle+(2L+L_{\boldsymbol{\phi}}\sqrt{md})(H-h+1)\zeta+2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}}\\
&=\left\{r_h(s,a,w)+\left\langle\hat{\boldsymbol{\xi}}_h^k,\boldsymbol{\psi}(s,a,w)\right\rangle+2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}}\right\}^+\\
&+(2L+L_{\boldsymbol{\phi}}\sqrt{md})(H-h+1)\zeta\\
&=Q_h^k(s,a,w)+(2L+L_{\boldsymbol{\phi}}\sqrt{md})(H-h+1)\zeta,
\end{aligned}$$

where the first equality follows from the fact that $Q_h^*(s,a,w)\ge 0$. This completes the proof.

□

D.6.2 Proof of Theorem 19

The proof for establishing the upper bound on the number of planning calls for misspecified settings follows exactly the steps as those in the proof of Theorem 10.

Now, we prove the regret bound. Let $\delta_h^k=V_h^{\tilde{k}}(s_h^k,w^k)-V_h^{\pi^k}(s_h^k,w^k)$ and $\xi_{h+1}^k=\mathbb{E}\left[\delta_{h+1}^k|s_h^k,a_h^k\right]-\delta_{h+1}^k$. Conditioned on events $\{\mathcal{E}_1(w)\}_{w\in\widetilde{\mathcal{W}}}$, for all $(s,a,w,h,k)\in\mathcal{S}\times\mathcal{A}\times\widetilde{\mathcal{W}}\times[H]\times[K]$, we have

$$\begin{aligned}
Q_h^{\tilde{k}}(s,a,w)-Q_h^{\pi^k}(s,a,w) &= r_h(s,a,w)+\left\langle\hat{\boldsymbol{\xi}}_h^{\tilde{k}},\boldsymbol{\psi}(s,a,w)\right\rangle-Q_h^{\pi^k}(s,a,w)+2L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{(\boldsymbol{\Lambda}_h^{\tilde{k}})^{-1}}\\
&\le\mathbb{P}_h\left[V_{h+1}^{\tilde{k}}(.,w)-V_{h+1}^{\pi^k}(.,w)\right](s,a)+(2L+L_{\boldsymbol{\phi}}\sqrt{md})\zeta\\
&+4L\beta\left\|\boldsymbol{\phi}(s,a)\right\|_{(\boldsymbol{\Lambda}_h^{\tilde{k}})^{-1}}. \qquad \text{(D.50)}
\end{aligned}$$

Note that $\delta_h^k\le Q_h^{\tilde{k}}(s_h^k,a_h^k,w^k)-Q_h^{\pi^k}(s_h^k,a_h^k,w^k)$. Thus, combining (D.50), Lemma 19, and a union bound over $\widetilde{\mathcal{W}}$, we conclude that for all $(h,k)\in[H]\times[K]$, with probability at least $1-\delta$, it holds that gives

$$\delta_h^k\le\xi_{h+1}^k+\delta_{h+1}^k+(2L+L_{\boldsymbol{\phi}}\sqrt{md})\zeta+4L\beta\left\|\boldsymbol{\phi}(s_h^k,a_h^k)\right\|_{(\boldsymbol{\Lambda}_h^{\tilde{k}})^{-1}}.$$

Now, we complete the regret analysis following similar steps as those of Theorem 9's

proof:

$$
\begin{aligned}
R_K &= \sum_{k=1}^{K} V_1^*(s_1^k, w^k) - V_1^{\pi^k}(s_1^k, w^k) \\
&\le (2L + L_{\boldsymbol{\phi}}\sqrt{md})HK\zeta + \sum_{k=1}^{K} V_1^{\tilde{k}}(s_1^k, w^k) - V_1^{\pi^k}(s_1^k, w^k) && \text{(Lemma 31)} \\
&= (2L + L_{\boldsymbol{\phi}}\sqrt{md})HK\zeta + \sum_{k=1}^{K} \delta_1^k \\
&\le (4L + 2L_{\boldsymbol{\phi}}\sqrt{md})HK\zeta + \sum_{k=1}^{K}\sum_{h=1}^{H} \xi_h^k + 4L\beta \sum_{k=1}^{K}\sum_{h=1}^{H} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\left(\boldsymbol{\Lambda}_h^{\tilde{k}}\right)^{-1}} \\
&\le (4L + 2L_{\boldsymbol{\phi}}\sqrt{md})HK\zeta + \sum_{k=1}^{K}\sum_{h=1}^{H} \xi_h^k + 4L\beta \sum_{k=1}^{K}\sum_{h=1}^{H} \left\| \boldsymbol{\phi}(s_h^k, a_h^k) \right\|_{\left(\boldsymbol{\Lambda}_h^{k}\right)^{-1}} \sqrt{\frac{\det \boldsymbol{\Lambda}_h^k}{\det \boldsymbol{\Lambda}_h^{\tilde{k}}}} \\
& && \text{(Eqn. (D.18))} \\
&\le (4L + 2L_{\boldsymbol{\phi}}\sqrt{md})HK\zeta + 2H\sqrt{T\log(dT/\delta)} + 8HL\beta\sqrt{2dK\log(1 + K/\lambda)} \\
&\le \tilde{\mathcal{O}}\left((L + L_{\boldsymbol{\phi}}\sqrt{md})HK\zeta + L\sqrt{\lambda(d^3 + md^2)H^3T}\right),
\end{aligned}
$$

where the last two inequalities follow from the similar steps in the proof of Theorem 9.

D.7 Auxiliary Lemmas

Notations. $\mathcal{N}_\epsilon(\mathcal{V})$ denotes the ϵ-covering number of the class $\mathcal{V}$ of functions mapping $\mathcal{S}$ to $\mathbb{R}$ with respect to the distance $\operatorname{dist}(V, V') = \sup_s \left|V(s) - V'(s)\right|$.

Lemma 32 (Bound on Weights $\boldsymbol{\theta}_h^k(w)$)**.** *Under Assumption 13, for any set of action-value functions $\{Q_h^k\}_{h\in[H]}$, and $(w, h, k) \in \mathcal{W} \times [H] \times [K]$, it holds that*

$$
\left\| \boldsymbol{\theta}_h^k(w) \right\|_2 \le H\sqrt{d}.
$$

Proof. Recall that $V_h^k(s, w) = \min\left\{\max_{a\in\mathcal{A}} Q_h^k(s, a, w), H\right\}$ and $\boldsymbol{\theta}_h^k(w) \coloneqq \int_{\mathcal{S}} V_{h+1}^k(s', w) d\boldsymbol{\mu}_h(s')$. Thus, we have

$$
\left\| \boldsymbol{\theta}_h^k(w) \right\|_2 = \left\| \int_{\mathcal{S}} V_{h+1}^k(s', w) d\boldsymbol{\mu}_h(s') \right\| \le H\sqrt{d}.
$$

□

Lemma 33 (Lemma D.4 in [64]). *Let $\{s_\tau\}_{\tau=1}^{\infty}$ be a stochastic process on state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}_\tau\}_{\tau=0}^{\infty}$. Let $\{\boldsymbol{\phi}_\tau\}_{\tau=0}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process where $\boldsymbol{\phi}_\tau \in \mathcal{F}_{\tau-1}$, and $\|\boldsymbol{\phi}_\tau\| \leq 1$. Let $\boldsymbol{\Lambda}_k = \lambda \mathbf{I}_d + \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_\tau \boldsymbol{\phi}_\tau^\top$. Then with probability at least $1-\delta$, for all $k \geq 0$ and $V \in \mathcal{V}$ such that $\sup_{s\in\mathcal{S}} \big|V(s)\big| \leq H$, we have*

$$\left\| \sum_{\tau=1}^{k} \boldsymbol{\phi}_\tau \cdot \Big(V(s_\tau) - \mathbb{E}\left[V(s_\tau) | \mathcal{F}_{\tau-1} \right] \Big) \right\|_{\boldsymbol{\Lambda}_k^{-1}}^2 \leq 4H^2 \left(\frac{d}{2} \log\left(\frac{k+\lambda}{\lambda} \right) + \log\left(\frac{\mathcal{N}_\epsilon(\mathcal{V})}{\delta} \right) \right) + \frac{8k^2\epsilon^2}{\lambda}.$$

Lemma 34. *For any $\epsilon > 0$, the ϵ-covering number of the Euclidean ball in $\mathbb{R}^d$ with radius $R > 0$ is upper bounded by $(1 + 2R/\epsilon)^d$.*

Lemma 35. *For a fixed w, let $\mathcal{V}$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with following parametric form*

$$V(.) = \min\left\{ \max_{a\in\mathcal{A}} \big\langle \mathbf{z}, \boldsymbol{\psi}(., a, w) \big\rangle + \big\langle \mathbf{y}, \boldsymbol{\phi}(., a) \big\rangle + \beta \sqrt{\boldsymbol{\phi}(., a)^\top \mathbf{Y} \boldsymbol{\phi}(., a)}, H \right\},$$

where the parameters $\beta \in \mathbb{R}$, $\mathbf{z} \in \mathbb{R}^{d'}$, $\mathbf{y} \in \mathbb{R}^d$, and $\mathbf{Y} \in \mathbb{R}^{d\times d}$ satisfy $0 \leq \beta \leq B$, $\|\mathbf{z}\| \leq z$, $\|\mathbf{y}\| \leq y$, and $\|\mathbf{Y}\| \leq \lambda^{-1}$. Assume $\big\|\boldsymbol{\phi}(s,a)\big\| \leq 1$ and $\big\|\boldsymbol{\psi}(s,a,w)\big\| \leq 1$ for all $(s,a,w) \in \mathcal{S}\times\mathcal{A}\times\mathcal{W}$. Then

$$\log\big(\mathcal{N}_\epsilon(\mathcal{V})\big) \leq d' \log(1 + 4z/\epsilon) + d \log(1 + 4y/\epsilon) + d^2 \log\left(\frac{1 + 8B^2\sqrt{d}}{\lambda\epsilon^2} \right).$$

Proof. First, we reparametrize $\mathcal{V}$ by letting $\tilde{\mathbf{Y}} = \beta^2 \mathbf{Y}$. We have

$$V(.) = \min\left\{ \max_{a\in\mathcal{A}} \big\langle \mathbf{z}, \boldsymbol{\psi}(., a, w) \big\rangle + \big\langle \mathbf{y}, \boldsymbol{\phi}(., a) \big\rangle + \sqrt{\boldsymbol{\phi}(., a)^\top \tilde{\mathbf{Y}} \boldsymbol{\phi}(., a)}, H \right\},$$

for $\|\mathbf{z}\| \leq z$, $\|\mathbf{y}\| \leq y$, and $\left\|\tilde{\mathbf{Y}}\right\| \leq \frac{B^2}{\lambda}$. For any two functions $V_1, V_2 \in \mathcal{V}$ with parameters

$\left(\mathbf{z}^1, \mathbf{y}^1, \tilde{\mathbf{Y}}^1\right)$ and $\left(\mathbf{z}^2, \mathbf{y}^2, \tilde{\mathbf{Y}}^2\right)$, respectively, we have

$$\begin{aligned}
\operatorname{dist}(V_1, V_2) \leq & \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| \left[\left\langle \mathbf{z}^1, \boldsymbol{\psi}(s,a,w) \right\rangle + \left\langle \mathbf{y}^1, \boldsymbol{\phi}(s,a) \right\rangle + \sqrt{\boldsymbol{\phi}(s,a)^\top \tilde{\mathbf{Y}}^1 \boldsymbol{\phi}(s,a)} \right] \right. \\
& \left. - \left[\left\langle \mathbf{z}^2, \boldsymbol{\psi}(s,a,w) \right\rangle + \left\langle \mathbf{y}^2, \boldsymbol{\phi}(s,a) \right\rangle + \sqrt{\boldsymbol{\phi}(s,a)^\top \tilde{\mathbf{Y}}^2 \boldsymbol{\phi}(s,a)} \right] \right| \\
\leq & \sup_{\boldsymbol{\psi}:\|\boldsymbol{\psi}\|\leq 1, \boldsymbol{\phi}:\|\boldsymbol{\phi}\|\leq 1} \left| \left[\left\langle \mathbf{z}^1, \boldsymbol{\psi} \right\rangle + \left\langle \mathbf{y}^1, \boldsymbol{\phi} \right\rangle + \sqrt{\boldsymbol{\phi}^\top \tilde{\mathbf{Y}}^1 \boldsymbol{\phi}} \right] - \left[\left\langle \mathbf{z}^2, \boldsymbol{\psi} \right\rangle + \left\langle \mathbf{y}^2, \boldsymbol{\phi} \right\rangle + \sqrt{\boldsymbol{\phi}^\top \tilde{\mathbf{Y}}^2 \boldsymbol{\phi}} \right] \right| \\
\leq & \sup_{\boldsymbol{\psi}:\|\boldsymbol{\psi}\|\leq 1} \left| \left\langle \mathbf{z}^1 - \mathbf{z}^2, \boldsymbol{\psi} \right\rangle \right| + \sup_{\boldsymbol{\phi}:\|\boldsymbol{\phi}\|\leq 1} \left| \left\langle \mathbf{y}^1 - \mathbf{y}^2, \boldsymbol{\phi} \right\rangle \right| + \sup_{\boldsymbol{\phi}:\|\boldsymbol{\phi}\|\leq 1} \sqrt{\left| \boldsymbol{\phi}^\top \left(\tilde{\mathbf{Y}}^1 - \tilde{\mathbf{Y}}^2 \right) \boldsymbol{\phi} \right|} \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(because } \left| \sqrt{a} - \sqrt{b} \right| \leq \sqrt{|a-b|} \text{ for } a, b \geq 0) \\
= & \left\| \mathbf{z}^1 - \mathbf{z}^2 \right\| + \left\| \mathbf{y}^1 - \mathbf{y}^2 \right\| + \sqrt{\left\| \tilde{\mathbf{Y}}^1 - \tilde{\mathbf{Y}}^2 \right\|} \\
\leq & \left\| \mathbf{z}^1 - \mathbf{z}^2 \right\| + \left\| \mathbf{y}^1 - \mathbf{y}^2 \right\| + \sqrt{\left\| \tilde{\mathbf{Y}}^1 - \tilde{\mathbf{Y}}^2 \right\|_F}. \qquad \text{(D.51)}
\end{aligned}$$

Let $\mathcal{C}_{\mathbf{z}}$ and $\mathcal{C}_{\mathbf{y}}$ be $\epsilon/2$-covers of $\{\mathbf{z} \in \mathbb{R}^{d'} : \|\mathbf{z}\| \leq z\}$ and $\{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{y}\| \leq y\}$, respectively, with respect to the 2-norm, and $\mathcal{C}_{\mathbf{Y}}$ be an $\epsilon^2/4$-cover of $\{\mathbf{Y} \in \mathbb{R}^{d\times d} : \|\mathbf{Y}\|_F \leq \frac{B^2\sqrt{d}}{\lambda}\}$, with respect to the Frobenius norm. By Lemma 34, we know

$$|\mathcal{C}_{\mathbf{z}}| \leq (1 + 4z/\epsilon)^{d'}, \quad \left|\mathcal{C}_{\mathbf{y}}\right| \leq (1 + 4y/\epsilon)^d, \quad |\mathcal{C}_{\mathbf{Y}}| \leq \left(\frac{1 + 8B^2\sqrt{d}}{\lambda\epsilon^2} \right)^{d^2}.$$

According to (D.51), it holds that $\mathcal{N}_\epsilon(\mathcal{V}) \leq |\mathcal{C}_{\mathbf{z}}| \left|\mathcal{C}_{\mathbf{y}}\right| |\mathcal{C}_{\mathbf{Y}}|$, and therefore

$$\log\left(\mathcal{N}_\epsilon(\mathcal{V})\right) \leq d' \log(1 + 4z/\epsilon) + d \log(1 + 4y/\epsilon) + d^2 \log\left(\frac{1 + 8B^2\sqrt{d}}{\lambda\epsilon^2} \right).$$

□

Lemma 36. *For a fixed w, let $\mathcal{V}$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with following parametric form*

$$V(.) = \min\left\{ \max_{a\in\mathcal{A}} \left\{ \left\langle \mathbf{z}, \boldsymbol{\psi}(., a, w) \right\rangle + 2L\beta\sqrt{\boldsymbol{\phi}(., a)^\top \mathbf{Y} \boldsymbol{\phi}(., a)} \right\}^+, H \right\},$$

where the parameters $\beta \in \mathbb{R}$, $\mathbf{z} \in \mathbb{R}^{d'}$ and $\mathbf{Y} \in \mathbb{R}^{d\times d}$ satisfy $0 \leq \beta \leq B$, $\|\mathbf{z}\| \leq z$, and $\|\mathbf{Y}\| \leq \lambda^{-1}$. Assume $\left\|\boldsymbol{\phi}(s,a)\right\| \leq 1$ and $\left\|\boldsymbol{\psi}(s,a,w)\right\| \leq 1$ for all $(s,a,w) \in \mathcal{S}\times\mathcal{A}\times\mathcal{W}$. Then

$$\log\left(\mathcal{N}_\epsilon(\mathcal{V})\right) \leq d' \log(1 + 4z/\epsilon) + d^2 \log\left(\frac{1 + 8B^2\sqrt{d}}{\lambda\epsilon^2} \right).$$

Proof. First, we reparametrize $\mathcal{V}$ by letting $\tilde{\mathbf{Y}} = \beta^2 \mathbf{Y}$. We have

$$V(.) = \min\left\{\max_{a\in\mathcal{A}} \left\langle \mathbf{z}, \boldsymbol{\psi}(.,a,w)\right\rangle + \sqrt{\boldsymbol{\phi}(.,a)^\top \tilde{\mathbf{Y}} \boldsymbol{\phi}(.,a)}, H\right\},$$

for $\|\mathbf{z}\| \le z$, and $\left\|\tilde{\mathbf{Y}}\right\| \le \frac{B^2}{\lambda}$. For any two functions $V_1, V_2 \in \mathcal{V}$ with parameters $\left(\mathbf{z}^1, \tilde{\mathbf{Y}}^1\right)$ and $\left(\mathbf{z}^2, \tilde{\mathbf{Y}}^2\right)$, respectively, we have

$$\begin{aligned}
\mathrm{dist}(V_1, V_2) &\le \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| \left[\left\langle \mathbf{z}^1, \boldsymbol{\psi}(s,a,w)\right\rangle + \sqrt{\boldsymbol{\phi}(s,a)^\top \tilde{\mathbf{Y}}^1 \boldsymbol{\phi}(s,a)}\right] - \left[\left\langle \mathbf{z}^2, \boldsymbol{\psi}(s,a,w)\right\rangle + \sqrt{\boldsymbol{\phi}(s,a)^\top \tilde{\mathbf{Y}}^2 \boldsymbol{\phi}(s,a)}\right] \right| \\
&\le \sup_{\boldsymbol{\psi}:\|\boldsymbol{\psi}\|\le 1, \boldsymbol{\phi}:\|\boldsymbol{\phi}\|\le 1} \left| \left[\left\langle \mathbf{z}^1, \boldsymbol{\psi}\right\rangle + \sqrt{\boldsymbol{\phi}^\top \tilde{\mathbf{Y}}^1 \boldsymbol{\phi}}\right] - \left[\left\langle \mathbf{z}^2, \boldsymbol{\psi}\right\rangle + \sqrt{\boldsymbol{\phi}^\top \tilde{\mathbf{Y}}^2 \boldsymbol{\phi}}\right] \right| \\
&\le \sup_{\boldsymbol{\psi}:\|\boldsymbol{\psi}\|\le 1} \left| \left\langle \mathbf{z}^1 - \mathbf{z}^2, \boldsymbol{\psi}\right\rangle \right| + \sup_{\boldsymbol{\phi}:\|\boldsymbol{\phi}\|\le 1} \sqrt{\left| \boldsymbol{\phi}^\top \left(\tilde{\mathbf{Y}}^1 - \tilde{\mathbf{Y}}^2\right) \boldsymbol{\phi} \right|} \\
&\qquad\qquad \text{(because } \left|\sqrt{a} - \sqrt{b}\right| \le \sqrt{|a-b|} \text{ for } a, b \ge 0) \\
&= \left\|\mathbf{z}^1 - \mathbf{z}^2\right\| + \sqrt{\left\|\tilde{\mathbf{Y}}^1 - \tilde{\mathbf{Y}}^2\right\|} \\
&\le \left\|\mathbf{z}^1 - \mathbf{z}^2\right\| + \sqrt{\left\|\tilde{\mathbf{Y}}^1 - \tilde{\mathbf{Y}}^2\right\|_F}. \qquad \text{(D.52)}
\end{aligned}$$

Let $\mathcal{C}_{\mathbf{z}}$ be an $\epsilon/2$-cover of $\{\mathbf{z} \in \mathbb{R}^{d'} : \|\mathbf{z}\| \le z\}$ with respect to the 2-norm, and $\mathcal{C}_{\mathbf{Y}}$ be an $\epsilon^2/4$-cover of $\{\mathbf{Y} \in \mathbb{R}^{d\times d} : \|\mathbf{Y}\|_F \le \frac{B^2\sqrt{d}}{\lambda}\}$, with respect to the Frobenius norm. By Lemma 34, we know

$$|\mathcal{C}_{\mathbf{z}}| \le (1 + 4z/\epsilon)^{d'}, \quad |\mathcal{C}_{\mathbf{Y}}| \le \left(\frac{1 + 8B^2\sqrt{d}}{\lambda\epsilon^2}\right)^{d^2}.$$

According to (D.52), it holds that $\mathcal{N}_\epsilon(\mathcal{V}) \le |\mathcal{C}_{\mathbf{z}}||\mathcal{C}_{\mathbf{Y}}|$, and therefore

$$\log\left(\mathcal{N}_\epsilon(\mathcal{V})\right) \le d' \log(1 + 4z/\epsilon) + d^2 \log\left(\frac{1 + 8B^2\sqrt{d}}{\lambda\epsilon^2}\right).$$

□

Lemma 37. *For a fixed w, let $\mathcal{V}$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with following parametric form*

$$V(.) = \min\left\{\max_{a\in\mathcal{A}} \left\{\left\langle \mathbf{z}, \boldsymbol{\psi}(.,a,w)\right\rangle + 2L\beta\sqrt{\boldsymbol{\phi}(.,a)^\top \mathbf{Y} \boldsymbol{\phi}(.,a)} + \tilde{\beta}\sqrt{\boldsymbol{\phi}(.,a,w)^\top \tilde{\mathbf{Y}} \boldsymbol{\phi}(.,a,w)}\right\}^+, H\right\},$$

where the parameters $\beta, \tilde{\beta} \in \mathbb{R}$, $\mathbf{z} \in \mathbb{R}^{d'}$, $\mathbf{Y} \in \mathbb{R}^{d\times d}$ and $\tilde{\mathbf{Y}} \in \mathbb{R}^{d'\times d'}$ satisfy $0 \le \beta \le B$, $0 \le \tilde{\beta} \le \tilde{B}$ $\|\mathbf{z}\| \le z$, $\|\mathbf{Y}\| \le \lambda^{-1}$ and $\left\|\tilde{\mathbf{Y}}\right\| \le \lambda^{-1}$. Assume $\left\|\boldsymbol{\phi}(s,a)\right\| \le 1$ and $\left\|\boldsymbol{\psi}(s,a,w)\right\| \le 1$

for all $(s,a,w) \in \mathcal{S}\times\mathcal{A}\times\mathcal{W}$. *Then*

$$\log\left(\mathcal{N}_\epsilon(\mathcal{V})\right) \le d'\log(1+4z/\epsilon) + d^2\log\left(\frac{1+8B^2\sqrt{d}}{\lambda\epsilon^2}\right) + d'^2\log\left(\frac{1+8\tilde{B}^2\sqrt{d'}}{\lambda\epsilon^2}\right).$$

Proof. First, we reparametrize $\mathcal{V}$ by letting $\mathbf{Z} = \beta^2\mathbf{Y}$ and $\tilde{\mathbf{Z}} = \tilde{\beta}^2\tilde{\mathbf{Y}}$. We have

$$V(.) = \min\left\{\max_{a\in\mathcal{A}}\left\langle \mathbf{z}, \boldsymbol{\psi}(.,a,w)\right\rangle + \sqrt{\boldsymbol{\phi}(.,a)^\top\mathbf{Z}\boldsymbol{\phi}(.,a)} + \sqrt{\boldsymbol{\phi}(.,a)^\top\tilde{\mathbf{Z}}\boldsymbol{\phi}(.,a)}, H\right\},$$

for $\|\mathbf{z}\| \le z$, $\|\mathbf{Z}\| \le \frac{B^2}{\lambda}$, and $\left\|\tilde{\mathbf{Z}}\right\| \le \frac{\tilde{B}^2}{\lambda}$. For any two functions $V_1, V_2 \in \mathcal{V}$ with parameters $\left(\mathbf{z}^1, \mathbf{Z}^1, \tilde{\mathbf{Z}}^1\right)$ and $\left(\mathbf{z}^2, \mathbf{Z}^2, \tilde{\mathbf{Z}}^2\right)$, respectively, we have

$$\begin{aligned}
\mathrm{dist}(V_1,V_2) \le & \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|\left[\left\langle \mathbf{z}^1, \boldsymbol{\psi}(s,a,w)\right\rangle + \sqrt{\boldsymbol{\phi}(s,a)^\top\mathbf{Z}^1\boldsymbol{\phi}(s,a)} + \sqrt{\boldsymbol{\psi}(s,a,w)^\top\tilde{\mathbf{Z}}^1\boldsymbol{\psi}(s,a,w)}\right]\right. \\
& \left. - \left[\left\langle \mathbf{z}^2, \boldsymbol{\psi}(s,a,w)\right\rangle + \sqrt{\boldsymbol{\phi}(s,a)^\top\mathbf{Z}^2\boldsymbol{\phi}(s,a)} + \sqrt{\boldsymbol{\psi}(s,a,w)^\top\tilde{\mathbf{Z}}^2\boldsymbol{\psi}(s,a,w)}\right]\right| \\
\le & \sup_{\boldsymbol{\psi}:\|\boldsymbol{\psi}\|\le 1, \boldsymbol{\phi}:\|\boldsymbol{\phi}\|\le 1}\left|\left[\left\langle \mathbf{z}^1, \boldsymbol{\psi}\right\rangle + \sqrt{\boldsymbol{\phi}^\top\mathbf{Z}^1\boldsymbol{\phi}} + \sqrt{\boldsymbol{\psi}^\top\tilde{\mathbf{Z}}^1\boldsymbol{\psi}}\right] - \left[\left\langle \mathbf{z}^2, \boldsymbol{\psi}\right\rangle + \sqrt{\boldsymbol{\phi}^\top\mathbf{Z}^2\boldsymbol{\phi}} + \sqrt{\boldsymbol{\psi}^\top\tilde{\mathbf{Z}}^2\boldsymbol{\psi}}\right]\right| \\
\le & \sup_{\boldsymbol{\psi}:\|\boldsymbol{\psi}\|\le 1}\left|\left\langle \mathbf{z}^1 - \mathbf{z}^2, \boldsymbol{\psi}\right\rangle\right| + \sup_{\boldsymbol{\phi}:\|\boldsymbol{\phi}\|\le 1}\sqrt{\left|\boldsymbol{\phi}^\top\left(\mathbf{Z}^1 - \mathbf{Z}^2\right)\boldsymbol{\phi}\right|} + \sup_{\boldsymbol{\psi}:\|\boldsymbol{\phi}\|\le 1}\sqrt{\left|\boldsymbol{\psi}^\top\left(\tilde{\mathbf{Z}}^1 - \tilde{\mathbf{Z}}^2\right)\boldsymbol{\psi}\right|} \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left(\text{because } \left|\sqrt{a}-\sqrt{b}\right| \le \sqrt{|a-b|} \text{ for } a,b\ge 0\right) \\
= & \left\|\mathbf{z}^1 - \mathbf{z}^2\right\| + \sqrt{\left\|\mathbf{Z}^1 - \mathbf{Z}^2\right\|} + \sqrt{\left\|\tilde{\mathbf{Z}}^1 - \tilde{\mathbf{Z}}^2\right\|} \\
\le & \left\|\mathbf{z}^1 - \mathbf{z}^2\right\| + \sqrt{\left\|\mathbf{Z}^1 - \mathbf{Z}^2\right\|_F} + \sqrt{\left\|\tilde{\mathbf{Z}}^1 - \tilde{\mathbf{Z}}^2\right\|_F}. \qquad \text{(D.53)}
\end{aligned}$$

Let $\mathcal{C}_{\mathbf{z}}$ be an $\epsilon/2$-cover of $\{\mathbf{z}\in\mathbb{R}^{d'} : \|\mathbf{z}\| \le z\}$ with respect to the 2-norm, $\mathcal{C}_{\mathbf{Z}}$ be an $\epsilon^2/4$-cover of $\{\mathbf{Z}\in\mathbb{R}^{d\times d} : \|\mathbf{Z}\|_F \le \frac{B^2\sqrt{d}}{\lambda}\}$, and $\mathcal{C}_{\tilde{\mathbf{Z}}}$ be an $\epsilon^2/4$-cover of $\{\tilde{\mathbf{Z}}\in\mathbb{R}^{d'\times d'} : \left\|\tilde{\mathbf{Z}}\right\|_F \le \frac{\tilde{B}^2\sqrt{d}}{\lambda}\}$ with respect to the Frobenius norm. By Lemma 34, we know

$$|\mathcal{C}_{\mathbf{z}}| \le (1+4z/\epsilon)^{d'}, \quad |\mathcal{C}_{\mathbf{Z}}| \le \left(\frac{1+8B^2\sqrt{d}}{\lambda\epsilon^2}\right)^{d^2}, \quad \left|\mathcal{C}_{\tilde{\mathbf{Z}}}\right| \le \left(\frac{1+8\tilde{B}^2\sqrt{d'}}{\lambda\epsilon^2}\right)^{d'^2}.$$

According to (D.53), it holds that $\mathcal{N}_\epsilon(\mathcal{V}) \le |\mathcal{C}_{\mathbf{z}}||\mathcal{C}_{\mathbf{Y}}|$, and therefore

$$\log\left(\mathcal{N}_\epsilon(\mathcal{V})\right) \le d'\log(1+4z/\epsilon) + d^2\log\left(\frac{1+8B^2\sqrt{d}}{\lambda\epsilon^2}\right) + d'^2\log\left(\frac{1+8\tilde{B}^2\sqrt{d'}}{\lambda\epsilon^2}\right).$$

□

Lemma 38. *Let $\mathcal{V}$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with following parametric form*

$$V(.,.) = \min\left\{\max_{a\in\mathcal{A}}\left\{\left\langle \mathbf{z}, \boldsymbol{\psi}(.,a,.)\right\rangle + 2L\beta\sqrt{\boldsymbol{\psi}(.,a,.)^\top \mathbf{Y}\boldsymbol{\psi}(.,a,.)}\right\}^+, H\right\},$$

where the parameters $\beta \in \mathbb{R}$, $\mathbf{z} \in \mathbb{R}^{d'}$ and $\mathbf{Y} \in \mathbb{R}^{d'\times d'}$ satisfy $0 \le \beta \le B$, $\|\mathbf{z}\| \le z$, and $\|\mathbf{Y}\| \le \lambda^{-1}$. Assume $\left\|\boldsymbol{\psi}(s,a,w)\right\| \le 1$ for all $(s,a,w) \in \mathcal{S}\times\mathcal{A}\times\mathcal{W}$. Then

$$\log\left(\mathcal{N}_\epsilon(\mathcal{V})\right) \le d'\log(1+4z/\epsilon) + {d'}^2\log\left(\frac{1+8B^2\sqrt{d'}}{\lambda\epsilon^2}\right).$$

Proof. First, we reparametrize $\mathcal{V}$ by letting $\tilde{\mathbf{Y}} = \beta^2\mathbf{Y}$. We have

$$V(.,.) = \min\left\{\max_{a\in\mathcal{A}}\left\langle \mathbf{z}, \boldsymbol{\psi}(.,a,.)\right\rangle + \sqrt{\boldsymbol{\psi}(.,a,.)^\top\tilde{\mathbf{Y}}\boldsymbol{\psi}(.,a,.)}), H\right\},$$

for $\|\mathbf{z}\| \le z$, and $\left\|\tilde{\mathbf{Y}}\right\| \le \frac{B^2}{\lambda}$. For any two functions $V_1, V_2 \in \mathcal{V}$ with parameters $\left(\mathbf{z}^1, \tilde{\mathbf{Y}}^1\right)$ and $\left(\mathbf{z}^2, \tilde{\mathbf{Y}}^2\right)$, respectively, we have

$$\begin{aligned}
\operatorname{dist}(V_1,V_2) \le & \sup_{(s,a,w)\in\mathcal{S}\times\mathcal{A}\times\mathcal{W}}\left|\left[\left\langle \mathbf{z}^1, \boldsymbol{\psi}(s,a,w)\right\rangle + \sqrt{\boldsymbol{\psi}(s,a)^\top\tilde{\mathbf{Y}}^1\boldsymbol{\psi}(s,a)}\right]\right. \\
& \left. - \left[\left\langle \mathbf{z}^2, \boldsymbol{\psi}(s,a,w)\right\rangle + \sqrt{\boldsymbol{\psi}(s,a,w)^\top\tilde{\mathbf{Y}}^2\boldsymbol{\psi}(s,a,w)}\right]\right| \\
\le & \sup_{\boldsymbol{\psi}:\|\boldsymbol{\psi}\|\le 1}\left|\left[\left\langle \mathbf{z}^1, \boldsymbol{\psi}\right\rangle + \sqrt{\boldsymbol{\psi}^\top\tilde{\mathbf{Y}}^1\boldsymbol{\psi}}\right] - \left[\left\langle \mathbf{z}^2, \boldsymbol{\psi}\right\rangle + \sqrt{\boldsymbol{\psi}^\top\tilde{\mathbf{Y}}^2\boldsymbol{\psi}}\right]\right| \\
\le & \sup_{\boldsymbol{\psi}:\|\boldsymbol{\psi}\|\le 1}\left|\left\langle \mathbf{z}^1 - \mathbf{z}^2, \boldsymbol{\psi}\right\rangle\right| + \sup_{\boldsymbol{\psi}:\|\boldsymbol{\psi}\|\le 1}\sqrt{\left|\boldsymbol{\psi}^\top\left(\tilde{\mathbf{Y}}^1 - \tilde{\mathbf{Y}}^2\right)\boldsymbol{\psi}\right|} \\
& \qquad\qquad (\text{because } \left|\sqrt{a}-\sqrt{b}\right| \le \sqrt{|a-b|} \text{ for } a,b\ge 0) \\
= & \left\|\mathbf{z}^1 - \mathbf{z}^2\right\| + \sqrt{\left\|\tilde{\mathbf{Y}}^1 - \tilde{\mathbf{Y}}^2\right\|} \\
\le & \left\|\mathbf{z}^1 - \mathbf{z}^2\right\| + \sqrt{\left\|\tilde{\mathbf{Y}}^1 - \tilde{\mathbf{Y}}^2\right\|_F}. \qquad \text{(D.54)}
\end{aligned}$$

Let $\mathcal{C}_\mathbf{z}$ be an $\epsilon/2$-cover of $\{\mathbf{z}\in\mathbb{R}^{d'} : \|\mathbf{z}\| \le z\}$ with respect to the 2-norm, and $\mathcal{C}_\mathbf{Y}$ be an $\epsilon^2/4$-cover of $\{\mathbf{Y}\in\mathbb{R}^{d'\times d'} : \|\mathbf{Y}\|_F \le \frac{B^2\sqrt{d'}}{\lambda}\}$, with respect to the Frobenius norm. By Lemma 34, we know

$$|\mathcal{C}_\mathbf{z}| \le (1+4z/\epsilon)^{d'}, \quad |\mathcal{C}_\mathbf{Y}| \le \left(\frac{1+8B^2\sqrt{d'}}{\lambda\epsilon^2}\right)^{{d'}^2}.$$

According to (D.54), it holds that $\mathcal{N}_\epsilon(\mathcal{V}) \leq |\mathcal{C}_{\mathbf{z}}||\mathcal{C}_{\mathbf{Y}}|$, and therefore

$$\log\left(\mathcal{N}_\epsilon(\mathcal{V})\right) \leq d' \log(1 + 4z/\epsilon) + {d'}^2 \log\left(\frac{1 + 8B^2\sqrt{d'}}{\lambda\epsilon^2}\right).$$

□

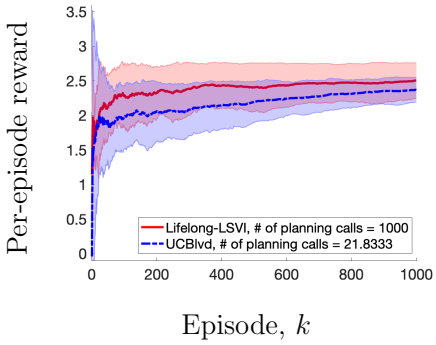
D.8 Details of the Experiments

In all the experiments, we have chosen $\delta = 0.01$, $\lambda = 1$, $d = 5$, and $H = 5$. The parameters $\{\boldsymbol{\eta}_h\}_{h\in[H]}$ are drawn from $\mathcal{N}(0, I_{d'})$. In order to tune parameters $\{\boldsymbol{\mu}_h(.)\}_{h\in[H]}$ and the feature mappings $\boldsymbol{\phi}$ such that they are compatible with Assumption 1, we consider that the feature space $\{\boldsymbol{\phi}(s,a) : (s,a) \in \mathcal{S} \times \mathcal{A}\}$ is a subset of the d-dimensional simplex, $\{\boldsymbol{\phi} \in \mathbb{R}^d : \sum_{i=1}^d \boldsymbol{\phi}_i = 1, \boldsymbol{\phi}_i \geq 0, \boldsymbol{\phi}_i \leq 1, \forall i \in [d]\}$, and $\mathbf{e}_i^\top \boldsymbol{\mu}_h(.)$ is an arbitrary probability measure over $\mathcal{S}$ for all $i \in [d]$.

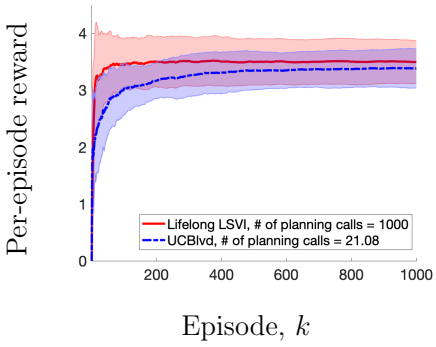
The results shown in Figure D.1a depict averages over 50 realizations for the main setup considered throughout the chapter with $m = 5$ and the results shown in Figure D.1b depict averages over 50 realizations, for the more general setup of Remark 3 with $d' = 10$. For the results shown in Figure D.1a, the mappings $\boldsymbol{\rho}(w)$ are drawn from $\mathcal{N}(0, I_m)$ except for the $n = m$ representative tasks $\{w^{(j)}\}_{j\in[m]}$ introduced in Assumption 15, for which we set $\boldsymbol{\rho}(w^{(j)}) = \mathbf{e}_j$ for $j \in [m]$. For the results shown in Figure D.1b, the mappings $\boldsymbol{\psi}(s,a,w)$ are drawn from $\mathcal{N}(0, I_{d'})$ and we set $\boldsymbol{\psi}(s,a,w^{(j)}) = \mathbf{e}_j$ for $j \in [d']$, where $\{w^{(j)}\}_{j\in[d']}$ are $n = d'$ representative tasks introduced in Assumption 20 in Appendix D.4. The parameters $\{\boldsymbol{\eta}_h\}_{h\in[H]}$ are drawn from $\mathcal{N}(0, I_{d'})$, where $d' = m \times d = 25$ in Figure D.1a. In our experiments, the exact same settings are used for both UCBlvd and Lifelong-LSVI in both Figures D.1a and D.1b. We chose fairly large d, m, and d' and by checking online, we noticed that the optimal value of QCQP in (5.8) happens always to be zero. All these together suggest that the assumptions made in the chapter approximately hold. Figures D.1a and D.1b depict the average per-episode reward of UCBlvd and state the average number of planning calls and compare them to those of baseline algorithm Lifelong-LSVI, a direct extension of LSVI-UCB in [64]. The results emphasize the value of UCBlvd in terms of requiring much smaller

numbers of planning calls. The plots verify that the performances of Lifelong-LSVI and UCBlvd are almost the same statistically, while UCBlvd uses much smaller numbers of planning calls (1000 vs ~ 20).

In Figure D.2, we plot UCBlvd's number of planning calls for different number of task episodes, K, while the setting is same as that in D.1a. In this figure, we empirically verify the logarithmic dependence of number of planning calls on K as suggested by Theorem 10.



(a) Setting of Theorem 10, $d = 5$, $m = 5$, $d' = 25$



(b) Setting of Remark 3, $d = 5$, $d' = 10$

Figure D.1: UCBlvd vs Lifelong-LSVI

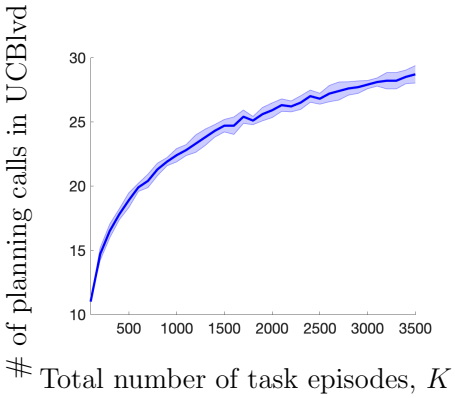


Figure D.2: Setting of Theorem 10, $d = 5$, $m = 5$, $d' = 25$

APPENDIX E

Proofs for Chapter 6

E.1 Proof of Lemma 8

Let $\boldsymbol{\mu} \sim \text{Unif}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$, where $\boldsymbol{\mu}_1 = [\Delta, 0]^\top, \boldsymbol{\mu}_2 = [-\Delta, 0]^\top$, $\mathbf{z} = \{z_t\}_{t=1}^T$ be the set of arm 1's reward, $H = \{a_t, y_t\}_{t=1}^T$ be the history over the course of T rounds, where a_t is the arm pulled and y_t is the observed reward at round t, $a_* = \arg\max_{a \in \{1,2\}} \boldsymbol{\mu}_a$, and $\hat{a} \sim \text{Unif}(\{a_1, a_2, \ldots, a_T\})$. We have

$$\begin{aligned} BR_T(\pi) = \mathbb{E}[R_T(\pi, \boldsymbol{\mu})] &= \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}(\hat{a} \neq a_*)\Delta] \\ &= \Delta T \mathbb{P}(\hat{a} \neq a_*). \qquad (\bigstar) \end{aligned}$$

Now, we lower bound $\mathbb{P}(\hat{a} \neq a_*)$ as follows

$$
\begin{aligned}
\mathbb{P}(\hat{a} \neq a_*) &= \sum_{a \in \{1,2\}} \mathbb{P}(a_* = a)\mathbb{P}(\hat{a} \neq a | a_* = a) \\
&= \sum_{a \in \{1,2\}} \mathbb{P}(a_* = a) \left[\mathbb{P}(\hat{a} \neq a) + \mathbb{P}(\hat{a} = a) - \mathbb{P}(\hat{a} = a | a_* = a)\right] \\
&\geq \sum_{a \in \{1,2\}} \mathbb{P}(a_* = a) \left[\mathbb{P}(\hat{a} \neq a) - \sqrt{\frac{1}{2}\mathrm{D}_{KL}(\mathbb{P}_{\hat{a}|a_*=a}, \mathbb{P}_{\hat{a}})}\right] \qquad \text{(Pinsker's inequality)} \\
&= \frac{1}{2} - \sum_{a \in \{1,2\}} \mathbb{P}(a_* = a)\sqrt{\frac{1}{2}\mathrm{D}_{KL}(\mathbb{P}_{\hat{a}|a_*=a}, \mathbb{P}_{\hat{a}})} \\
&\geq \frac{1}{2} - \sqrt{\frac{1}{2}\sum_{a \in \{1,2\}} \mathbb{P}(a_* = a)\mathrm{D}_{KL}(\mathbb{P}_{\hat{a}|a_*=a}, \mathbb{P}_{\hat{a}})} \qquad \text{(Jensen's inequality)} \\
&= \frac{1}{2} - \sqrt{\frac{1}{2}I(\hat{a}; a_*)} \\
&\geq \frac{1}{2} - \sqrt{\frac{1}{2}I(M, H; a_*)} \qquad \text{(Data processing)} \\
&\geq \frac{1}{2} - \sqrt{\frac{1}{2}\left(I(M; a_*) + I(H; a_*)\right)} \\
&\geq \frac{1}{2} - \sqrt{\frac{1}{2}\left(\frac{1}{16} + I(H; a_*)\right)}. \qquad (\bigstar\bigstar)
\end{aligned}
$$

In our next step towards lower bounding $\mathbb{P}(\hat{a} \neq a_*)$, we upper bound $I(H; a_*)$, as follows

$$
\begin{aligned}
I(H; a_*) &\leq I(\mathbf{z}; a_*) \qquad \text{(Data processing)} \\
&= \sum_{a \in \{1,2\}} \frac{1}{2}\mathrm{D}_{KL}(\mathbb{P}(\mathbf{z}|a_* = a), \mathbb{P}(\mathbf{z})) \\
&\leq \sum_{b \in \{1,2\}} \sum_{a \in \{1,2\}} \frac{1}{2}\mathrm{D}_{KL}\left(\mathbb{P}(\mathbf{z}|a_* = a), \mathbb{P}(\mathbf{z}|a_* = b)\right) \\
&= \frac{1}{2}\mathrm{D}_{KL}\left(\mathbb{P}(\mathbf{z}|a_* = 1), \mathbb{P}(\mathbf{z}|a_* = 2)\right) + \frac{1}{2}\mathrm{D}_{KL}\left(\mathbb{P}(\mathbf{z}|a_* = 2), \mathbb{P}(\mathbf{z}|a_* = 1)\right) \\
&= \frac{1}{2}\left[T(2\Delta)^2 + T(2\Delta)^2\right] \\
&= 4T\Delta^2. \qquad (\bigstar\bigstar\bigstar)
\end{aligned}
$$

Combining ★, ★★, and ★★★, we have

$$BR_T(\pi) \geq \Delta T \left(\frac{1}{2} - \sqrt{\frac{1}{2}\left(\frac{1}{16} + 4T\Delta^2\right)} \right),$$

which concludes the lemma.

E.2 Proof of Theorem 12

In this section, we give a complete outline of the proof of Theorem 12 which starts with the proof of Lemma 9.

E.2.1 Proof of Lemma 9

For each batch $m \in [M]$, let $\mathbf{b}_m = \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{i=1}^{N} \mathbf{x}_t^i y_t^i$ and $\mathbf{V}_m = \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{i=1}^{N} \mathbf{x}_t^i \mathbf{x}_t^{i\top}$. We have

$$\begin{aligned}
\mathbf{\Lambda}_m^i &= \lambda \mathbf{I} + \frac{NT_m}{2} \mathbb{E}_{\mathcal{X} \sim \mathcal{D}_m^i} \mathbb{E}_{\mathbf{x} \sim \pi_{m-1}^i(\mathcal{X})} [\mathbf{x}\mathbf{x}^\top] \\
&= \lambda \mathbf{I} + \frac{NT_m}{4} \left(2\mathbb{E}_{\mathcal{X} \sim \mathcal{D}_m^i} \mathbb{E}_{\mathbf{x} \sim \pi_{m-1}^i(\mathcal{X})} [\mathbf{x}\mathbf{x}^\top] + 6\gamma I \right) - 1.5 NT_m \gamma I. \qquad \text{(E.1)}
\end{aligned}$$

By choosing $\gamma = \frac{3\log(\frac{4dT}{\delta})}{NT_m}$ and $\lambda = 5\log\left(\frac{4dT}{\delta}\right)$, combining (E.1) and Lemma 44, for all $m \in [M]$, with probability at least $1 - \delta/2$, we have

$$\begin{aligned}
\mathbf{\Lambda}_m^i &\succeq \left(\lambda - 5\log\left(\frac{4dT}{\delta}\right) \right) I + \frac{1}{2} \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{i=1}^{N} \mathbf{x}_t^i \mathbf{x}_t^{i\top} \\
&= \frac{1}{2} \mathbf{V}_m. \qquad \text{(E.2)}
\end{aligned}$$

Moreover, for a fixed $\mathbf{x} \in \mathcal{X}_t^i$ and $(i,t) \in [N] \times [T]$, let $z_{t,m}^{j,i} =$

$\mathbf{x}^\top \left(\mathbf{\Lambda}_m^i\right)^{-1} \left(\mathbf{x}_t^j y_t^j - \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i}\mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta}\right)$. Thus, we have

$$\begin{aligned}
\left|\left\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta}\right\rangle\right| &= \left|\left\langle \mathbf{x}, \left(\mathbf{\Lambda}_m^i\right)^{-1}\mathbf{b}_m - \boldsymbol{\theta}\right\rangle\right| \\
&= \left|\left\langle \mathbf{x}, \left(\mathbf{\Lambda}_m^i\right)^{-1}\mathbf{b}_m\right\rangle - \left\langle \mathbf{x}, \left(\mathbf{\Lambda}_m^i\right)^{-1}\mathbf{\Lambda}_m^i\boldsymbol{\theta}\right\rangle\right| \\
&\le \left|\left\langle \mathbf{x}, \left(\mathbf{\Lambda}_m^i\right)^{-1}\mathbf{b}_m\right\rangle - \left\langle \mathbf{x}, \left(\mathbf{\Lambda}_m^i\right)^{-1}\left(\mathbf{\Lambda}_m^i - \lambda\mathbf{I}\right)\boldsymbol{\theta}\right\rangle\right| + \left|\lambda\langle \mathbf{x}, \left(\mathbf{\Lambda}_m^i\right)^{-1}\boldsymbol{\theta}\rangle\right| \\
&\le \left|\mathbf{x}^\top\left(\mathbf{\Lambda}_m^i\right)^{-1}\left(\mathbf{b}_m - \frac{NT_m}{2}\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i}\mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta}\right)\right| + \sqrt{\lambda}\|\mathbf{x}\|_{\left(\mathbf{\Lambda}_m^i\right)^{-1}} \\
&\qquad\qquad \text{(Cauchy Schwarz inequality and Assumption 16)} \\
&= \left|\sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2}\sum_{j=1}^{N} z_{t,m}^{j,i}\right| + \sqrt{\lambda}\|\mathbf{x}\|_{\left(\mathbf{\Lambda}_m^i\right)^{-1}}.
\end{aligned}$$

Note that

$$\mathbb{E}\left[z_{t,m}^{j,i}\right] = \mathbb{E}\left[\mathbf{x}^\top\left(\mathbf{\Lambda}_m^i\right)^{-1}\left(\mathbf{x}_t^j({\mathbf{x}_t^j}^\top\boldsymbol{\theta} + \eta_t^j) - \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i}\mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta}\right)\right] = 0,$$

(Noise η_t^j is zero-mean and independent of $\mathbf{x}_t^j$)

By Azuma's inequality, for a fixed $\mathbf{x} \in \mathcal{X}_t^i$ and $(i,t) \in [N]\times[T]$, we have

$$\mathbb{P}\left(\left|\sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2}\sum_{j=1}^{N} z_{t,m}^{j,i}\right| \ge \alpha\|\mathbf{x}\|_{\left(\mathbf{\Lambda}_m^i\right)^{-1}}\right) \le 2\exp\left(\frac{-\alpha^2\|\mathbf{x}\|^2_{\left(\mathbf{\Lambda}_m^i\right)^{-1}}}{2c_m^i}\right), \tag{E.3}$$

where

$$
\begin{aligned}
c_m^i &= \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} \left| \mathbf{x}^\top \left(\mathbf{\Lambda}_m^i\right)^{-1} \left(\mathbf{x}_t^j y_t^j - \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i} \mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta} \right) \right|^2 \\
&\leq 2 \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} \left| \mathbf{x}^\top \left(\mathbf{\Lambda}_m^i\right)^{-1} \mathbf{x}_t^j y_t^j \right|^2 + NT_m \left| \mathbf{x}^\top \left(\mathbf{\Lambda}_m^i\right)^{-1} \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i} \mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta} \right|^2 \\
&\leq 2 \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} \left| \mathbf{x}^\top \left(\mathbf{\Lambda}_m^i\right)^{-1} \mathbf{x}_t^j \right|^2 + \frac{4}{NT_m} \left| \mathbf{x}^\top \left(\mathbf{\Lambda}_m^i\right)^{-1} \left(\mathbf{\Lambda}_m^i - \lambda \mathbf{I}\right) \boldsymbol{\theta} \right|^2 \text{ (Assumption 16)} \\
&= 2 \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} \mathbf{x}^\top \left(\mathbf{\Lambda}_m^i\right)^{-1} \mathbf{x}_t^j \mathbf{x}_t^{j\top} \left(\mathbf{\Lambda}_m^i\right)^{-1} \mathbf{x} + \frac{4}{NT_m} \left| \mathbf{x}^\top \boldsymbol{\theta} - \lambda \mathbf{x}^\top \left(\mathbf{\Lambda}_m^i\right)^{-1} \boldsymbol{\theta} \right|^2 \\
&\leq 2\mathbf{x}^\top \left(\mathbf{\Lambda}_m^i\right)^{-1} \mathbf{V}_m \left(\mathbf{\Lambda}_m^i\right)^{-1} \mathbf{x} + \left(\frac{4\|\boldsymbol{\theta}\|_{\mathbf{\Lambda}_m^i}^2}{NT_m} + \frac{4\lambda}{NT_m} \right) \|\mathbf{x}\|_{(\mathbf{\Lambda}_m^i)^{-1}}^2 \\
&\qquad \text{(Cauchy Schwarz inequality and Assumption 16)} \\
&\leq 4\mathbf{x}^\top \left(\mathbf{\Lambda}_m^i\right)^{-1} \mathbf{\Lambda}_m^i \left(\mathbf{\Lambda}_m^i\right)^{-1} \mathbf{x} + \left(\frac{4\|\boldsymbol{\theta}\|_{\mathbf{\Lambda}_m^i}^2}{NT_m} + \frac{4\lambda}{NT_m} \right) \|\mathbf{x}\|_{(\mathbf{\Lambda}_m^i)^{-1}}^2 \\
&\qquad \text{(Conditioned on the event in Eqn. (E.2))} \\
&\leq \left(6 + \frac{8\lambda}{NT_m} \right) \|\mathbf{x}\|_{(\mathbf{\Lambda}_m^i)^{-1}}^2 , \qquad \text{(E.4)}
\end{aligned}
$$

where the last inequity follows from the fact that

$$
\|\boldsymbol{\theta}\|_{\mathbf{\Lambda}_m^i}^2 \leq \|\boldsymbol{\theta}\|_2^2 \, \lambda_{\max}\left(\mathbf{\Lambda}_m^i\right) \leq \lambda + \frac{NT_m}{2}. \qquad \text{(Assumption 16)}
$$

Combining (E.3) and (E.4), and by a union bound, we have

$$
\mathbb{P}\left(\left| \left\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta} \right\rangle \right| \leq \left(6\sqrt{\log\left(\frac{2KNT}{\delta}\right)} + \sqrt{\lambda} \right) \|\mathbf{x}\|_{(\mathbf{\Lambda}_m^i)^{-1}}, \ \forall \mathbf{x} \in \mathcal{X}_t^i, i \in [N], t \in [T], m \in [M] \right) \geq 1 - \delta. \tag{E.5}
$$

E.2.2 Completing the Proof of Theorem 12

Next, we state the following lemma, which we borrow from Theorem 5 in [103] and is used in the proof analysis of Theorem 12.

Lemma 39 ([103]). *Let $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_L \sim \mathcal{D}$ be i.i.d drawn from a distribution $\mathcal{D}$ and input of Algorithm 14 and let π be the output policy of Algorithm 14. For any $\lambda \in (0,1)$, we have*

$$\mathbb{P}\left[\mathbb{V}_{\mathcal{D}}^{\lambda}(\pi) \leq \mathcal{O}\left(\sqrt{d\log d\log(\lambda^{-1})}\right)\right] \geq 1 - exp\left(\mathcal{O}(d^3\log d\log(d\lambda^{-1})) - Ld^{-2c}2^{-16}\right),$$

where we define the λ-deviation of policy π over $\mathcal{D}$ by

$$\mathbb{V}_{\mathcal{D}}^{\lambda}(\pi) := \mathbb{E}_{\mathcal{X}\sim\mathcal{D}}\left[\max_{\mathbf{x}\in\mathcal{X}}\sqrt{\mathbf{x}^\top\left(\lambda\mathbf{I} + \mathbb{E}_{\mathcal{X}\sim\mathcal{D}}\mathbb{E}_{\mathbf{y}\sim\pi(\mathcal{X})}[\mathbf{y}\mathbf{y}^\top]\right)^{-1}\mathbf{x}}\right]. \tag{E.6}$$

Corollary 2. *As a direct corollary of Lemma 39, if $T \geq \Omega\left(d^{22}\log^2(\frac{NT}{\delta})\log^2 d\log^2(dNT\lambda^{-1})\right)$, then for all $m \geq 2$ and $i \in [N]$, with probability at least $1-\delta$, it holds that*

$$\mathbb{V}_{\mathcal{D}_m^i}^{(\frac{2\lambda}{NT_m})}(\pi_{m-1}^i) \leq \mathcal{O}(\sqrt{d\log d\log(NT\lambda^{-1})}). \tag{E.7}$$

Now, we focus on the regret of the i-th agent at m-th batch for any $m \geq 3$. Let $\mathcal{D}_m^i$ be the distribution based on which the surviving sets $\mathcal{X}_t^{i(m)}$ for all $t \in [\mathcal{T}_{m-1}+1 : \mathcal{T}_m]$ are generated when conditioned on the first $m-1$ batches. For any $t \in [\mathcal{T}_{m-1}+1 : \mathcal{T}_m]$, conditioned on the event that the confidence intervals in Lemma 9 hold, we have

$$\begin{aligned}
r_t^i &= \mathbb{E}\left[\langle\boldsymbol{\theta}, \mathbf{x}_{*,t}^i\rangle - \langle\boldsymbol{\theta}, \mathbf{x}_t^i\rangle\right] \\
&\leq \mathbb{E}\left[\langle\boldsymbol{\theta}_{m-1}^i, \mathbf{x}_{*,t}^i\rangle - \langle\boldsymbol{\theta}_{m-1}^i, \mathbf{x}_t^i\rangle + \beta\left\|\mathbf{x}_{*,t}^i\right\|_{\left(\boldsymbol{\Lambda}_{m-1}^i\right)^{-1}} + \beta\left\|\mathbf{x}_t^i\right\|_{\left(\boldsymbol{\Lambda}_{m-1}^i\right)^{-1}}\right] && \text{(Lemma 9)} \\
&\leq 2\beta\mathbb{E}\left[\left\|\mathbf{x}_{*,t}^i\right\|_{\left(\boldsymbol{\Lambda}_{m-1}^i\right)^{-1}} + \left\|\mathbf{x}_t^i\right\|_{\left(\boldsymbol{\Lambda}_{m-1}^i\right)^{-1}}\right] && (\mathbf{x}_{*,t}^i \in \mathcal{X}_t^{i(m)}) \\
&\leq 4\beta\mathbb{E}\left[\max_{\mathbf{x}\in\mathcal{X}_t^{i(m)}}\|\mathbf{x}\|_{\left(\boldsymbol{\Lambda}_{m-1}^i\right)^{-1}}\right] \\
&\leq 4\beta\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i}\left[\max_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x}\|_{\left(\boldsymbol{\Lambda}_{m-1}^i\right)^{-1}}\right] \\
&\leq 4\beta\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\left[\max_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x}\|_{\left(\boldsymbol{\Lambda}_{m-1}^i\right)^{-1}}\right] \\
&\leq \frac{8\beta}{\sqrt{NT_{m-1}}}\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\left[\max_{\mathbf{x}\in\mathcal{X}}\sqrt{\mathbf{x}^\top\left(\frac{2\lambda}{NT_{m-1}}I + \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\mathbb{E}_{\mathbf{y}\sim\pi_{m-2}^i(\mathcal{X})}[\mathbf{y}\mathbf{y}^\top]\right)^{-1}\mathbf{x}}\right] \\
&= \frac{8\beta}{\sqrt{NT_{m-1}}}\mathbb{V}_{\mathcal{D}_{m-1}^i}^{(\frac{2\lambda}{NT_{m-1}})}(\pi_{m-2}^i), && \text{(E.8)}
\end{aligned}$$

where the third inequality follows from our established confidence intervals in Lemma 9 guaranteeing that $\mathbf{x}^i_{*,t} \in \mathcal{X}_t^{i(m)}$ for all $(i,t,m) \in [N] \times [\mathcal{T}_{m-1}+1 : \mathcal{T}_m] \times [M]$ with probability at least $1-\delta$. Now, continuing form (E.8), we bound the cumulative regret of batches $m \geq 3$, as follows:

$$\begin{aligned}
\sum_{t=\mathcal{T}_2+1}^{T} \sum_{i=1}^{N} r_t^i &\leq \sum_{m=3}^{M} \frac{8\beta N T_m}{\sqrt{NT_{m-1}}} \mathbb{V}_{\mathcal{D}^i_{m-1}}^{(\frac{2\lambda}{NT_{m-1}})}(\pi^i_{m-2}) \\
&\leq 8\beta\sqrt{dN\log d \log(NT\lambda^{-1})} \sum_{m=2}^{M} \frac{T_m}{\sqrt{T_{m-1}}} \\
&\qquad\qquad \text{(Conditioned on the event in Eqn. (E.7))} \\
&= 8\beta M a \sqrt{dN \log d \log(NT\lambda^{-1})}. \qquad \text{(E.9)}
\end{aligned}$$

Next, we bound cumulative regret of the first two batches. Under Assumption 16, during the first two batches, the instantaneous regret of each agent i at any round t is at most 2. Therefore

$$\sum_{t=1}^{\mathcal{T}_2} \sum_{i=1}^{N} r_t^i \leq 2N\mathcal{T}_2 = 4a\sqrt{dN}. \qquad \text{(E.10)}$$

Note that for any $m \geq 3$, we can write T_m as

$$\begin{aligned}
T_m = a T_{m-1}^{\frac{1}{2}} = a^{\frac{3}{2}} T_{m-2}^{\frac{1}{4}} = \ldots = a^{\frac{2^{m-2}-1}{2^{m-3}}} T_2^{\frac{1}{2^{m-2}}} \\
T_m &= a^{\frac{1}{2^{m-2}}} a^{\frac{2^{m-2}-1}{2^{m-3}}} \left(\frac{T_2}{a}\right)^{\frac{1}{2^{m-2}}} \\
&= a^{\frac{2^{m-1}-1}{2^{m-2}}} \left(\sqrt{\frac{d}{N}}\right)^{\frac{1}{2^{m-2}}} \\
&= \left(\frac{a^{2^{m-1}-1} d^{\frac{1}{2}}}{N^{\frac{1}{2}}}\right)^{\frac{1}{2^{m-2}}}.
\end{aligned}$$

Our choice of a in the algorithm ensures that for any $M > 0$, $T_M = T$ and $\sum_{m=1}^{M} T_m \geq T_M = T$, and thus the choice of grid $\{\mathcal{T}_1, \ldots, \mathcal{T}_M\}$ is valid. If we let $M = 1 + \log\left(\frac{\log\left(\frac{NT}{d}\right)}{2} + 1\right)$,

from (E.9) and (E.10), we conclude that, with probability at least $1-2\delta$, it holds that

$$R_T \leq 4\sqrt{dNT}\left(\frac{NT}{d}\right)^{\frac{1}{2(2^{M-1}-1)}} + 8\beta M\sqrt{dNT\log d\log(NT\lambda^{-1})}\left(\frac{NT}{d}\right)^{\frac{1}{2(2^{M-1}-1)}}$$
$$\leq \mathcal{O}\left(\sqrt{dNT\log d\log^2\left(\frac{KNT}{\delta\lambda}\right)}\log\log\left(\frac{NT}{d}\right)\right). \tag{E.11}$$

E.2.3 Communication Cost as Number of Bits Transmitted

In this section, we consider the number of bits transmitted in a slightly modified version of DisBE-LUCB. To this end, we make the following minor modification to DisBE-LUCB. Let ϵ_0 be an additional input to the algorithm. In Line 9 of DisBE-LUCB, agent i sends vector $\tilde{\mathbf{u}}_m^i$ which is an ϵ_0-precise rounded version of $\mathbf{u}_m^i$. In particular, if it rounds each entry of $\mathbf{u}_m^i$ with precision ϵ_0, vector $\tilde{\mathbf{u}}_m^i$ will be obtained. Now, we observe how this extra rounding step affects confidence intervals in Lemma 9. In fact, we are interested in upper bounds on $\left|\left\langle \mathbf{x}, \tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}\right\rangle\right|$, where $\tilde{\boldsymbol{\theta}}_m^i = \left(\boldsymbol{\Lambda}_m^i\right)^{-1}\sum_{i=1}^N \tilde{\mathbf{u}}_m^i$.

For $\delta\in(0,1)$, let $\beta = 6\sqrt{\log\left(\frac{2KNT}{\delta}\right)} + \sqrt{\lambda}$. Then for all $\mathbf{x}\in\mathcal{X}_t^i, i\in[N], t\in[T], m\in[M]$, with probability at least $1-\delta$, it holds that

$$\begin{aligned}
\left|\left\langle \mathbf{x}, \tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}\right\rangle\right| &= \left|\left\langle \mathbf{x}, \tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}_m^i + \boldsymbol{\theta}_m^i - \boldsymbol{\theta}\right\rangle\right| \\
&\leq \left|\left\langle \mathbf{x}, \tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}_m^i\right\rangle\right| + \left|\left\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta}\right\rangle\right| \\
&\leq \left(\left\|\tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}_m^i\right\|_{\boldsymbol{\Lambda}_m^i} + \beta\right)\|\mathbf{x}\|_{\left(\boldsymbol{\Lambda}_m^i\right)^{-1}} \\
&\qquad\qquad \text{(Lemma 9 and Cauchy Schwarz inequality)} \\
&\leq \left(\sqrt{\boldsymbol{\Lambda}_{\max}(\boldsymbol{\Lambda}_m^i)}\left\|\tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}_m^i\right\|_2 + \beta\right)\|\mathbf{x}\|_{\left(\boldsymbol{\Lambda}_m^i\right)^{-1}} \\
&\leq \left(N\sqrt{dT}\epsilon_0 + \beta\right)\|\mathbf{x}\|_{\left(\boldsymbol{\Lambda}_m^i\right)^{-1}}. \qquad (E.12)
\end{aligned}$$

Therefore, letting $\epsilon_0 = \frac{\beta}{N\sqrt{dT}}$, we have

$$\left|\left\langle \mathbf{x}, \tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}\right\rangle\right| \leq 2\beta\|\mathbf{x}\|_{\left(\boldsymbol{\Lambda}_m^i\right)^{-1}}, \tag{E.13}$$

which implies that replacing β in DisBE-LUCB with 2β, will result in the same order of regret as that of DisBE-LUCB for our modified algorithm. Moreover, since for transmission of each

real number $\log(dNT)$ bits is used, the communication cost of our modified algorithm in terms of number of bits is same as that stated in Theorem 12 with an additional multiplicative factor $\log(dNT)$.

E.2.4 Relaxing the Assumption on Knowledge of $\mathcal{D}$

In this section, we relax this assumption and consider more realistic settings where each agent i can estimate matrix $\mathbf{\Lambda}_m^i$ in batch m up to an ϵ_m error, i.e.,

$$(1-\epsilon_m)\mathbf{\Lambda}_m^i \preceq \tilde{\mathbf{\Lambda}}_m^i \preceq (1+\epsilon_m)\mathbf{\Lambda}_m^i, \tag{E.14}$$

where $\tilde{\mathbf{\Lambda}}_m^i$ is an estimation of $\mathbf{\Lambda}_m^i$. Given this estimation, we define

$$\tilde{\boldsymbol{\theta}}_m^i = \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \sum_{j=1}^{N} \mathbf{u}_m^j, \tag{E.15}$$

as the new estimation of $\boldsymbol{\theta}$ computed by agent i at batch m in this modified version of DisBE-LUCB.

We note that if the inequalities hold component-wise, i.e., $(1-\epsilon_m)\mathbf{\Lambda}_m^i \leq \tilde{\mathbf{\Lambda}}_m^i \leq (1+\epsilon_m)\mathbf{\Lambda}_m^i$, this concludes that (E.14) holds. This is because for any positive semi-definite matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ such that $\mathbf{A} = \mathbf{B} + \mathbf{C}$, we have:

$$\mathbf{A} \succeq \mathbf{B},\ \mathbf{A} \succeq \mathbf{C}. \tag{E.16}$$

This combined with the fact that all $(1-\epsilon_m)\mathbf{\Lambda}_m^i$, $\tilde{\mathbf{\Lambda}}_m^i$, and $(1+\epsilon_m)\mathbf{\Lambda}_m^i$ are positive semi-definite symmetric matrices ensures that (E.14) holds if $(1-\epsilon_m)\mathbf{\Lambda}_m^i \leq \tilde{\mathbf{\Lambda}}_m^i \leq (1+\epsilon_m)\mathbf{\Lambda}_m^i$, and therefore, (E.14) is a weaker assumption than the component-wise assumption $(1-\epsilon_m)\mathbf{\Lambda}_m^i \leq \tilde{\mathbf{\Lambda}}_m^i \leq (1+\epsilon_m)\mathbf{\Lambda}_m^i$.

Now, we define corresponding modified confidence intervals in the following lemma.

Lemma 40. *Suppose* $\|\boldsymbol{\theta}\|_2 \leq 1$, $\left\|\mathbf{x}_{t,a}^i\right\|_2 \leq 1$, $\left|y_t^i\right| \leq 1$ *for all* $(a,i,t) \in [K]\times[N]\times[T]$ *and* $\epsilon_m \leq \sqrt{\frac{\lambda}{NT_m}}$ *for all* $m \in [M]$. *For* $\delta \in (0,1)$, *let* $\beta_m = 6\sqrt{\frac{\log\left(\frac{2KNT}{\delta}\right)}{1-\epsilon_m}} + 4\sqrt{\lambda}$. *Then for all* $\mathbf{x} \in \mathcal{X}_t^i, i \in [N], t \in [T], m \in [M]$, *with probability at least* $1-\delta$, *it holds that* $\left|\left\langle \mathbf{x}, \tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}\right\rangle\right| \leq \beta_m \|\mathbf{x}\|_{\left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1}}$.

Proof. The proof closely follows the steps in the proof of Lemma 9. For each batch $m \in [M]$, let $\mathbf{b}_m = \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{i=1}^{N} \mathbf{x}_t^i y_t^i$ and $\mathbf{V}_m = \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{i=1}^{N} \mathbf{x}_t^i \mathbf{x}_t^{i\top}$. For a fixed $\mathbf{x} \in \mathcal{X}_t^i$ and $(i,t) \in [N] \times [T]$, let $z_{t,m}^{j,i} = \mathbf{x}^\top \left(\tilde{\mathbf{\Lambda}}_m^{\,i}\right)^{-1} \left(\mathbf{x}_t^j y_t^j - \mathbb{E}_{\mathcal{X} \sim \mathcal{D}_m^i} \mathbb{E}_{\mathbf{x} \sim \pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta}\right)$. Thus, we have

$$
\begin{aligned}
\left|\left\langle \mathbf{x}, \tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta} \right\rangle\right| &= \left|\left\langle \mathbf{x}, \left(\tilde{\mathbf{\Lambda}}_m^{\,i}\right)^{-1} \mathbf{b}_m - \boldsymbol{\theta} \right\rangle\right| \\
&= \left|\left\langle \mathbf{x}, \left(\tilde{\mathbf{\Lambda}}_m^{\,i}\right)^{-1} \mathbf{b}_m \right\rangle - \left\langle \mathbf{x}, \left(\tilde{\mathbf{\Lambda}}_m^{\,i}\right)^{-1} \tilde{\mathbf{\Lambda}}_m^{\,i} \boldsymbol{\theta} \right\rangle\right| \\
&= \left|\left\langle \mathbf{x}, \left(\tilde{\mathbf{\Lambda}}_m^{\,i}\right)^{-1} \mathbf{b}_m \right\rangle - \left\langle \mathbf{x}, \left(\tilde{\mathbf{\Lambda}}_m^{\,i}\right)^{-1} \left(\mathbf{\Lambda}_m^i - \lambda \mathbf{I}\right) \boldsymbol{\theta} \right\rangle + \left\langle \mathbf{x}, \left(\tilde{\mathbf{\Lambda}}_m^{\,i}\right)^{-1} \left(\mathbf{\Lambda}_m^i - \tilde{\mathbf{\Lambda}}_m^{\,i} - \lambda \mathbf{I}\right) \boldsymbol{\theta} \right\rangle\right| \\
&\le \left|\left\langle \mathbf{x}, \left(\tilde{\mathbf{\Lambda}}_m^{\,i}\right)^{-1} \mathbf{b}_m \right\rangle - \left\langle \mathbf{x}, \left(\tilde{\mathbf{\Lambda}}_m^{\,i}\right)^{-1} \left(\mathbf{\Lambda}_m^i - \lambda \mathbf{I}\right) \boldsymbol{\theta} \right\rangle\right| + \left|\left\langle \mathbf{x}, \left(\tilde{\mathbf{\Lambda}}_m^{\,i}\right)^{-1} \left(\mathbf{\Lambda}_m^i - \tilde{\mathbf{\Lambda}}_m^{\,i} - \lambda \mathbf{I}\right) \boldsymbol{\theta} \right\rangle\right| \\
&\le \left|\mathbf{x}^\top \left(\tilde{\mathbf{\Lambda}}_m^{\,i}\right)^{-1} \left(\mathbf{b}_m - \frac{NT_m}{2} \mathbb{E}_{\mathcal{X} \sim \mathcal{D}_m^i} \mathbb{E}_{\mathbf{x} \sim \pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta}\right)\right| + 4\sqrt{\lambda}\|\mathbf{x}\|_{\left(\tilde{\mathbf{\Lambda}}_m^{\,i}\right)^{-1}} \\
&\qquad \text{(Cauchy Schwarz inequality)} \\
&= \left|\sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} z_{t,m}^{j,i}\right| + 4\sqrt{\lambda}\|\mathbf{x}\|_{\left(\tilde{\mathbf{\Lambda}}_m^{\,i}\right)^{-1}}, \qquad \text{(E.17)}
\end{aligned}
$$

where the second inequality follows from

$$\begin{aligned}
\|\boldsymbol{\theta}\|_{\left(\tilde{\boldsymbol{\Lambda}}_m^i\right)^{-1}\left(\boldsymbol{\Lambda}_m^i-\tilde{\boldsymbol{\Lambda}}_m^i-\lambda\mathbf{I}\right)^2} &= \sqrt{\boldsymbol{\theta}^\top\left(\tilde{\boldsymbol{\Lambda}}_m^i\right)^{-1}\left(\boldsymbol{\Lambda}_m^i-\tilde{\boldsymbol{\Lambda}}_m^i-\lambda\mathbf{I}\right)^2\boldsymbol{\theta}} \\
&\le \|\boldsymbol{\theta}\|_2\sqrt{\lambda_{\max}\left(\left(\tilde{\boldsymbol{\Lambda}}_m^i\right)^{-1}\left(\boldsymbol{\Lambda}_m^i-\tilde{\boldsymbol{\Lambda}}_m^i-\lambda\mathbf{I}\right)^2\right)} \\
&\le \sqrt{\lambda_{\max}\left(\left(\tilde{\boldsymbol{\Lambda}}_m^i\right)^{-1}\left(\boldsymbol{\Lambda}_m^i-\tilde{\boldsymbol{\Lambda}}_m^i\right)^2+\lambda^2\left(\tilde{\boldsymbol{\Lambda}}_m^i\right)^{-1}\right)} \qquad (\|\boldsymbol{\theta}\|_2\le 1) \\
&\le \sqrt{\lambda_{\max}\left(\left(\tilde{\boldsymbol{\Lambda}}_m^i\right)^{-1}\left(\boldsymbol{\Lambda}_m^i-\tilde{\boldsymbol{\Lambda}}_m^i\right)^2+\lambda^2\left(\tilde{\boldsymbol{\Lambda}}_m^i\right)^{-1}\right)} \\
&\le \sqrt{\lambda_{\max}\left(\left(\tilde{\boldsymbol{\Lambda}}_m^i\right)^{-1}\left(\boldsymbol{\Lambda}_m^i-\tilde{\boldsymbol{\Lambda}}_m^i\right)^2\right)}+\sqrt{\lambda} \\
&\qquad\qquad \text{(Cauchy Schwarz inequality)} \\
&\le \epsilon_m\sqrt{\lambda_{\max}\left(\tilde{\boldsymbol{\Lambda}}_m^i\right)}+\sqrt{\lambda} \qquad \text{(Eqn. (E.14))} \\
&\le 2\epsilon_m\sqrt{\lambda_{\max}\left(\boldsymbol{\Lambda}_m^i\right)}+\sqrt{\lambda} \qquad \text{(Eqn. (E.14))} \\
&\le \epsilon_m\sqrt{NT_m}+3\sqrt{\lambda} \\
&\le 4\sqrt{\lambda}. \qquad (\epsilon_m\le\sqrt{\tfrac{\lambda}{NT_m}})
\end{aligned}$$

Note that

$$\mathbb{E}\left[z_{t,m}^{j,i}\right]=\mathbb{E}\left[\mathbf{x}^\top\left(\tilde{\boldsymbol{\Lambda}}_m^i\right)^{-1}\left(\mathbf{x}_t^j(\mathbf{x}_t^{j\top}\boldsymbol{\theta}+\eta_t^j)-\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i}\mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta}\right)\right]=0,$$

(Noise η_t^j is zero-mean and independent of $\mathbf{x}_t^j$)

By Azuma's inequality, for a fixed $\mathbf{x}\in\mathcal{X}_t^i$ and $(i,t)\in[N]\times[T]$, we have

$$\mathbb{P}\left(\left|\sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2}\sum_{j=1}^{N}z_{t,m}^{j,i}\right|\ge\alpha\|\mathbf{x}\|_{\left(\tilde{\boldsymbol{\Lambda}}_m^i\right)^{-1}}\right)\le 2\exp\left(\frac{-\alpha^2\|\mathbf{x}\|^2_{\left(\tilde{\boldsymbol{\Lambda}}_m^i\right)^{-1}}}{2c_m^i}\right), \tag{E.18}$$

where

$$
\begin{aligned}
c_m^i &= \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} \left| \mathbf{x}^\top \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \left(\mathbf{x}_t^j y_t^j - \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i} \mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta} \right) \right|^2 \\
&\le 2 \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} \left| \mathbf{x}^\top \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \mathbf{x}_t^j y_t^j \right|^2 + NT_m \left| \mathbf{x}^\top \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i} \mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta} \right|^2 \\
&\le 2 \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} \left| \mathbf{x}^\top \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \mathbf{x}_t^j \right|^2 + \frac{4}{NT_m} \left| \mathbf{x}^\top \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \left(\mathbf{\Lambda}_m^i - \lambda \mathbf{I} \right) \boldsymbol{\theta} \right|^2 \\
&= 2 \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} \mathbf{x}^\top \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \mathbf{x}_t^j \mathbf{x}_t^{j\top} \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \mathbf{x} + \frac{4}{NT_m} \left| \mathbf{x}^\top \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \left(\mathbf{\Lambda}_m^i \right) \boldsymbol{\theta} - \lambda \mathbf{x}^\top \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \boldsymbol{\theta} \right|^2 \\
&\le 2\mathbf{x}^\top \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \mathbf{V}_m \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \mathbf{x} + \frac{1}{1-\epsilon_m} \left(4 + \frac{8\lambda}{NT_m}\right) \|\mathbf{x}\|^2_{\left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1}} \qquad \text{(Cauchy Schwarz inequality)} \\
&\le 4\mathbf{x}^\top \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \mathbf{\Lambda}_m^i \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \mathbf{x} + \frac{1}{1-\epsilon_m} \left(4 + \frac{8\lambda}{NT_m}\right) \|\mathbf{x}\|^2_{\left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1}} \\
&\qquad \text{(Conditioned on the event in Eqn. (E.2))} \\
&\le \frac{4}{1-\epsilon_m} \mathbf{x}^\top \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \mathbf{x} + \frac{1}{1-\epsilon_m} \left(4 + \frac{8\lambda}{NT_m}\right) \|\mathbf{x}\|^2_{\left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1}} \qquad ((1-\epsilon_m)\mathbf{\Lambda}_m^i \preceq \tilde{\mathbf{\Lambda}}_m^i) \\
&= \frac{8}{1-\epsilon_m} \left(1 + \frac{\lambda}{NT_m}\right) \|\mathbf{x}\|^2_{\left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1}} \\
&\le \frac{16}{1-\epsilon_m} \|\mathbf{x}\|^2_{\left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1}}, \qquad \text{(E.19)}
\end{aligned}
$$

where the third inequity follows from the fact that

$$
\begin{aligned}
\boldsymbol{\theta}^\top \left(\mathbf{\Lambda}_m^i \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \mathbf{\Lambda}_m^i \right) \boldsymbol{\theta} &\le \|\boldsymbol{\theta}\|^2 \lambda_{\max} \left(\mathbf{\Lambda}_m^i \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \mathbf{\Lambda}_m^i \right) \\
&\le \lambda_{\max} \left(\mathbf{\Lambda}_m^i \left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1} \mathbf{\Lambda}_m^i \right) \qquad (\|\boldsymbol{\theta}\|_2 \le 1) \\
&\le \frac{1}{1-\epsilon_m} \lambda_{\max} \left(\mathbf{\Lambda}_m^i \right) \qquad ((1-\epsilon_m)\mathbf{\Lambda}_m^i \preceq \tilde{\mathbf{\Lambda}}_m^i) \\
&\le \frac{\lambda + NT_m}{1-\epsilon_m}.
\end{aligned}
$$

Combining (E.17), (E.18) and

(E.19), and by a union bound, we have

$$
\mathbb{P}\left(\left| \left\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta} \right\rangle \right| \le \left(6\sqrt{\frac{\log\left(\frac{2KNT}{\delta}\right)}{1-\epsilon_m}} + 4\sqrt{\lambda} \right) \|\mathbf{x}\|_{\left(\tilde{\mathbf{\Lambda}}_m^i\right)^{-1}}, \ \forall \mathbf{x} \in \mathcal{X}_t^i, i \in [N], t \in [T], m \in [M] \right) \ge 1 - \delta. \qquad \text{(E.20)}
$$

□

Now, we state the regret bound for DisBE-LUCB with $\tilde{\mathbf{\Lambda}}_m^i$ and $\tilde{\boldsymbol{\theta}}_m^i$.

Theorem 20. *Fix* $M = 1 + \log\Big(\log\big(NT/d\big)/2 + 1\Big)$. *Under the setting of Lemma 40, if* $T \geq \Omega\left(d^{22}\log^2(NT/\delta)\log^2 d\log^2(d\lambda^{-1})\right)$ *and* $\beta = \max_{m\in[M]}\beta_m$, *then with probability at least* $1-2\delta$, *it holds that* $R_T \leq \mathcal{O}\left(\frac{1}{1-\max_{m\in[M]}\epsilon_m}\sqrt{dNT\log d\log^2\left(\frac{KNT}{\delta\lambda}\right)}\log\log\left(\frac{NT}{d}\right)\right)$, *where the communication cost is measured by the number of real numbers communicated by the agents.*

Proof. The proof follows similar steps to those in the proof of Theorem 12.

We focus on the regret of the i-th agent at m-th batch for any $m \geq 3$. Let $\mathcal{D}_m^i$ be the distribution based on which the surviving sets $\mathcal{X}_t^{i(m)}$ for all $t \in [\mathcal{T}_{m-1}+1 : \mathcal{T}_m]$ are generated when conditioned on the first $m-1$ batches. For any $t \in [\mathcal{T}_{m-1}+1 : \mathcal{T}_m]$, conditioned on the event that the confidence intervals in Lemma 9 hold, we have

$$
\begin{aligned}
r_t^i &= \mathbb{E}\left[\langle\boldsymbol{\theta}, \mathbf{x}_{*,t}^i\rangle - \langle\boldsymbol{\theta}, \mathbf{x}_t^i\rangle\right] \\
&\leq \mathbb{E}\left[\langle\tilde{\boldsymbol{\theta}}_{m-1}^i, \mathbf{x}_{*,t}^i\rangle - \langle\tilde{\boldsymbol{\theta}}_{m-1}^i, \mathbf{x}_t^i\rangle + \beta\left\|\mathbf{x}_{*,t}^i\right\|_{\left(\tilde{\mathbf{\Lambda}}_{m-1}^i\right)^{-1}} + \beta\left\|\mathbf{x}_t^i\right\|_{\left(\tilde{\mathbf{\Lambda}}_{m-1}^i\right)^{-1}}\right] && \text{(Lemma 40)} \\
&\leq 2\beta\mathbb{E}\left[\left\|\mathbf{x}_{*,t}^i\right\|_{\left(\tilde{\mathbf{\Lambda}}_{m-1}^i\right)^{-1}} + \left\|\mathbf{x}_t^i\right\|_{\left(\tilde{\mathbf{\Lambda}}_{m-1}^i\right)^{-1}}\right] && (\mathbf{x}_{*,t}^i \in \mathcal{X}_t^{i(m)}) \\
&\leq 4\beta\mathbb{E}\left[\max_{\mathbf{x}\in\mathcal{X}_t^{i(m)}}\|\mathbf{x}\|_{\left(\tilde{\mathbf{\Lambda}}_{m-1}^i\right)^{-1}}\right] \\
&\leq 4\beta\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i}\left[\max_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x}\|_{\left(\tilde{\mathbf{\Lambda}}_{m-1}^i\right)^{-1}}\right] \\
&\leq 4\beta\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\left[\max_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x}\|_{\left(\tilde{\mathbf{\Lambda}}_{m-1}^i\right)^{-1}}\right] \\
&\leq \frac{4\beta}{\sqrt{1-\epsilon_m}}\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\left[\max_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x}\|_{\left(\mathbf{\Lambda}_{m-1}^i\right)^{-1}}\right] && ((1-\epsilon_m)\mathbf{\Lambda}_m^i \preceq \tilde{\mathbf{\Lambda}}_m^i) \\
&\leq \frac{8\beta}{\sqrt{NT_{m-1}(1-\epsilon_m)}}\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\left[\max_{\mathbf{x}\in\mathcal{X}}\sqrt{\mathbf{x}^\top\left(\frac{2\lambda}{NT_{m-1}}I + \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\mathbb{E}_{\mathbf{y}\sim\pi_{m-2}^i(\mathcal{X})}[\mathbf{y}\mathbf{y}^\top]\right)^{-1}\mathbf{x}}\right] \\
&= \frac{8\beta}{\sqrt{NT_{m-1}(1-\epsilon_m)}}\mathbb{V}_{\mathcal{D}_{m-1}^i}^{(\frac{2\lambda}{NT_{m-1}})}(\pi_{m-2}^i), && \text{(E.21)}
\end{aligned}
$$

where the third inequality follows from our established confidence intervals in Lemma 40 guaranteeing that $\mathbf{x}_{*,t}^i \in \mathcal{X}_t^{i(m)}$ for all $(i,t,m) \in [N]\times[\mathcal{T}_{m-1}+1:\mathcal{T}_m]\times[M]$ with probability

at least $1-\delta$. The rest of the proof follows the steps as those in the proof of Theorem 12 with an additional $\frac{1}{\sqrt{1-\epsilon_m}}$ multiplicative factor in the bound.

Therefore, we conclude that, with probability at least $1-2\delta$, it holds that

$$\begin{aligned} R_T &\leq 4\sqrt{dNT}\left(\frac{NT}{d}\right)^{\frac{1}{2(2^{M-1}-1)}} + 8\beta M\sqrt{\frac{dNT\log d\log(NT\lambda^{-1})}{1-\max_{m\in[M]}\epsilon_m}}\left(\frac{NT}{d}\right)^{\frac{1}{2(2^{M-1}-1)}} \\ &\leq \mathcal{O}\left(\frac{1}{1-\max_{m\in[M]}\epsilon_m}\sqrt{dNT\log d\log^2\left(\frac{KNT}{\delta\lambda}\right)}\log\log\left(\frac{NT}{d}\right)\right). \end{aligned} \tag{E.22}$$

□

E.3 Decentralized Batch Elimination LUCB without Server

In this environment, the agents are represented by the nodes of an undirected and connected graph G. Each agent i can send and receive messages only to and from its immediate neighbors $j \in \mathcal{N}(i)$.

Definition 4 (Communication Matrix)**.** *For an undirected connected graph G with N nodes, $\mathbf{P} \in \mathbb{R}^{N\times N}$ is a symmetric communication matrix if it satisfies the following three conditions: (i) $\mathbf{P}_{i,j} = 0$ if there is no connection between nodes i and j; (ii) the sum of each row and column of $\mathbf{P}$ is 1; (iii) the eigenvalues are real and their magnitude is less than 1, i.e., $1 = |\lambda_1| > |\lambda_2| \geq \ldots |\lambda_N| \geq 0$.*

We assume that $\mathbf{P}$ is known to the agents. We remark that $\mathbf{P}$ can be constructed with little global information about the graph, such as its adjacency matrix and the graph's maximal degree; For example, one can compute it as $\mathbf{P} = I_N - \frac{1}{\delta_{\max}+1}\mathbf{D}^{-1/2}\mathcal{L}\mathbf{D}^{-1/2}$, where $\delta_{\max}$ is the maximum degree of the graph, $\mathcal{L} \in \mathbb{R}^{N\times N}$ is the graph Laplacian, and $\mathbf{D} \in \mathbb{R}^{N\times N}$ is a diagonal matrix whose entries are the degrees of the nodes (see [47] for details).

Running consensus. In order to share information about agents' past actions among the network, we rely on *running consensus*, e.g., [89, 139]. The goal of running consensus is that after enough rounds of communication, each agent has an accurate estimate of the

average (over all agents) of the initial values of each agent. Precisely, let $\boldsymbol{\nu}_0 \in \mathbb{R}^N$ be a vector, where each entry $\boldsymbol{\nu}_{0,i}, i \in [N]$ represents agent's i information at some initial round. Then, running consensus aims at providing an accurate estimate of the average $\frac{1}{N}\sum_{i\in[N]} \boldsymbol{\nu}_{0,i}$ for each agent. It turns out that the communication matrix $\mathbf{P}$ defined in Definition 4 plays a key role in reaching consensus. The details are standard in the rich related literature [139, 89]. Here, we only give a brief explanation of the high-level principles. Roughly speaking, a consensus algorithm updates $\boldsymbol{\nu}_0$ by $\boldsymbol{\nu}_1 = \mathbf{P}\boldsymbol{\nu}_0$, $\boldsymbol{\nu}_2 = \mathbf{P}\boldsymbol{\nu}_1$ and so on. Note that this operation respects the network structure since the updated value $\boldsymbol{\nu}_{1,j}$ is a weighted average of only $\boldsymbol{\nu}_{0,j}$ itself and neighbor-only values $\boldsymbol{\nu}_{0,i}, i \in \mathcal{N}(j)$. Thus, after S rounds, agent j has access to entry j of $\boldsymbol{\nu}_S = \mathbf{P}^S\boldsymbol{\nu}_0$. We adapt *polynomial filtering* introduced in [90, 107] to speed up the mixing of information by following an approach whose convergence rate is faster than the standard multiplication method above. Specifically, after S communication rounds, instead of $\mathbf{P}^S$, agents compute and apply to the initial vector $\boldsymbol{\nu}_0$ an appropriate re-scaled *Chebyshev polynomial* $q_S(\mathbf{P})$ of degree S of the communication matrix. Recall that Chebyshev polynomials are defined recursively. It turns out that the Chebyshev polynomial of degree ℓ for a communication matrix $\mathbf{P}$ is also given by a recursive formula as follows: $q_{\ell+1}(\mathbf{P}) = \frac{2w_\ell}{|\lambda_2|w_{\ell+1}}\mathbf{P}q_\ell(\mathbf{P}) - \frac{w_{\ell-1}}{w_{\ell+1}}q_{\ell-1}(\mathbf{P})$, where $w_0 = 0, w_1 = 1/|\lambda_2|$, $w_{\ell+1} = 2w_\ell/|\lambda_2| - w_{\ell-1}$, $q_0(\mathbf{P}) = I$ and $q_1(\mathbf{P}) = \mathbf{P}$. Specifically, in a Chebyshev-accelerated gossip protocol [90], the agents update their estimates of the average of the initial vector's $\boldsymbol{\nu}_0$ entries as follows:

$$\boldsymbol{\nu}_{\ell+1} = (2w_\ell)/(|\lambda_2|w_{\ell+1})\mathbf{P}\boldsymbol{\nu}_\ell - (w_{\ell-1}/w_{\ell+1})\boldsymbol{\nu}_{\ell-1}. \tag{E.23}$$

DecBE-LUCB, presented in Algorithm 12, implements the Chebyshev-accelerated gossip protocol outlined above for every entry of vectors $\mathbf{u}_m^i = \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \mathbf{x}_t^i y_t^i$ at the end of m-th batch.

The accelerated consensus algorithm, summarized in Algorithm 13, guarantees fast mixing of information thanks to the following key property stated in Lemma 3 of [90]: for $\epsilon \in (0,1)$ and any vector $\boldsymbol{\nu_0}$ in the N-dimensional simplex, it holds that

$$\|Nq_S(\mathbf{P})\boldsymbol{\nu_0} - \mathbf{1}\|_2 \leq \epsilon, \text{ if } S = \frac{\log(2N/\epsilon)}{\sqrt{2\log(1/|\lambda_2|)}}. \tag{E.24}$$

In view of this, DecBE-LUCB properly implements the accelerated consensus algorithm such that for every $i \in [N]$ and $m \in [M]$, the vector $\mathbf{u}_m^i$ is communicated within the network during the last S rounds of batch m. At round $\mathcal{T}_m+1$, agent i has access to $\sum_{j=1}^N a_{i,j}\mathbf{u}_m^j$, where $a_{i,j} = N[q_S(\mathbf{P})]_{i,j}$. Thanks to (E.24), $a_{i,j}$ is ϵ close to 1, thus, these are good approximations of the true $\sum_{j=1}^N \mathbf{u}_m^j$. Furthermore, the choice of grid $\mathcal{T} = \{\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_M\}$ in DecBE-LUCB is slightly different than what used in DisBE-LUCB.

E.3.1 Theoretical Guarantees of DecBE-LUCB

As the first step in regret analysis of DecBE-LUCB, we establish the following confidence intervals.

Lemma 41 (Confidence intervals for DecBE-LUCB)**.** *Suppose Assumption 16 holds. Fix $\delta \in (0,1)$ and let $\epsilon = \frac{\beta}{\sqrt{d}}$ and $\gamma = 2\beta$, where β is defined in Lemma 9. Then*

$$\mathbb{P}\left(\left|\left\langle \mathbf{x}, \hat{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}\right\rangle\right| \le \gamma\|\mathbf{x}\|_{\left(\boldsymbol{\Lambda}_m^i\right)^{-1}},\ \forall \mathbf{x} \in \mathcal{X}_t^i, i \in [N], t \in [T], m \in [M]\right) \ge 1-\delta. \tag{E.25}$$

Proof. Recall the definition of $\boldsymbol{\theta}_m^i$ in (6.4). For a fixed $\mathbf{x} \in \mathcal{X}_t^i$ and $(i,t) \in [N] \times [T]$, we have

$$\begin{aligned}
\left|\left\langle \mathbf{x}, \hat{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}\right\rangle\right| &\le \left|\left\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta}\right\rangle\right| + \left|\left\langle \mathbf{x}, \hat{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}_m^i\right\rangle\right| \\
&\le \left|\left\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta}\right\rangle\right| + \|\mathbf{x}\|_{\left(\boldsymbol{\Lambda}_m^i\right)^{-2}} \left\|\bar{\mathbf{u}}_{m,i} - \sum_{j=1}^N \mathbf{u}_m^j\right\|_2 \quad \text{(Cauchy Schwarz inequality)} \\
&\le \left|\left\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta}\right\rangle\right| + \epsilon\sqrt{d}\|\mathbf{x}\|_{\left(\boldsymbol{\Lambda}_m^i\right)^{-1}} \\
&\qquad \text{(Assumption 16 and choice of } S \text{ in (E.24))} \\
&= \left|\left\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta}\right\rangle\right| + \beta\|\mathbf{x}\|_{\left(\boldsymbol{\Lambda}_m^i\right)^{-1}}. \qquad \text{(E.26)}
\end{aligned}$$

Combining Lemma 9 and (E.26), we have

$$\mathbb{P}\left(\left|\left\langle \mathbf{x}, \hat{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}\right\rangle\right| \le 2\beta\|\mathbf{x}\|_{\left(\boldsymbol{\Lambda}_m^i\right)^{-1}},\ \forall \mathbf{x} \in \mathcal{X}_t^i, i \in [N], t \in [T], m \in [M]\right) \ge 1-\delta. \tag{E.27}$$

□

Algorithm 12 DecBE-LUCB for agent i

Input: N, d, δ, T, M, λ, ϵ

Initialization: $S = \frac{\log(2N/\epsilon)}{\sqrt{2\log(1/|\lambda_2|)}}$, $a = \sqrt{T+S}\left(\frac{N(T+S)}{d}\right)^{\frac{1}{2(2^{M-1}-1)}}$, $T_1 = T_2 = a\sqrt{\frac{d}{N}} + S$, $T_m = \lfloor a\sqrt{T_{m-1}-S} + S \rfloor$, $\boldsymbol{\theta}_0^i = \mathbf{0}$, $\boldsymbol{\Lambda}_0^i = \lambda\mathbf{I}$, $\mathcal{T}_0 = 0$, $\mathcal{T}_m = \mathcal{T}_{m-1} + T_m$, $\lambda = 5\log\left(\frac{4dT}{\delta}\right)$, $\gamma = 12\sqrt{\log\left(\frac{2KNT}{\delta}\right)} + 2\sqrt{\lambda}$, arbitrary policy π_0^i

for $m = 1, \ldots, M$ **do**

  for $t = \mathcal{T}_{m-1}+1, \ldots, \min\{\mathcal{T}_m, T\}$ **do**

    Let $\mathcal{X}_t^{i(m)} = \cap_{k=0}^{m-1}\mathcal{E}\left(\mathcal{X}_t^i; (\boldsymbol{\Lambda}_k^i, \hat{\boldsymbol{\theta}}_k^i, \gamma)\right)$

    Play arm $a_{i,t}$ associated with feature vector $\mathbf{x}_t^i \sim \pi_{m-1}\left(\mathcal{X}_t^{i(m)}\right)$ and observe y_t^i.

  end for Set $\mathcal{K}_0^i = \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+(T_m-S)/2} \mathbf{x}_t^i y_t^i$

  for $t = \mathcal{T}_m - S + 1$ **do**

    Let $\mathcal{X}_t^{i(m)} = \cap_{k=0}^{m-1}\mathcal{E}\left(\mathcal{X}_t^i; (\boldsymbol{\Lambda}_k^i, \hat{\boldsymbol{\theta}}_k^i, \gamma)\right)$

    Play arm $a_{i,t}$ associated with feature vector $\mathbf{x}_t^i \sim \pi_{m-1}\left(\mathcal{X}_t^{i(m)}\right)$ and observe y_t^i.

    Send each entry of $\mathcal{K}_0^i$, i.e., $[\mathcal{K}_0^i]_n$, $\forall n \in [d]$ to your neighbors $\mathcal{N}(j)$ and receive the corresponding values from them. For each $n \in [d]$, update $[\mathcal{K}_1^i]_n = \mathbf{P}_{i,i}[\mathcal{K}_0^i]_n + \sum_{j\in\mathcal{N}(i)} \mathbf{P}_{i,j}[\mathcal{K}_0^j]_n$

  end for

  Set $s = 1$

  for $t = \mathcal{T}_m - S + 2, \ldots, \mathcal{T}_m$ **do**

    Construct set $\mathcal{X}_t^{i(m)} = \cap_{k=0}^{m-1}\mathcal{E}\left(\mathcal{X}_t^i; (\boldsymbol{\Lambda}_k^i, \hat{\boldsymbol{\theta}}_k^i, \gamma)\right)$.

    Play arm $a_{i,t}$ associated with feature vector $\mathbf{x}_t^i \sim \pi_{m-1}\left(\mathcal{X}_t^{i(m)}\right)$ and observe y_t^i. $[\mathcal{K}_{s+1}^i]_n = \text{Comm}([\mathcal{K}_s^i]_n, [\mathcal{K}_{s-1}^i]_n, s+1)$, $\forall n \in [d]$

    $s = s + 1$

  end for

  Compute/construct

$$\begin{aligned}
\boldsymbol{\Lambda}_m^i &= \lambda\mathbf{I} + \frac{N(T_m - S)}{2}\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i}\mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top], \\
\hat{\boldsymbol{\theta}}_m^i &= \left(\boldsymbol{\Lambda}_m^i\right)^{-1}\bar{\mathbf{u}}_{m,i}, \\
\mathcal{S}_m^i &= \left\{\mathcal{X}_t^{i(m+1)}\right\}_{t=\mathcal{T}_{m-1}+(T_m-S)/2+1}^{\mathcal{T}_m}, \\
\pi_m^i &= \text{ ExplorationPolicy}\left(\frac{2\lambda}{N(T_m - S)}, \mathcal{S}_m^i\right).
\end{aligned}$$

end for

Theorem 21. *Fix* $M = 1 + \log\left(\frac{\log\left(\frac{N(T+S)}{d}\right)}{2} + 1\right)$, *with* S *defined in* (E.24) *for* $\epsilon = 6\sqrt{\frac{\log\left(\frac{2dKNT}{\delta}\right)}{d}}$ *in Algorithm 8. Suppose Assumption 16 holds. If* $T \geq \Omega\left(d^{22}\log^2(\frac{NT}{\delta})\log^2 d\log^2(d\lambda^{-1})\right)$, *then with probability at least* $1-2\delta$, *it holds that*

$$R_T \leq \mathcal{O}\left(\left(\frac{N\log(dN)}{\sqrt{1/|\lambda_2|}} + \sqrt{dN\left(T + \frac{\log(dN)}{\sqrt{1/|\lambda_2|}}\right)\log d\log^2\left(\frac{KN\left(T+\frac{\log(dN)}{\sqrt{1/|\lambda_2|}}\right)}{\delta\lambda}\right)}\right)\log\log\left(\frac{NT}{d}\right)\right), \tag{E.28}$$

and

$$\textit{Communication Cost} \leq \mathcal{O}\left(\frac{\delta_{max}dN\log(dN)}{\sqrt{\log(1/|\mathbf{\Lambda}_2|)}}\right). \tag{E.29}$$

Proof. The proof follows similar steps as those of Theorem 12's proof. We focus on the regret of m-th batch for any $m \geq 3$. For any $i \in [N]$, $t \in [\mathcal{T}_{m-1}+1 : \mathcal{T}_m]$, conditioned on the event that the confidence intervals in Lemma 41 hold, we have

$$\begin{aligned}
r_t^i &= \mathbb{E}\left[\langle\boldsymbol{\theta}, \mathbf{x}_{*,t}^i\rangle - \langle\boldsymbol{\theta}, \mathbf{x}_t^i\rangle\right] \\
&\leq \mathbb{E}\left[\langle\hat{\boldsymbol{\theta}}_{m-1}^i, \mathbf{x}_{*,t}^i\rangle - \langle\hat{\boldsymbol{\theta}}_{m-1}^i, \mathbf{x}_t^i\rangle + \beta\left\|\mathbf{x}_{*,t}^i\right\|_{(\mathbf{\Lambda}_{m-1}^i)^{-1}} + \beta\left\|\mathbf{x}_t^i\right\|_{(\mathbf{\Lambda}_{m-1}^i)^{-1}}\right] && \text{(Lemma 41)} \\
&\leq 2\gamma\mathbb{E}\left[\left\|\mathbf{x}_{*,t}^i\right\|_{(\mathbf{\Lambda}_{m-1}^i)^{-1}} + \left\|\mathbf{x}_t^i\right\|_{(\mathbf{\Lambda}_{m-1}^i)^{-1}}\right] && (\mathbf{x}_{*,t}^i \in \mathcal{X}_t^{i(m)}) \\
&\leq 4\gamma\mathbb{E}\left[\max_{\mathbf{x}\in\mathcal{X}_t^{i(m)}}\|\mathbf{x}\|_{(\mathbf{\Lambda}_{m-1}^i)^{-1}}\right] \\
&\leq 4\gamma\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i}\left[\max_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x}\|_{(\mathbf{\Lambda}_{m-1}^i)^{-1}}\right] \\
&\leq 4\gamma\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\left[\max_{\mathbf{x}\in\mathcal{X}}\|\mathbf{x}\|_{(\mathbf{\Lambda}_{m-1}^i)^{-1}}\right] \\
&\leq \frac{8\gamma}{\sqrt{N(T_{m-1}-S)}}\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\left[\max_{\mathbf{x}\in\mathcal{X}}\sqrt{\mathbf{x}^\top\left(\frac{2\lambda}{N(T_{m-1}-S)}I + \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\mathbb{E}_{\mathbf{y}\sim\pi_{m-2}^i(\mathcal{X})}[\mathbf{y}\mathbf{y}^\top]\right)^{-1}\mathbf{x}}\right] \\
&= \frac{8\gamma}{\sqrt{N(T_{m-1}-S)}}\mathbb{V}_{\mathcal{D}_{m-1}^i}^{(\frac{2\lambda}{N(T_{m-1}-S)})}(\pi_{m-2}^i), && \text{(E.30)}
\end{aligned}$$

where the third inequality follows from our established confidence intervals in Lemma 41 guaranteeing that $\mathbf{x}_{*,t}^i \in \mathcal{X}_t^{i(m)}$ for all $(i,t,m) \in [N]\times[\mathcal{T}_{m-1}+1:\mathcal{T}_m]\times[M]$ with probability

at least $1-\delta$. Now, continuing form (E.8), we bound the cumulative regret of batches $m \geq 3$, as follows:

$$\begin{aligned}
\sum_{t=\mathcal{T}_2+1}^{T}\sum_{i=1}^{N} r_t^i &\leq 2MSN + \sum_{m=3}^{M}\sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_m - S}\sum_{i=1}^{N} r_t^i \\
&\leq 2MSN + \frac{8\gamma MN(T_m - S)}{\sqrt{N(T_{m-1}-S)}}\mathbb{V}_{\mathcal{D}_{m-1}^i}^{(\frac{2\lambda}{N(T_{m-1}-S)})}(\pi_{m-2}^i) \\
&\leq 2MSN + +8\gamma M\sqrt{dN\log d\log(NT\lambda^{-1})}\sum_{m=2}^{M}\frac{T_m - S}{\sqrt{T_{m-1}-S}} \\
&\qquad\qquad \text{(Conditioned on the event in Eqn. (E.7))} \\
&= 2MSN + 8\gamma Ma\sqrt{dN\log d\log(NT\lambda^{-1})}. \qquad \text{(E.31)}
\end{aligned}$$

Next, we bound cumulative regret of the first two batches. Under Assumption 16, during the first two batches, the instantaneous regret of each agent i at any round t is at most 2. Therefore

$$\sum_{t=1}^{\mathcal{T}_2}\sum_{i=1}^{N} r_t^i \leq 2N\mathcal{T}_2 = 4a\sqrt{dN}. \qquad \text{(E.32)}$$

Note that the choice of a in the algorithm ensures that for any $M > 0$, $T_M = T$ and $\sum_{m=1}^{M} T_m \geq T_M = T$, and thus the choice of grid $\{\mathcal{T}_1, \ldots, \mathcal{T}_M\}$ is valid. If we let $M = 1 + \log\left(\frac{\log\left(\frac{N(T+S)}{d}\right)}{2} + 1\right)$, from (E.31) and (E.32), we conclude that, with probability at least $1-2\delta$, it holds that

$$\begin{aligned}
R_T &\leq 2MSN + 4\sqrt{dN(T+S)}\left(\frac{NT}{d}\right)^{\frac{1}{2(2^{M-1}-1)}} + 8\gamma M\sqrt{dNT\log d\log(NT\lambda^{-1})}\left(\frac{N(T+S)}{d}\right)^{\frac{1}{2(2^{M-1}-1)}} \\
&\leq \mathcal{O}\left(\left(\frac{N\log(dN)}{\sqrt{1/|\lambda_2|}} + \sqrt{dN\left(T + \frac{\log(dN)}{\sqrt{1/|\lambda_2|}}\right)\log d\log^2\left(\frac{KN\left(T + \frac{\log(dN)}{\sqrt{1/|\lambda_2|}}\right)}{\delta\lambda}\right)}\right)\log\log\left(\frac{NT}{d}\right)\right). \qquad \text{(E.33)}
\end{aligned}$$

□

E.3.2 Communication Step

In this section, we summarize the accelerated Chebyshev communication step, discussed above, in Algorithm 13, which follows the same steps as those of the communication algorithm

presented in [90].

Algorithm 13 Comm for Agent i

1: **Input:** x_{now}, x_{prev}, ℓ

2: **Output:** $x_{i,\text{next}}$

3: **Initialization:** $w_0 = 0, w_1 = 1/|\lambda_2|, w_r = 2w_{r-1}/|\lambda_2| - w_{r-2},\ \forall 2 \leq r \leq S$, $x_{i,\text{now}} = x_{\text{now}}$, $x_{i,\text{prev}} = x_{\text{prev}}$

4: Send $x_{i,\text{now}}$ and receive the corresponding $x_{j,\text{now}}$ to and from $j \in \mathcal{N}(i)$ // Recall that all agents run Comm in parallel.

5: $x_{i,\text{next}} = \frac{2w_{\ell-1}}{|\lambda_2|w_\ell}\mathbf{P}_{i,i}x_{i,\text{now}} + \frac{2w_{\ell-1}}{|\lambda_2|w_\ell}\sum_{j\in\mathcal{N}(i)}\mathbf{P}_{i,j}x_{j,\text{now}} - \frac{w_{\ell-2}}{w_\ell}x_{i,\text{prev}}$

Chebyshev polynomials [148] are defined as $T_0(x) = 1, T_1(x) = x$ and $T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$. Define:

$$q_\ell(\mathbf{P}) = \frac{T_\ell(\mathbf{P}/|\lambda_2|)}{T_\ell(1/|\lambda_2|)}. \tag{E.34}$$

By the properties of Chebyshev polynomial [20], it can be shown that:

$$q_{\ell+1}(\mathbf{P}) = \frac{2w_\ell}{|\lambda_2|w_{\ell+1}}\mathbf{P}q_\ell(\mathbf{P}) - \frac{w_{\ell-1}}{w_{\ell+1}}q_{\ell-1}(\mathbf{P}), \tag{E.35}$$

where $w_0 = 1, w_1 = 1/|\lambda_2|$, $w_{\ell+1} = 2w_\ell/|\lambda_2| - w_{\ell-1}$, $q_0(\mathbf{P}) = I$ and $q_1(\mathbf{P}) = \mathbf{P}$. This implies that when agents share an specific quantity, whose initial values given by agents are denoted by vector $\boldsymbol{\nu}_0 \in \mathbb{R}^N$, by using the recursive Chebyshev-accelerated updating rule, they have:

$$\boldsymbol{\nu}_{\ell+1} = \frac{2w_\ell}{|\lambda_2|w_{\ell+1}}\mathbf{P}\boldsymbol{\nu}_\ell - \frac{w_{\ell-1}}{w_{\ell+1}}\boldsymbol{\nu}_{\ell-1}. \tag{E.36}$$

In light of the above mentioned recursive procedure, the accelerated communication step is summarized in Algorithm 13 below for agent i. We denote the inputs by: 1) x_{now}, which is the quantity of interest that agent i wants to update at the current round, 2) x_{prev}, which is the estimated value for a quantity of interest that agent i updated at the previous round, and 3) ℓ which is the current round of communication. Note that inputs are scalars, however matrices and vectors also can be passed as inputs with Comm running for each of their entries.

E.4 Omitted Algorithms

In this section, we present a definition and necessary algorithms, that are borrowed from [103] and are used as subroutines in DisBE-LUCB and DecBE-LUCB.

Definition 5 ([103]). *Fix $\alpha = \log K$. For a given positive semi-definite matrix $\mathbf{M}$, we define the softmax policy $\pi^S_{\mathbf{M}}(\mathcal{X})$ over a set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ with $k \leq K$ with*

$$\pi^S_{\mathbf{M}}(\mathbf{x}_i) = \frac{(\mathbf{x}_i^\top \mathbf{M} \mathbf{x}_i)^\alpha}{\sum_{i=1}^k (\mathbf{x}_i^\top \mathbf{M} \mathbf{x}_i)^\alpha}. \tag{E.37}$$

Now, suppose we are given a set $\mathcal{M} = \left\{(p_i, \mathbf{M}_i)\right\}_{i=1}^n$ such that $p_i \geq 0$ and $\sum_{i=1}^n p_i = 1$. We define the mixed-softmax policy $\pi^{MS}_{\mathcal{M}}(\mathcal{X})$ over $\mathcal{X}$ as

$$\pi^{MS}_{\mathcal{M}}(\mathbf{x}_i) = \begin{cases} \pi^G(\mathcal{X}), & \text{with probability } 1/2, \\ \pi^S_{\mathbf{M}_i}(\mathcal{X}), & \text{with probability } p_i/2, \end{cases} \tag{E.38}$$

where $\pi^G(\mathcal{X})$ is called G-optimal design and is the minimizer of $g(\pi) = \max_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x}\|^2_{\mathbf{V}(\pi)^{-1}}$, where $\mathbf{V}(\pi) = \sum_{\mathbf{x} \in \mathcal{X}} \pi(\mathbf{x}) \mathbf{x}\mathbf{x}^\top$; see Section 21 in [78] for details.

Algorithm 14 ExplorationPolicy

1: **Input:** λ, $\mathcal{S} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_L\}$

2: **Output:** A mixed-softmax policy π Using Algorithm 15 find a core $\mathcal{C} \subseteq \mathcal{S}$ such that

$$\max_{\mathcal{X}_i \in \mathcal{C}, \mathbf{x} \in \mathcal{X}_i} \mathbf{x}^\top \mathbf{A}(\mathcal{C})^{-1} \mathbf{x} > d^5 \tag{E.39}$$

and

$$\frac{|\mathcal{C}|}{L} < 1 - \mathcal{O}(d^{-2} \log \lambda^{-1}) \tag{E.40}$$

where $\mathbf{A}(\mathcal{C}) := \lambda \mathbf{I} + \frac{1}{L} \sum_{\mathcal{X}_i \in \mathcal{C}} \mathbb{E}_{\mathbf{x} \sim \pi^{\mathrm{G}}(\mathcal{X}_i)}[\mathbf{x}\mathbf{x}^\top]$, and for any set $\mathcal{X} \subset \mathbb{R}^d$, $\pi^{\mathrm{G}}(\mathcal{X})$ is called G-optimal design and is the maximizer of $g(\pi) = \max_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x}\|^2_{\mathbf{V}(\pi)^{-1}}$, where $\mathbf{V}(\pi) = \sum_{\mathbf{x} \in \mathcal{X}} \pi(\mathbf{x}) \mathbf{x}\mathbf{x}^\top$.

3: Return the mixed-softmax policy π by calling MixedSoftMax$(\lambda, \mathcal{C})$.

Algorithm 15 CoreIdentification (Algorithm 4 in [103])

Input: λ, $\mathcal{S} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_L\}$

Output: A core set $\mathcal{C} \subseteq \mathcal{S}$

Initialization: $\mathcal{C}_1 = \mathcal{S}$

for $\xi = 1, 2, \ldots$ **do**

  if $\max_{\mathcal{X}_i \in \mathcal{C}_\xi, \mathbf{x} \in \mathcal{X}_i} \mathbf{x}^\top \mathbf{A}(\mathcal{C}_\xi)^{-1}\mathbf{x} > d^5$ **then**

    Return $\mathcal{C}_\xi$.

  else

$$\mathcal{C}_{\xi+1} = \left\{ \mathcal{X}_i \in \mathcal{C}_\xi : \max_{\mathbf{x} \in \mathcal{X}_i} \mathbf{x}^\top \mathbf{A}(\mathcal{C}_\xi)^{-1}\mathbf{x} \leq \frac{1}{2} d^5 \right\},$$

  where $\mathbf{A}(\mathcal{C}) := \lambda \mathbf{I} + \frac{1}{L} \sum_{\mathcal{X}_i \in \mathcal{C}} \mathbb{E}_{\mathbf{x} \sim \pi^{\mathrm{G}}(\mathcal{X}_i)}[\mathbf{x}\mathbf{x}^\top]$, and for any set $\mathcal{X} \subset \mathbb{R}^d$, $\pi^{\mathrm{G}}(\mathcal{X})$ is called G-optimal design and is the maximizer of $g(\pi) = \max_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x}\|^2_{\mathbf{V}(\pi)^{-1}}$, where $\mathbf{V}(\pi) = \sum_{\mathbf{x} \in \mathcal{X}} \pi(\mathbf{x}) \mathbf{x}\mathbf{x}^\top$.

  end if

end for

Algorithm 16 MixedSoftMax

Input: λ, $\mathcal{S} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_L\}$

Output: A mixed-softmax policy π

Initialization: $Q = 2d^2 \log d$, $\mathcal{X}_{(i-1)L+j} = \mathcal{X}_j$, $\forall (i,j) \in [Q] \times L$, $\mathbf{U}_0 = \lambda QLI + \frac{Q}{2} \sum_{i=1}^{L} \mathbb{E}_{\mathbf{x} \sim \pi^{\mathrm{G}}(\mathcal{X}_i)}[\mathbf{x}\mathbf{x}^\top]$, $n = 1$, $\tau_n = \emptyset$, $\mathbf{W}_n = \mathbf{U}_0$

for $s = 1, \ldots, QL$ **do**

  $\tau_n = \tau_n \cup \{s\}$

  $\mathbf{U}_s = \mathbf{U}_{s-1} + \mathbb{E}_{\mathbf{x} \sim \pi^{\mathrm{S}}_{\mathbf{W}_n^{-1}}(\mathcal{X}_s)}[\mathbf{x}\mathbf{x}^\top]$, where $\pi^{\mathrm{S}}_{\mathbf{W}_n^{-1}}(\mathcal{X}_s)$ is computed as in Definition 5.

  if $\frac{\det \mathbf{U}_s}{\det \mathbf{W}_n} > 2$ **then**

    $n = n + 1$, $\tau_n = \emptyset$, $\mathbf{W}_n = \mathbf{U}_s$

  end if

end for

$p_i = \frac{\mathbb{I}\{|\tau_i| \geq L\}|\tau_i|}{\sum_{i=1}^{n} \mathbb{I}\{|\tau_i| \geq L\}|\tau_i|}$ and $\mathbf{M}_i = QL\mathbf{W}_i^{-1}$, $\forall i \in [n]$

Return the mixed-softmax policy with parameters $\mathcal{M} = \left\{(p_i, \mathbf{M}_i)\right\}_{i=1}^{n}$ as in Definition 5.

E.5 Auxiliary Lemmas

Lemma 42 ([125], Theorem 5.1.1)**.** *Consider a finite sequence $\mathbf{X}_k$ of independent, random, Hermitian matrices with common dimension d. Assume that $0 \leq \lambda_{min}(\mathbf{X}_k)$ and $\lambda_{max}(\mathbf{X}_k) \leq L$ for each index k. Introduce the random matrix*

$$\mathbf{Y} = \sum_{k=1}^{n} \mathbf{X}_k \tag{E.41}$$

Define the minimum eigenvalue μ_{min} and maximum eigenvalue μ_{max} of the expectation $\mathbb{E}[\mathbf{Y}]$:

$$\mu_{min} = \lambda_{min}(\mathbb{E}[\mathbf{Y}]), \quad \mu_{max} = \lambda_{max}(\mathbb{E}[\mathbf{Y}]). \tag{E.42}$$

Then

$$\mathbb{P}\left(\lambda_{min}(\mathbf{Y}) \leq (1-\varepsilon)\mu_{min}\right) \leq d \left(\frac{\exp(-\varepsilon)}{(1-\varepsilon)^{1-\varepsilon}}\right)^{\frac{\mu_{min}}{L}}, \quad \text{for } \varepsilon \in [0,1) \tag{E.43}$$

$$\mathbb{P}\left(\lambda_{max}(\mathbf{Y}) \geq (1+\varepsilon)\mu_{max}\right) \leq d \left(\frac{\exp(\varepsilon)}{(1+\varepsilon)^{1+\varepsilon}}\right)^{\frac{\mu_{max}}{L}}, \quad \text{for } \varepsilon \geq 0. \tag{E.44}$$

Lemma 43. *Suppose $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \sim \mathcal{D}$ are d-dimensional vectors that are i.i.d. drawn from a distribution $\mathcal{D}$ and $\|\mathbf{x}_k\|_2 \le L$ for all $k \in [n]$ almost surely. Let $\gamma = \lambda_{min}\left(\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top]\right) > 0$ be the smallest eigenvalue of the co-variance matrix. We have that*

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k\mathbf{x}_k^\top \preceq 2\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top]\right) \ge 1 - d\exp\left(\frac{-\gamma n}{3}\right). \tag{E.45}$$

Proof. Let $\mathbf{\Sigma} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top]$ and $\mathbf{y}_k = \mathbf{\Sigma}^{\frac{-1}{2}}\mathbf{x}_k$ for all $k \in [n]$. Also, we have $\lambda_{\max}(\mathbf{y}_k\mathbf{y}_k^\top) = \|\mathbf{y}_k\|_2^2 \le \frac{1}{\gamma}$ almost surely, and $\mathbb{E}[\mathbf{y}_k\mathbf{y}_k^\top] = I$. Therefore, plugging $\varepsilon = 1$ in (E.44), we have

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k\mathbf{x}_k^\top \preceq 2\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top]\right) &= \mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbf{y}_k\mathbf{y}_k^\top \preceq 2\mathbf{\Sigma}^{\frac{-1}{2}}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top]\mathbf{\Sigma}^{\frac{-1}{2}}\right) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbf{y}_k\mathbf{y}_k^\top \preceq 2I\right) \\
&= \mathbb{P}\left(\lambda_{\max}\left(\sum_{k=1}^{n}\mathbf{y}_k\mathbf{y}_k^\top\right) \le 2n\right) \\
&\ge 1 - d\left(\frac{e}{4}\right)^{n\gamma} \ge 1 - d\exp\left(\frac{-\gamma n}{3}\right).
\end{aligned} \tag{E.46}$$

□

Lemma 44. *Suppose $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \sim \mathcal{D}$ are d-dimensional vectors that are i.i.d. drawn from a distribution $\mathcal{D}$ and $\|\mathbf{x}_k\|_2 \le 1$ for all $k \in [n]$ almost surely. For any cutoff level $\gamma > 0$, we have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k\mathbf{x}_k^\top \preceq 2\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top] + 6\gamma I\right) \ge 1 - 2d\exp\left(\frac{-\gamma n}{3}\right). \tag{E.47}$$

Proof. Suppose $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top] = \sum_{i=1}^{d}\lambda_i\boldsymbol{\nu}_i\boldsymbol{\nu}_i^\top$, where $\{\boldsymbol{\nu}_i\}_{i=1}^{d}$ is a set of orthonormal basis. Let $\mathbb{P}_+ = \sum_{i=1}^{d}\boldsymbol{\nu}_i\boldsymbol{\nu}_i^\top\mathbb{1}(\lambda_i \ge \gamma)$ and $\mathbb{P}_- = \sum_{i=1}^{d}\boldsymbol{\nu}_i\boldsymbol{\nu}_i^\top\mathbb{1}(\lambda_i < \gamma)$, so that $\mathbb{P}_+\mathbb{P}_- = I$. We observe that the eigenvalues of $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbb{P}_+\mathbf{x}\mathbf{x}^\top\mathbb{P}_+^\top]$ are greater than or equal to γ when restricted to the space spanned by the $\mathbb{P}_+$. Therefore, by Lemmas 43 and 42 (Eqn. (E.44)), we respectively

have

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbb{P}_{+}\mathbf{x}_k\mathbf{x}_k^{\top}\mathbb{P}_{+}^{\top} \preceq 2\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbb{P}_{+}\mathbf{x}\mathbf{x}^{\top}\mathbb{P}_{+}^{\top}]\right) \geq 1 - d\exp\left(\frac{-\gamma n}{3}\right) \tag{E.48}$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbb{P}_{-}\mathbf{x}_k\mathbf{x}_k^{\top}\mathbb{P}_{-}^{\top} \preceq 2\gamma I\right) \geq 1 - d\exp\left(\frac{-\gamma n}{3}\right). \tag{E.49}$$

Now, we observe that

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k\mathbf{x}_k^{\top} &= \frac{1}{n}\left(\sum_{k=1}^{n}\mathbb{P}_{+}\mathbf{x}_k\mathbf{x}_k^{\top}\mathbb{P}_{+}^{\top} + \sum_{k=1}^{n}\mathbb{P}_{+}\mathbf{x}_k\mathbf{x}_k^{\top}\mathbb{P}_{-}^{\top} + \sum_{k=1}^{n}\mathbb{P}_{-}\mathbf{x}_k\mathbf{x}_k^{\top}\mathbb{P}_{+}^{\top} + \sum_{k=1}^{n}\mathbb{P}_{-}\mathbf{x}_k\mathbf{x}_k^{\top}\mathbb{P}_{-}^{\top}\right) \\
&= \frac{1}{n}\left(\sum_{k=1}^{n}\mathbb{P}_{+}\mathbf{x}_k\mathbf{x}_k^{\top}\mathbb{P}_{+}^{\top} + \sum_{k=1}^{n}\mathbb{P}_{+}\mathbb{P}_{+}\mathbb{P}_{-}\mathbf{x}_k\mathbf{x}_k^{\top}\mathbb{P}_{-}^{\top} + \sum_{k=1}^{n}\mathbb{P}_{-}\mathbf{x}_k\mathbf{x}_k^{\top}\mathbb{P}_{-}^{\top}\mathbb{P}_{+}^{\top}\mathbb{P}_{+}^{\top} + \sum_{k=1}^{n}\mathbb{P}_{-}\mathbf{x}_k\mathbf{x}_k^{\top}\mathbb{P}_{-}^{\top}\right) \\
&\preceq \frac{1}{n}\left(\sum_{k=1}^{n}\mathbb{P}_{+}\mathbf{x}_k\mathbf{x}_k^{\top}\mathbb{P}_{+}^{\top} + \sum_{k=1}^{n}\mathbb{P}_{-}\mathbf{x}_k\mathbf{x}_k^{\top}\mathbb{P}_{-}^{\top} + \sum_{k=1}^{n}\mathbb{P}_{-}\mathbf{x}_k\mathbf{x}_k^{\top}\mathbb{P}_{-}^{\top} + \sum_{k=1}^{n}\mathbb{P}_{-}\mathbf{x}_k\mathbf{x}_k^{\top}\mathbb{P}_{-}^{\top}\right) \\
&= \frac{1}{n}\sum_{k=1}^{n}\mathbb{P}_{+}\mathbf{x}_k\mathbf{x}_k^{\top}\mathbb{P}_{+}^{\top} + \frac{3}{n}\sum_{k=1}^{n}\mathbb{P}_{-}\mathbf{x}_k\mathbf{x}_k^{\top}\mathbb{P}_{-}^{\top}
\end{aligned} \tag{E.50}$$

Also, note that

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbb{P}_{+}\mathbf{x}\mathbf{x}^{\top}\mathbb{P}_{+}^{\top}] = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbf{x}\mathbf{x}^{\top} - \mathbb{P}_{+}\mathbf{x}\mathbf{x}^{\top}\mathbb{P}_{-}^{\top} - \mathbb{P}_{-}\mathbf{x}\mathbf{x}^{\top}\mathbb{P}_{+}^{\top} - \mathbb{P}_{-}\mathbf{x}\mathbf{x}^{\top}\mathbb{P}_{-}^{\top}\right] \preceq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbf{x}\mathbf{x}^{\top}\right]. \tag{E.51}$$

Therefore, combining (E.49) and (E.50) and (E.51), we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k\mathbf{x}_k^{\top} \preceq 2\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^{\top}] + 6\gamma I\right) \geq 1 - 2d\exp\left(\frac{-\gamma n}{3}\right). \tag{E.52}$$

□

REFERENCES

[1] Yasin Abbasi-Yadkori and Gergely Neu. Online learning in mdps with side information. *arXiv preprint arXiv:1406.6812*, 2014.

[2] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[3] Naoki Abe, Prem Melville, Cezar Pendus, Chandan K Reddy, David L Jensen, Vince P Thomas, James J Bennett, Gary F Anderson, Brent R Cooley, Melissa Kowalczyk, et al. Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84, 2010.

[4] David Abel, Dilip Arumugam, Lucas Lehnert, and Michael Littman. State abstractions for lifelong reinforcement learning. In *International Conference on Machine Learning*, pages 10–19. PMLR, 2018.

[5] David Abel, Yuu Jinnai, Sophie Yue Guo, George Konidaris, and Michael Littman. Policy and value transfer in lifelong reinforcement learning. In *International Conference on Machine Learning*, pages 20–29. PMLR, 2018.

[6] Axel Abels, Diederik Roijers, Tom Lenaerts, Ann Nowé, and Denis Steckelmacher. Dynamic weights in multi-objective deep reinforcement learning. In *International Conference on Machine Learning*, pages 11–20. PMLR, 2019.

[7] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 22–31. JMLR. org, 2017.

[8] Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, et al. Making contextual decisions with low technical debt. *arXiv preprint arXiv:1606.03966*, 2016.

[9] Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3450–3458. Curran Associates, Inc., 2016.

[10] A. K. Akametalu, J. F. Fisac, J. H. Gillula, S. Kaynama, M. N. Zeilinger, and C. J. Tomlin. Reachability-based safe learning with gaussian processes. In *53rd IEEE Conference on Decision and Control*, pages 1424–1431, Dec 2014.

[11] Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems*, pages 9252–9262, 2019.

[12] Sanae Amani, Tor Lattimore, András György, and Lin Yang. Distributed contextual linear bandits with minimax optimal communication cost. In *International Conference on Machine Learning*, pages 691–717. PMLR, 2023.

[13] Sanae Amani and Christos Thrampoulidis. Decentralized multi-agent linear bandits with safety constraints. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6627–6635, May 2021.

[14] Sanae Amani, Christos Thrampoulidis, and Lin Yang. Safe reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 243–253. PMLR, 2021.

[15] Sanae Amani, Christos Thrampoulidis, and Lin F. Yang. Safe reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 243–253. PMLR, 2021.

[16] Sanae Amani, Lin Yang, and Ching-An Cheng. Provably efficient lifelong reinforcement learning with linear representation. In *The Eleventh International Conference on Learning Representations*, 2022.

[17] Sanae Amani and Lin F Yang. Doubly pessimistic algorithms for strictly safe off-policy optimization. In *2022 56th Annual Conference on Information Sciences and Systems (CISS)*, pages 113–118. IEEE, 2022.

[18] Haitham Bou Ammar, Eric Eaton, Paul Ruvolo, and Matthew Taylor. Online multi-task learning for policy gradient methods. In *International conference on machine learning*, pages 1206–1214. PMLR, 2014.

[19] Haitham Bou Ammar, Rasul Tutunov, and Eric Eaton. Safe policy search for lifelong reinforcement learning with sublinear regret. In *International Conference on Machine Learning*, pages 2361–2369. PMLR, 2015.

[20] Mario Arioli and Jennifer Scott. Chebyshev acceleration of iterative refinement. *Numerical Algorithms*, 66(3):591–608, 2014.

[21] Anil Aswani, Humberto Gonzalez, S Shankar Sastry, and Claire Tomlin. Provably safe and robust learning-based model predictive control. *Automatica*, 49(5):1216–1226, 2013.

[22] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

[23] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Mach. Learn.*, 47(2-3):235–256, May 2002.

[24] Orly Avner and Shie Mannor. Multi-user communication networks: A coordinated multi-armed bandit approach. *IEEE/ACM Transactions on Networking*, 27(6):2192–2207, 2019.

[25] A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 207–216, Oct 2013.

[26] Ashwinkumar Badanidiyuru, John Langford, and Aleksandrs Slivkins. Resourceful contextual bandits. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 1109–1134, Barcelona, Spain, 13–15 Jun 2014. PMLR.

[27] Yu Bai, Tengyang Xie, Nan Jiang, and Yu-Xiang Wang. Provably efficient q-learning with low switching cost. *arXiv preprint arXiv:1905.12849*, 2019.

[28] Felix Berkenkamp, Andreas Krause, and Angela P Schoellig. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *arXiv preprint arXiv:1602.04450*, 2016.

[29] Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *Advances in neural information processing systems*, pages 908–918, 2017.

[30] Dimitri P Bertsekas et al. *Dynamic programming and optimal control: Vol. 1*. Athena scientific Belmont, 2000.

[31] Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1-3):33–57, 1996.

[32] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[33] Emma Brunskill and Lihong Li. Sample complexity of multi-task reinforcement learning. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 122–131, 2013.

[34] Emma Brunskill and Lihong Li. Pac-inspired option discovery in lifelong reinforcement learning. In *International conference on machine learning*, pages 316–324. PMLR, 2014.

[35] Emma Brunskill and Lihong Li. The online coupon-collector problem and its application to lifelong reinforcement learning. *arXiv preprint arXiv:1506.03379*, 2015.

[36] Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

[37] Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. In *Advances in neural information processing systems*, pages 8092–8101, 2018.

[38] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

[39] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

[40] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

[41] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

[42] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo R Jovanović. Provably efficient safe exploration via primal-dual policy optimization. *arXiv preprint arXiv:2003.00534*, 2020.

[43] Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33, 2020.

[44] Simon S Du, Ruosong Wang, Mengdi Wang, and Lin F Yang. Continuous control with contexts, provably. *arXiv preprint arXiv:1910.13614*, 2019.

[45] Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

[46] Abhimanyu Dubey and AlexSandy' Pentland. Differentially-private federated linear bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

[47] John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2011.

[48] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

[49] Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[50] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

[51] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062. PMLR, 2019.

[52] Minbo Gao, Tianle Xie, Simon S Du, and Lin F Yang. A provably efficient algorithm for linear markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494*, 2021.

[53] Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *arXiv preprint arXiv:1904.01763*, 2019.

[54] Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. Conservative exploration in reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1431–1441. PMLR, 2020.

[55] Mohammad Ghavamzadeh, Marek Petrik, and Yinlam Chow. Safe policy improvement by minimizing robust baseline regret. *Advances in Neural Information Processing Systems*, 29:2298–2306, 2016.

[56] J. H. Gillulay and C. J. Tomlin. Guaranteed safe online learning of a bounded system. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2979–2984, Sep. 2011.

[57] Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

[58] Yanjun Han, Zhengqing Zhou, Zhengyuan Zhou, Jose Blanchet, Peter W Glynn, and Yinyu Ye. Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321*, 2020.

[59] Osama A Hanna, Lin F Yang, and Christina Fragouli. Contexts can be cheap: Solving stochastic contextual bandits with linear bandit algorithms. *arXiv preprint arXiv:2211.05632*, 2022.

[60] Osama A Hanna, Lin F Yang, and Christina Fragouli. Learning in distributed contextual linear bandits without sharing the context. *arXiv preprint arXiv:2206.04180*, 2022.

[61] Matteo Hessel, Hubert Soyer, Lasse Espeholt, Wojciech Czarnecki, Simon Schmitt, and Hado van Hasselt. Multi-task deep reinforcement learning with popart. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3796–3803, 2019.

[62] Ruiquan Huang, Weiqiang Wu, Jing Yang, and Cong Shen. Federated linear contextual bandits. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[63] David Isele, Mohammad Rostami, and Eric Eaton. Using task features for zero-shot knowledge transfer in lifelong learning. In *IJCAI*, volume 16, pages 1620–1626, 2016.

[64] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

[65] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? *arXiv preprint arXiv:2012.15085*, 2020.

[66] Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33:15312–15325, 2020.

[67] Krishna C Kalagarla, Rahul Jain, and Pierluigi Nuzzo. A sample-efficient algorithm for episodic finite-horizon mdp with constraints. *arXiv preprint arXiv:2009.11348*, 2020.

[68] Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi, and Benjamin Van Roy. Conservative contextual linear bandits. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3910–3919. Curran Associates, Inc., 2017.

[69] Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *arXiv preprint arXiv:2012.13490*, 2020.

[70] Nathan Korda, Balázs Szörényi, and Li Shuai. Distributed clustering of linear bandits in peer to peer networks. In *Journal of machine learning research workshop and conference proceedings*, volume 48, pages 1301–1309. International Machine Learning Societ, 2016.

[71] Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *arXiv preprint arXiv:1906.00949*, 2019.

[72] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

[73] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 167–172. IEEE, 2016.

[74] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. On distributed cooperative decision-making in multiarmed bandits. In *2016 European Control Conference (ECC)*, pages 243–248. IEEE, 2016.

[75] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Social imitation in cooperative multiarmed bandits: Partition-based algorithms with strictly local information. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 5239–5244. IEEE, 2018.

[76] Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer, 2012.

[77] Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pages 3652–3661. PMLR, 2019.

[78] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[79] Erwan Lecarpentier, David Abel, Kavosh Asadi, Yuu Jinnai, Emmanuel Rachelson, and Michael L Littman. Lipschitz lifelong reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8270–8278, 2021.

[80] Orin Levy and Yishay Mansour. Learning efficiently function approximation for contextual mdp. *arXiv preprint arXiv:2203.00995*, 2022.

[81] Chuanhao Li, Huazheng Wang, Mengdi Wang, and Hongning Wang. Communication efficient distributed learning for kernelized contextual bandits. *arXiv preprint arXiv:2206.04835*, 2022.

[82] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[83] Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2071–2080. JMLR. org, 2017.

[84] Shuai Li, Fei Hao, Mei Li, and Hee-Cheol Kim. Medicine rating prediction and recommendation in mobile social networks. In *International conference on grid and pervasive computing*, pages 216–223. Springer, 2013.

[85] Keqin Liu and Qing Zhao. Decentralized multi-armed bandit with multiple distributed players. In *2010 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2010.

[86] Keqin Liu and Qing Zhao. Distributed learning in cognitive radio networks: Multi-armed bandit with distributed multiple players. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3010–3013. IEEE, 2010.

[87] Keqin Liu and Qing Zhao. Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*, 58(11):5667–5681, 2010.

[88] Tao Liu, Ruida Zhou, Dileep Kalathil, PR Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. *arXiv preprint arXiv:2106.02684*, 2021.

[89] Nancy A Lynch. *Distributed algorithms*. Elsevier, 1996.

[90] David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 4531–4542, 2019.

[91] Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, pages 597–618. PMLR, 2018.

[92] Aditya Modi and Ambuj Tewari. No-regret exploration in contextual reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pages 829–838. PMLR, 2020.

[93] Teodor Mihai Moldovan and Pieter Abbeel. Safe exploration in markov decision processes. *arXiv preprint arXiv:1205.4810*, 2012.

[94] Ahmadreza Moradipari, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Safe linear thompson sampling. *arXiv preprint arXiv:1911.02156*, 2019.

[95] Chris J Ostafew, Angela P Schoellig, and Timothy D Barfoot. Robust constrained learning-based nmpc enabling reliable mobile robot path tracking. *The International Journal of Robotics Research*, 35(13):1547–1563, 2016.

[96] Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. *arXiv preprint arXiv:2006.10185*, 2020.

[97] Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 2827–2835. PMLR, 2021.

[98] Santiago Paternain, Miguel Calvo-Fullana, Luiz FO Chamon, and Alejandro Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *arXiv preprint arXiv:1911.09101*, 2019.

[99] Santiago Paternain, Luiz FO Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. *arXiv preprint arXiv:1910.13393*, 2019.

[100] Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual optimization: Stochastically constrained markov decision processes with adversarial losses and unknown transitions. *arXiv preprint arXiv:2003.00660*, 2020.

[101] Rafael Rafailov, Tianhe Yu, Aravind Rajeswaran, and Chelsea Finn. Offline reinforcement learning from images with latent space models. *arXiv preprint arXiv:2012.11547*, 2020.

[102] Nicholas Roy, Ingmar Posner, Tim Barfoot, Philippe Beaudoin, Yoshua Bengio, Jeannette Bohg, Oliver Brock, Isabelle Depatie, Dieter Fox, Dan Koditschek, et al. From machine learning to robotics: Challenges and opportunities for embodied intelligence. *arXiv preprint arXiv:2110.15245*, 2021.

[103] Yufei Ruan, Jiaqi Yang, and Yuan Zhou. Linear bandits with limited adaptivity and learning distributional optimal design. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 74–87, 2021.

[104] Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

[105] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[106] Jens Schreiter, Duy Nguyen-Tuong, Mona Eberts, Bastian Bischoff, Heiner Markert, and Marc Toussaint. Safe exploration for active learning with gaussian processes. In Albert Bifet, Michael May, Bianca Zadrozny, Ricard Gavalda, Dino Pedreschi, Francesco Bonchi, Jaime Cardoso, and Myra Spiliopoulou, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 133–149, Cham, 2015. Springer International Publishing.

[107] Kevin Seaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3027–3036. JMLR. org, 2017.

[108] Daniel L Silver, Qiang Yang, and Lianghao Li. Lifelong machine learning systems: Beyond learning algorithms. In *2013 AAAI spring symposium series*, 2013.

[109] Shagun Sodhani, Amy Zhang, and Joelle Pineau. Multi-task reinforcement learning with context-based representations. In *International Conference on Machine Learning*, pages 9767–9779. PMLR, 2021.

[110] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. 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*, pages 1015–1022. Omnipress, 2010.

[111] Krishnan Srinivasan, Benjamin Eysenbach, Sehoon Ha, Jie Tan, and Chelsea Finn. Learning to be safe: Deep rl with a safety critic. *arXiv preprint arXiv:2010.14603*, 2020.

[112] Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning*, pages 9133–9143. PMLR, 2020.

[113] Yanan Sui, Joel Burdick, Yisong Yue, et al. Stagewise safe bayesian optimization with gaussian processes. In *International Conference on Machine Learning*, pages 4788–4796, 2018.

[114] Yanan Sui, Alkis Gotovos, Joel W. Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 997–1005. JMLR.org, 2015.

[115] Yanan Sui, Vincent Zhuang, Joel W Burdick, and Yisong Yue. Stagewise safe bayesian optimization with gaussian processes. *arXiv preprint arXiv:1806.07555*, 2018.

[116] Yanchao Sun, Xiangyu Yin, and Furong Huang. Temple: Learning template of transitions for sample efficient multi-task rl. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9765–9773, 2021.

[117] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[118] Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

[119] Balázs Szörényi, Róbert Busa-Fekete, István Hegedűs, Róbert Ormándi, Márk Jelasity, and Balázs Kégl. Gossip-based distributed stochastic bandit algorithms. In *Journal of Machine Learning Research Workshop and Conference Proceedings*, volume 2, pages 1056–1064. International Machine Learning Societ, 2013.

[120] Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*, 2018.

[121] Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.

[122] Brijen Thananjeyan, Ashwin Balakrishna, Suraj Nair, Michael Luo, Krishnan Srinivasan, Minho Hwang, Joseph E Gonzalez, Julian Ibarz, Chelsea Finn, and Ken Goldberg. Recovery rl: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, 6(3):4915–4922, 2021.

[123] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, pages 2380–2388. PMLR, 2015.

[124] Sebastian Thrun and Tom M Mitchell. Lifelong robot learning. *Robotics and autonomous systems*, 15(1-2):25–46, 1995.

[125] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[126] Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe exploration in finite markov decision processes with gaussian processes. *arXiv preprint arXiv:1606.04753*, 2016.

[127] Matteo Turchetta, Andrey Kolobov, Shital Shah, Andreas Krause, and Alekh Agarwal. Safe reinforcement learning via curriculum induction. *arXiv preprint arXiv:2006.12136*, 2020.

[128] Ilnura Usmanova, Andreas Krause, and Maryam Kamgarpour. Safe convex learning under uncertain constraints. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2106–2114. PMLR, 16–18 Apr 2019.

[129] Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained markov decision processes. In *International Conference on Machine Learning*, pages 9797–9806. PMLR, 2020.

[130] Akifumi Wachi, Yanan Sui, Yisong Yue, and Masahiro Ono. Safe exploration and optimization of constrained mdps using gaussian processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[131] Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

[132] Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *Advances in Neural Information Processing Systems*, 34:13524–13536, 2021.

[133] Yining Wang, Ruosong Wang, Simon Shaolei Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2020.

[134] Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang. Distributed bandit learning: How much communication is needed to achieve (near) optimal regret. *arXiv preprint arXiv:1904.06309*, 2019.

[135] Aaron Wilson, Alan Fern, Soumya Ray, and Prasad Tadepalli. Multi-task reinforcement learning: a hierarchical bayesian approach. In *Proceedings of the 24th international conference on Machine learning*, pages 1015–1022, 2007.

[136] Huasen Wu, R. Srikant, Xin Liu, and Chong Jiang. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 433–441. Curran Associates, Inc., 2015.

[137] Jingfeng Wu, Vladimir Braverman, and Lin Yang. Accommodating picky customers: Regret bound and exploration complexity for multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, 34:13112–13124, 2021.

[138] Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1254–1262. JMLR.org, 2016.

[139] Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.

[140] Annie Xie and Chelsea Finn. Lifelong robotic reinforcement learning by retaining experiences. *arXiv preprint arXiv:2109.09180*, 2021.

[141] Tengyu Xu, Yingbin Liang, and Guanghui Lan. A primal approach to constrained policy optimization: Global optimality and finite-time analysis. *arXiv preprint arXiv:2011.05869*, 2020.

[142] Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

[143] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. *Advances in Neural Information Processing Systems*, 33, 2020.

[144] Ruihan Yang, Huazhe Xu, Yi Wu, and Xiaolong Wang. Multi-task reinforcement learning with soft modularization. *Advances in Neural Information Processing Systems*, 33:4767–4777, 2020.

[145] Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J Ramadge. Projection-based constrained policy optimization. *arXiv preprint arXiv:2010.03152*, 2020.

[146] Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1567–1575. PMLR, 2021.

[147] Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3948–3958. PMLR, 2020.

[148] David M Young. *Iterative solution of large linear systems*. Elsevier, 2014.

[149] Ming Yu, Zhuoran Yang, Mladen Kolar, and Zhaoran Wang. Convergent policy optimization for safe reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 3127–3139, 2019.

[150] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

[151] Yusen Zhan, Haitham Bou Ammar, and Matthew E Taylor. Scalable lifelong reinforcement learning. *Pattern Recognition*, 72:407–418, 2017.

[152] Chicheng Zhang and Zhi Wang. Provably efficient multi-task reinforcement learning with model transfer. *Advances in Neural Information Processing Systems*, 34, 2021.

[153] Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020.

[154] Liyuan Zheng and Lillian J Ratliff. Constrained upper confidence reinforcement learning. *arXiv preprint arXiv:2001.09377*, 2020.