University of California
Santa Barbara

Learning and Pricing Algorithms for Human-Cyber-Physical Systems

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Electrical and Computer Engineering

by

Ahmadreza Moradipari

Committee in charge:

Professor Mahnoosh Alizadeh, Chair
Professor Christos Thrampoulidis
Professor Ramtin Pedarsani
Professor João P. Hespanha

December 2022

The Dissertation of Ahmadreza Moradipari is approved.

Professor Christos Thrampoulidis

Professor Ramtin Pedarsani

Professor João P. Hespanha

Professor Mahnoosh Alizadeh, Committee Chair

November 2022

Learning and Pricing Algorithms for Human-Cyber-Physical Systems

by

Ahmadreza Moradipari

To my Mom and Dad.

Acknowledgements

First and Foremost, I would like to express my sincerest gratitude and appreciation to my advisor, Prof. Mahnoosh Alizadeh. Without her guidance and persistent help this thesis would not have been possible. I was always amazed by her thorough insight in research and finding deep connections between abstract mathematical frameworks and practical engineering problems. I was also inspired by her patience in teaching me the academic writing from the scratch. She is also extremely kind and understanding. She genuinely cares about the well-being of her students. She was not only my academic advisor but also a great friend over the past years. I simply could not wish for a better advisor.

I would like to thank Prof. Ramtin Pedarsani. I always admire him as an academic person and as a friend. He is kind, supportive and very insightful in mathematically formulating the practical engineering problems. His Stochastic Processes course was my most favourite course at UCSB, which significantly helped me with my research. I also thank him for his chats, stories, and motivational conversations.

I have been very fortunate to collaborate with Prof. Christos Thrampoulidis. I truly appreciate his guidance and his amazing ability to come up with simple tricks to solve complicated problems. Chapter 4 and 5 of this thesis is the result of collaboration with Christos.

I would like to extend my gratitude to my thesis committee member Prof. Joao Hespanha for his insights and great feedback.

I would also like to thank Dr. Mohammad Ghavamzadeh. I truly appreciate his guidance and I learned a lot from his great experience and his attitude towards research. I would like to thank Dr. Yasin Abbasi-Yadkori. He has been very helpful to me with his knowledge and patience. Chapter 6 is the result of collaboration with Mohammad

and Yasin.

I also thank Dr. David Isele with whom I collaborated as an intern at Honda Research Institute. I had a great experience doing research with him, and leaned many practical topics on autonomous driving research.

Last but not least, I owe this thesis to my beloved family. Ehsan, I thank you for your consistent support and help in all my life. There are no words that can fully express my gratitude to you. Hossein, even though we have been away for the last few years, I always feel your support. You taught me how to approach scientific problems and the right path to become an engineer. Yasaman, I thank you for your support and kindness. Even though you are my younger sister, I have learned a lot from you. My dad, who genuinely loves science. He taught me how to be persistence in working on solving problems and not to give up easily. You are the kindest person I know in my life. My mom, whose sacrifice and support helped me to get this far. Even though we have been away for the last few years, I have never felt that you are not by my side. This thesis is dedicated to my parents.

Curriculum Vitæ

Ahmadreza Moradipari

Education

2022	Ph.D. in Electrical and Computer Engineering (Expected), University of California, Santa Barbara.
2017	B.S. in Electrical Engineering, Sharif University of Technology.

Publications

Journal Publications

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

2. **A. Moradipari**, N. Tucker, and M. Alizadeh, "Mobility-Aware Electric Vehicle Fast Charging Load Models with Geographical Price Variations", *IEEE Transactions on Transportation Electrification*, 2020.

3. **A. Moradipari**, and M. Alizadeh, "Pricing and Routing Mechanisms for Differentiated Services in an Electric Vehicle Public Charging Station Network", *IEEE Transactions on Smart Grid*, 2019.

4. N. Tucker, **A. Moradipari**, and M. Alizadeh, "Constrained Thompson Sampling for Real-Time Electricity Pricing with Grid Reliability Constraints", *IEEE Transactions on Smart Grid*, 2019.

Conferences Proceedings

1. **A. Moradipari**, B. Turan, Y. Abbasi-Yadkori, M. Alizadeh, and M. Ghavamzadeh "Feature and Parameter Selection in Stochastic Linear Bandits", *International Conference on Machine Learning (**ICML**)*, **(Spotlight)**, 2022.

2. **A. Moradipari**, S. Bae, M. Alizadeh, E. Moradi Pari, and D. Isele, "Predicting Parameters for Modeling Traffic Participants", *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2022.

3. **A. Moradipari**, M. Ghavamzadeh, T. Rajabzadeh, C. Thrampoulidis, and M. Alizadeh, "Multi-Environment Meta-Learning in Linear Stochastic Bandits", *IEEE International Symposium on Information Theory (ISIT)*, 2022.

4. **A. Moradipari**, M. Ghavamzadeh, and M. Alizadeh, "Collaborative Multi-agent Stochastic Linear Bandits", *American Control Conference (ACC)*, 2022.

5. **A. Moradipari**, S. Amani, M. Alizadeh, and C. Thrampoulidis "Safe Linear Bandits", *55th Annual Conference on Information Sciences and Systems (CISS)*, 2021.

6. **A. Moradipari**, M. Alizadeh, and C. Thrampoulidis "Stage-wise Conservative Linear Bandits", *Conference on Neural Information Processing Systems (**NeurIPS**)*, 2020.

7. **A. Moradipari**, M. Alizadeh, and C. Thrampoulidis "Linear Thompson sampling under unknown linear constraints", *IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

8. **A. Moradipari**, N. Tucker, T. Zhang, G. Cezar, and M. Alizadeh, "Mobility-Aware Smart Charging of Electric Bus Fleets", *IEEE Power & Energy Society General Meeting (PESGM)*, 2020.

9. **A. Moradipari** and M. Alizadeh, "Pricing Differentiated Services in an Electric Vehicle Public Charging Station Network", *IEEE 56th Conference on Decision and Control (CDC)*, Miami, FL, pp. 6488-6494, 2018.

10. **A. Moradipari** and M. Alizadeh, "Learning to Dynamically Price Electricity Demand Based on Multi-armed Bandits", *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Anaheim, CA, pp. 917-921, 2018.

11. **A. Moradipari**, S. Shahsavari, A. Esmaeili, and F. Marvasti, "Using Empirical Covariance Matrix in Enhancing Prediction Accuracy of Linear Models with Missing Information", *IEEE International Conference on Sampling Theory and Applications (SampTA)*, pp. 446-450, 2017.

Abstract

Learning and Pricing Algorithms for Human-Cyber-Physical Systems

by

Ahmadreza Moradipari

Nowadays with the growth of large-scale societal infrastructure systems, there has been significant research attention on improving efficiency, guaranteeing safety, reducing operational costs, and decreasing the carbon footprint of these systems. In particular, this thesis is focused on Human-Cyber-Physical Systems (H-CPS) (e.g. smart grid, electric transportation networks, autonomous driving). An H-CPS is any physical system in which a mechanism is controlled by both computer-based algorithms and human inputs. With the increasing complexity of human-machine interfaces, the traditional engineering and operating strategies are not adequate to manage. In fact, a mix of tools from stochastic control, distributed optimization, machine learning, and game theory is required. For example, in modern electric transportation systems, without appropriate demand management and coordination schemes, Electric Vehicle (EV) charging patterns could create problems for power transmission and distribution networks, and hence reduce the environmental benefits of transportation electrification.

Furthermore, when managing demand to reduce costs in a power system, it is necessary to ensure that the operating constraints of the power grid are not violated as a result of our actions. Additionally, because of the availability of real-time data from these infrastructure systems, training a large-scale model over a vast amount of data requires sophisticated techniques that accelerate the training of learning models. Therefore, it becomes important to develop algorithms that are computationally efficient and consider the critical safety requirements of these systems.

The aforementioned problems are characterized by many challenges including: How can we encourage customers to act in a way that is more likely to benefit society even when it may conflict with their own interests? How do we make sure that the infrastructure systems' safety criteria are not disregarded while we are learning the proper procedures to optimize customer behavior? How do we make sure that our proposed algorithms are computationally efficient?

This thesis is focused on developing optimization and machine learning frameworks that promote efficiency and flexibility in large-scale societal infrastructure systems with the active involvement of humans. In the first part of the thesis, we focus on designing optimal price and routing mechanisms for a public charging stations network in electric transportation systems to coordinate customers (i.e., EV drives) towards a socially optimal behavior given their heterogeneity. In the second part of the thesis, we provide two theoretical learning guarantees for online decision-making problems in safety-constrained unknown linear systems. Moving on to the third part, we develop two methods to speed up the learning process of the online learning algorithms in new tasks based on their limited past experience with unknown linear systems. We also support our theoretical results in all three parts by significant improvement in numerical experiments.

Contents

Chapter 1

Introduction

Large-scale societal infrastructure systems powered by artificial intelligence (AI) and human input are shaping the future of our society. Unlike traditional engineering approaches, modern and adaptive operating strategies are required to handle these complex systems. In fact, many large-scale societal systems with the active involvement of humans in the control loop require advanced methods to account for the stochastic behavior of humans while considering critical infrastructure constraints. Furthermore, with the availability of real-time data from these infrastructure systems, training a large-scale model over a massive data set is computationally intensive, and hence it requires modern strategies that speed up the training of learning models.

In this thesis, we focus on two frameworks: designing pricing and routing mechanisms for demand management in electric transportation systems as well as designing machine learning algorithms for systems with the active involvement of humans in the control loop, with a shared goal in mind: how do we adapt recent advances in distributed optimization, statistical learning theory, and control theory for efficient and safe data-driven decision-making in such safety-constrained complex systems?

1.1 Motivation

Electric Vehicle Management in Public Charging Station Network

According to the U.S. Department of Energy, more than 2.2 million electric vehicles (EVs) were on the road in 2021 and there are nearly 47,000 charging station locations with more than 120,000 ports across the nation [1]. It is important to note that both the EVs and stations are not evenly distributed throughout the country. There are certain regions where EV adoption rates are much higher than average yet the number of charging stations are lacking. Additionally, the recharging process of an EV is significantly slower than the refueling process for an internal combustion engine vehicle (ICEV), meaning that EVs occupy chargers for long periods of time. As a result of the limited infrastructure and long charging times, EV owners in populated urban areas are experiencing high levels of congestion at public charging stations during peak usage hours [2]. Therefore, without appropriate demand management schemes in place, Electric Vehicle (EV) charging patterns could create problems for power transmission and distribution networks, and reduce the environmental benefits of transportation electrification.

In this thesis, our objective is to guide EV drivers to drive into the right charging station in a mobility-aware fashion, in order to 1) manage the effect of EVs on the grid (e.g., on capacity-constrained feeders or integration of behind-the-meter solar) and 2) ensure fair service to customers with proper capacity allocation and short station wait times (admission control), considering heterogeneous user preferences and needs. In particular, we consider a Charging Network Operator (CNO) that owns a network of EV public charging stations and wishes to offer a menu of differentiated service options for access to its stations. This involves designing optimal pricing and routing schemes for the setting where users cannot directly choose which station they use. Instead, they choose their priority level and energy request amount from the differentiated service

menu, and then the CNO directly assigns them to a station on their path. This allows higher-priority users to experience lower wait times at stations and allows the CNO to directly manage demand, exerting a higher level of control that can be used to manage the effect of EVs on the grid and control station wait times. We consider the scenarios where the CNO is a social welfare-maximizing or a profit-maximizing entity, and in both cases, design incentive-compatible price-routing policies that ensure users reveal their true private information to the CNO.

Bandit Algorithms for Safety-Critical H-CPS

Stochastic bandit optimization algorithms have long found applications in many fields where some characteristics of the users' responses are not known and can only be learned through a limited number of noisy observations, including recommendation engines, advertisement placement, personalized medicine, etc. This shares a similar challenge to the one in H-CPS due to the involvement of humans in the control loop. For example, bandit optimization inherently maximizes a reward function (e.g., efficiency in societal systems) where some characteristics are not known and have to be learned while interacting with a user. This is similar to the electricity pricing in societal-scale infrastructure systems such as power grids or transportation networks which minimize the operational costs with a limited number of user interactions. However, the existing bandit algorithms might not be directly applicable to these cases due to the existence of infrastructural safety constraints that have to be met at each round of user interactions.

In this thesis, we formulate a linear stochastic bandit problem with safety constraints that depend linearly on an unknown parameter vector. As such, especially in the earlier rounds, there is a need to choose actions with caution, while at the same time making sure that the chosen action provides sufficient learning opportunities about the set of safe actions. For this bandit problem, we propose a Thompson-sampling algorithm,

which includes necessary modifications to respect the safety constraints with provable performance guarantees. Furthermore, we formulate related problem variations with stage-wise baseline constraints, in which the learner must choose actions that not only maximize cumulative reward across the entire time horizon but further satisfy a linear baseline constraint taking the form of a lower bound on the instantaneous reward.

Model Selection in the Bandit Problem

With the availability of real-time data from the societal infrastructure systems, training a large-scale model over a massive data set is extremely computational, and hence it requires modern strategies that speed up the training of learning models. One approach is to leverage the similarity of the new online learning tasks with the previous experiences and transfer proper information in order to adapt faster to the new situations. For example, in autonomous driving research, accurately modeling the behavior of traffic participants is essential for safely and efficiently navigating an autonomous vehicle through heavy traffic. Using the previously recorded traffic data set, we could design/train different models for different driver behaviors (e.g., conservative or aggressive drivers) [3]. Then, we could use those offline trained models, in online decision-making scenarios with a limited number of possible interactions to determine which model could be a good representation for each of the traffic participants. In practice, the number of trained models could be very large, which requires the online strategies to be efficient in the number of models.

In this thesis, we discuss model selection in stochastic linear bandits (LB), where the LB problem at hand is selected from a set of M models. The agent has information about the models but does not know the identity of the one(s) that the new LB problem has been selected. The goal of the agent is to identify the true model(s) and transfer its (their) collected experience to speed up the learning of the task at hand. It is a common

scenario in many application domains that the new task belongs to a family of models that are either known accurately or with misspecification. For example, it is reasonable to assume that the customers of an online marketing website, the users of an app, or the patients in a medical trial belong to a certain number of categories based on their shopping and browsing habits or their genetic signatures. It is also common these days that websites, apps, and clinics have a large amount of information from each of these categories that can be used to build a model. For another example, we could consider different driving behaviors could belong to a certain number of categories in autonomous driving research, and there exists an excessive amount of recorded traffic data that can be used to build these categories.

1.2 Chapter Overviews

1.2.1 Chapter 1

Chapter 1 presents the motivation for this thesis and summary of main contributions.

1.2.2 Chapter 2

Chapter 2 presents our results on electric vehicle (EV) demand management in public charging stations network.

In Section 2.2, we present a decision problem of a Charging Network Operator (CNO) for managing EVs in a public charging station network through differentiated services. In this case, EV users cannot directly choose which charging station they will charge at. Instead, they choose their energy demand and their priority level, as well as their traveling preferences (which stations they are willing to visit) from among a menu of service options that are offered to them, and the CNO then assigns them to the charging

stations directly to control station wait times and electricity costs. This is reminiscent of incentive-based direct load control algorithms that are very popular in demand response.

Sections 2.3 and 2.4 present incentive-compatible pricing and routing policies for maximizing the social welfare or the profit of the CNO. We first formulate the CNO's goal for choosing a routing policy that maximizes the social welfare in Section 2.3 as well as a profit of the system in Section 2.4 of all EV users with access to the network, which is generally a non-convex problem. We proposed an incentive-compatible pricing policy that enforces the socially optimal routing policy as an equilibrium. Then, we propose an algorithm that finds the globally optimal solution of the non-convex routing policy problem in both social welfare and profit maximization scenarios in the special case of hard capacity constraints. We also highlighted the benefits of our algorithms towards behind-the-meter solar integration at the station level.

1.2.3 Chapter 3

Chapter 3 presents a brief summary of the Stochastic Linear Bandit (LB) problem.

Section 3.2 presents the LB problem setting. Then we present the two well-known algorithms OFUL and Thompson Sampling for the LB problem with their regret bounds in Sections 3.4 and 3.5, respectively.

1.2.4 Chapter 4

Chapter 4 presents our theoretical results on guaranteeing safety in H-CPS in certain instances. Section 4.2 presents the setting for a linear stochastic bandit (LB) problem in which the environment is subject to unknown linear safety constraints that need to be satisfied at each round with a follow-up motivational example.

In Section 4.3, we present the first safe Linear Thompson Sampling (Safe-LTS) algo-

rithm with provable regret guarantees for the linear bandit problem with linear safety constraints. We also present the challenges risen by the presence of the safety constraints in the traditional linear bandit problem and the mechanisms that Safe-LTS employs to address these challenges. Moreover, in Section 4.4, we show that the Safe-LTS achieves the same frequentist regret of order $\mathcal{O}(d^{3/2}\log^{1/2} d \cdot T^{1/2}\log^{3/2} T)$ as the original LTS without safety constraint. Hence, the dependence of our regret bound on the time horizon T cannot be improved by modulo logarithmic factors. We also compare Safe-LTS with several UCB-type safe algorithms in Section 4.5. We show that our algorithm has: better regret in the worst-case ($\widetilde{\mathcal{O}}(T^{1/2})$ vs. $\widetilde{\mathcal{O}}(T^{2/3})$), fewer parameters to tune and often superior empirical performance.

1.2.5 Chapter 5

Chapter 5 presents another notation of safety among other applications. In particular, we study stage-wise conservative linear stochastic bandits: an instance of bandit optimization, which accounts for (unknown) "safety constraints" that appear in applications such as online advertising and medical trials. At each stage, the learner must choose actions that not only maximize cumulative reward across the entire time horizon but further satisfy a linear baseline constraint that takes the form of a lower bound on the instantaneous reward. For this problem, in Section 5.3 and C.7, we present two novel algorithms, *stage-wise conservative linear Thompson Sampling* (SCLTS) and *stage-wise conservative linear UCB* (SCLUCB), that respect the baseline constraints and enjoy probabilistic regret bounds of order $\mathcal{O}(\sqrt{T}\log^{3/2} T)$ and $\mathcal{O}(\sqrt{T}\log T)$, respectively.

Notably, the proposed algorithms can be adjusted with only minor modifications to tackle different problem variations, such as constraints with bandit feedback, or an unknown sequence of baseline rewards. We discuss these and other improvements over

the state-of-the-art. For instance, compared to existing solutions, in Section 5.4, we show that SCLTS plays the (non-optimal) baseline action at most $\mathcal{O}(\log T)$ times (compared to $\mathcal{O}(\sqrt{T})$). Finally, we make connections to another studied form of "safety constraints" that takes the form of an *upper bound* on the instantaneous reward. While this incurs additional complexity to the learning process as the optimal action is not guaranteed to belong to the "safe set" at each round, we show that SCLUCB can properly adjust in this setting via a simple modification in Appendix 5-Section C.8.

1.2.6 Chapter 6

Chapter 6 presents our results on model selection settings in stochastic linear bandits.

In Section 6.2, we study two model selection settings in stochastic linear bandits (LB). In the first setting, which we refer to as *feature selection*, the expected reward of the LB problem is in the linear span of at least one of M feature maps (models). In the second setting, the reward parameter of the LB problem is arbitrarily selected from M models represented as (possibly) overlapping balls in $\mathbb{R}^d$. However, the agent only has access to misspecified models, i.e., estimates of the centers and radii of the balls. We refer to this setting as *parameter selection*.

For each setting in Sections 6.3 and 6.4, we develop and analyze a computationally efficient algorithm that is based on a reduction from bandits to full-information problems. This allows us to obtain regret bounds that are not worse (up to a $\sqrt{\log M}$ factor) than the case where the true model is known. This is the best-reported dependence on the number of models M in these settings. Finally, in Section 6.5, we empirically show the effectiveness of our algorithms using synthetic and real-world experiments.

Chapter 2

Pricing and Routing Mechanisms for Electric Vehicle Public Charging Stations Network

2.1 Introduction

It is well-known that without appropriate demand management schemes in place, Electric Vehicle (EV) charging patterns could create problems for power transmission and distribution networks, and reduce the environmental benefits of transportation electrification. Hence, the past decade has seen significant research advances in the design of EV demand management algorithms. Broadly speaking, most available smart charging approaches focus on optimizing residential and commercial charging profiles when the duration of charge events allows for temporal load shifting. However, our focus in this chapter is on public charging station networks, which are fundamentally different from residential and commercial charging in two ways: 1) Temporal load shifting after a plug-in event is not feasible, unless battery swapping methods are employed. Most drivers

would want to leave the station as soon as possible, quite similar to a gas station stop; 2) Access to EV supply equipment (EVSE) is open to the public, which creates congestion effects and results in wait times at popular stations.

Prior art: We categorize the rich literature on mobility-aware charge management of EVs into three categories. The first category considers using the mobility pattern of EVs in order to optimize EV charging load in an economic dispatch problem and manage EVs' effects on transmission systems (see, e.g., [4, 5, 6, 7, 8]) or distribution systems (see, e.g., [9, 10]). In [11], the authors study the dynamic impact of EV movements on integrated power and traffic systems. They propose Nodal Time-of-Use (NTOU) and Road Traffic Congestion (RTC) prices to control the driving pattern of EV loads. In [12], the authors study the extended Pickup Delivery Problems (PDPs) for an EV fleet containing EV customers with different service requests. They propose a mixed-integer quadratic constraint optimization for solving the offline pre-trip scheduling problem. This line of work is not focused on public charging stations and mostly adopts traffic assignment models. The second category of related work focuses on the problem of routing EV users to stations (see, e.g., [13, 14, 15, 16, 17, 18]). Naturally, given the stochastic nature of EV arrivals and the limited number of EVSEs at each station, one can consider the problem of managing access to public charging stations as a queuing network, where previous works have considered various objectives such as revenue maximization or waiting time minimization (see, e.g. [19, 20, 21] and the references therein). The main focus of these papers is the design of optimal *routing* policies to directly send users to stations given heterogeneous user needs and not on designing pricing strategies. The advantage of using our proposed mechanism compared to these papers is that we jointly design incentive-compatible pricing and routing policies. This means that our work does not assume that customers will have to follow our routing orders without considering customers' incentives to deviate from the posted assignment. The downside is that our algorithm is

more complex than one that is solely focused on optimal routing without any incentive issues. The third category of work, which is most intimately connected to our work, considers the design of pricing strategies to manage users' access to charging networks, where individuals decide which station to use based on prices (self-routing) [22]. In [23], the authors study the waiting times of charging station queues and the profit of the CNO under flat rate charging prices as well as a threshold-based pricing policy that penalizes higher demand. In [24], the authors propose a Stackelberg framework to design prices for charging stations that incentives more uniform station utilization. In [25], the authors study the joint charging and navigation problem of EVs. They formulate en-route charging navigation problems using Dynamic Programming (DP). They propose a so-called Simplified Charge Control (SCC) algorithm for deterministic traffic networks. Moreover, for the stochastic case, they propose an online state recursion algorithm.

Our objective is to guide EV drivers to drive *into the right station* in a mobility-aware fashion, in order to 1) manage the effect of EVs on the grid (e.g., on capacity-constrained feeders or integration of behind-the-meter solar) and 2) ensure fair service to customers with proper capacity allocation and short station wait times (admission control), considering heterogeneous user preferences and needs. This is not an easy task to achieve merely through pricing algorithms, mainly due to the complexity of the price response structure of users and its dependence on the users' mobility needs and preferences, information that is not readily available and is very hard to obtain. Hence, we take a different path here, which allows us to somewhat separate the pricing and admission control aspects of the problem. We assume that customers cannot directly choose which charging station they will charge at. Instead, a Charging Network Operator (CNO) is in charge of directly assigning users to charge stations given their respective *value of time* (VoT), *charging demand* and *travels preferences*. We believe that this is reasonable given that, even today, access to public charging stations is only allowed

for specific vehicle types or with users with prepaid charging plans/subscriptions. A customer's travel preferences specify which charging stations they are willing to visit. The CNO's goal is to design a menu of *differentiated service options* with service qualities that are tailored to the characteristics of heterogeneous users. Each service option is tailored to users with given VoT, charging demand, and travel preferences, and is associated with a routing policy (i.e., the probability of that customer type being assigned to each of the stations on their path), as well as an appropriate price. The CNO wishes to optimize these differentiated routing policies and prices in order to optimally use capacity-limited charging stations and minimize electricity costs. Furthermore, the CNO's goal is to design *incentive-compatible* pricing-routing policies, which ensures that individual users reveal their true needs and preferences to the CNO. Such differentiated pricing mechanisms have been studied before in the context of residential demand response in recent years (see, e.g., [26, 27]) in order to incentivize the participation of loads in direct load control programs, analogous to what we are trying to achieve here for fast charging networks.

The contributions of the work presented in this chapter are as follows:

- Modeling the decision problem faced by a CNO for managing EVs in a public charging station network through differentiated services.
- Proposing incentive-compatible pricing and routing policies for maximizing the social welfare or the profit of the CNO considering users' mobility patterns, distribution network constraints, or behind-the-meter solar generation.
- Proposing an algorithm that finds the globally optimal solution for the CNO's non-convex objective in the special case of hard capacity.

2.2 System Model

2.2.1 Individual User Model

We first describe the individual EV users' parameters and decision-making model.

User types

We assume that users belong to one of $V \times E \times B$ types. A type (i, j, ℓ) customer has a value of time (VoT) v_i with $i \in \mathcal{V} = \{1, \ldots, V\}$, an energy demand e_j with $j \in \mathcal{E} = \{1, \ldots, E\}$, and a traveling preference $\mathcal{G}_\ell$, with $\ell \in \mathcal{B} = \{1, \ldots, B\}$. The value of time is often used to model the heterogeneity of users' utility and choice when optimizing their response in the presence of travel time variations. The set of traveling preferences $\mathcal{G}_\ell$ declares the set of stations to which customers with preference ℓ have access on their path. More specifically, for each traveling preferences ℓ, we define the vector $\mathbf{y}_\ell$ with length Q (number of charging stations) such that $\mathbf{y}_\ell(q) = 1$ if station $q \in \mathcal{G}_\ell$ and 0 otherwise. For convenience, we order the customer types such that both VoT and energy demand are in ascending order, i.e., $v_1 < v_2 < \ldots < v_V$, and $e_1 < e_2 < \ldots < e_E$. In this chapter, we assume that users do not act strategically in choosing the amount of energy they need, i.e., they fully charge their EV if they enter a charging station.

We assume that type (i, j, ℓ) customers arrive in the system with a given 3-dimensional expected (average) arrival rate matrix $\mathbf{\Lambda} = [\Lambda_{i,j,\ell}]_{i \in \mathcal{V}, j \in \mathcal{E}, \ell \in \mathcal{B}}$, which we consider as an inelastic and known parameter. In each *potential arrival*, the customers can choose to either purchase a service option from the differentiated service options offered by CNO, or choose to not buy any charging services. Note that we are in a static setting, i.e., the expected rate of arrival of users of different types is assumed as a constant variable when designing pricing/routing policies. While the arrival rate can vary across time, we will assume that the dynamics of the charging process at fast charging stations are faster

than the dynamics of average traffic conditions.

Service options

We assume that the number of differentiated service options that are available matches the three-dimensional user types $(i, j, \ell) \in \mathcal{V} \times \mathcal{E} \times \mathcal{B}$. The CNO will sell each service option (i, j, ℓ) with price $P_{i,j,\ell}$. Moreover, service options are differentiated in terms of a routing policy $\mathbf{r}_{i,j,\ell} = [r^q_{i,j,\ell}]_{q=1,\dots,Q}$, which is a column vector of routing probabilities of customers that purchase service option (i, j, ℓ) to each charging station $q \in \mathcal{G}_\ell$.

The joint choice of these *pricing-routing policies* $(P_{i,j,\ell}, \mathbf{r}_{i,j,\ell})$ would affect the proportion of users that choose to purchase each service option, which would, in turn, affect the arrival rate and average charging demand per EV at each charging station. As a result, the average total electricity demand and waiting times at the station are determined through the design of these pricing-routing policies. Hence, the design of the pricing-routing policy to be employed directly affects the social welfare (or the CNO's profit). To concretely model this connection, we first model how users choose which service type to purchase (if any).

User decision model

In general, users have no obligation to buy the services option corresponding to their own true type (why would I tell a CNO that I have a low value of time and be assigned a longer wait?). The total utility of a user from purchasing charging services is the reward they receive from charging minus the expected waiting cost (which is the product of VoT with the expected waiting time) and the price paid for charging services. Let us assume that customers with the value of time v_i and traveling preference ℓ will get a reward R^ℓ_i for receiving a full battery charge. Furthermore, we assume that information about the expected wait time $W_{i,j,\ell}$ of each service option (i, j, ℓ) in the menu is available to users.

Throughout this chapter, we assume that the time it takes to drive to a station from the main corridor (denoted by d_q) is included in the "wait time" corresponding to that station (on top of the queuing time ϱ_q), i.e., we have

$$W_{i,j,\ell} = \sum_{q=1}^{Q} \Big(d_q + \varrho_q \Big) r_{i,j,\ell}^{(q)}. \tag{2.2.1}$$

We will assume that the users do not observe the current exact realization of wait times, i.e., the expected wait time $W_{i,j,\ell}$ is not conditioned on the realization of the random arrival rate of the user and will be constant at the equilibrium. Therefore, customers of type (i, j, ℓ) will choose their service option (m, k, t) by solving:

$$\max_{m \in \mathcal{V}, j \le k \le E, t \in \mathcal{B}_\ell} R_i^\ell - v_i W_{m,k,t} - P_{m,k,t}. \tag{2.2.2}$$

According to our assumption on the inelasticity of user's charging needs, a customer of type (i, j, ℓ) can only choose a service option (m, k, ℓ) if $e_j \le e_k$. Moreover, we assume users of type (i, j, ℓ) may only choose a travel preference $t \in \mathcal{B}_\ell$, where $\mathcal{B}_\ell$ is defined as the set of all preferences $t \in \mathcal{B}$ such that $\mathcal{G}_t \subset \mathcal{G}_\ell$ (otherwise the user would have to change their travel origin-destination pair). If the total utility defined in (2.2.2) is not positive for any available service option (m, k, t), then that customer will not purchase charging services. We would like to note that our scheme is not forcing any user to accept the CNO's routing to different stations. It only provides lower prices for more flexibility in regard to waiting time and station choice. If a user is not willing to provide this flexibility, they may choose to select the service option that only includes the specific station they would like to visit and naturally pay a higher price for receiving service.

The aggregate effect of each individual customer's decision of whether to buy service or not and their choice of service option will lead to a Nash Equilibrium (NE) of *effective*

expected arrival rates in the charging station network, denoted by $\boldsymbol{\lambda} = [\lambda_{i,j,\ell}]_{i\in\mathcal{V},j\in\mathcal{E},\ell\in\mathcal{B}}$. Our goal in this chapter is to design a pricing routing policy such that 1) the resulting NE is optimal for maximizing social welfare or CNO profit; 2) we belong to the family of incentive-compatible (IC) pricing policies, i.e., policies where every user can achieve the best outcome for themselves by acting according to their true preferences. Next, we characterize conditions that should hold at equilibrium for such policies.

2.2.2 Incentive Compatible (IC) Pricing-Routing Policies

In this chapter, we would like to focus on Incentive Compatible (IC) pricing-routing policies. A pricing-routing policy is IC if, for each user type (i, j, ℓ), it is always optimal to choose the service option that matches their user type, i.e., service option (i, j, ℓ). Hence, no users will have any incentive to lie about their user type to the CNO, which can be desirable for system design purposes. Mathematically, given the user's decision problem in (2.2.2), this condition will be satisfied for a pricing routing policy if the following conditions are satisfied at equilibrium:

$$\forall k, t \in \mathcal{V}, t \neq k, \forall j \in \mathcal{E}, \forall \ell \in \mathcal{B}$$

$$P_{k,j,\ell} + v_k W_{k,j,\ell} \leq P_{t,j,\ell} + v_k W_{t,j,\ell}, \tag{2.2.3}$$

$$P_{t,j,\ell} + v_t W_{t,j,\ell} \leq P_{k,j,\ell} + v_t W_{k,j,\ell}, \tag{2.2.4}$$

$$\forall i \in \mathcal{V}, \forall t, k \in \mathcal{E}, t > k, \forall \ell \in \mathcal{B}$$

$$P_{i,k,\ell} + v_i W_{i,k,\ell} \leq P_{i,t,\ell} + v_i W_{i,t,\ell}, \tag{2.2.5}$$

$$\forall i \in \mathcal{V}, \forall j \in \mathcal{E}, \forall \ell \in \mathcal{B}, \forall t \in \mathcal{B}_\ell$$

$$P_{i,j,\ell} + v_i W_{i,j,\ell} \leq P_{i,j,t} + v_i W_{i,j,t}, \tag{2.2.6}$$

These conditions ensure that no user receives a higher utility by joining the system under any type other than their own. For convenience, we refer to (2.2.3)-(2.2.4) as *vertical IC* constraints, and (2.2.5) as the *horizontal IC* constraint. Note that while the service options' prices $P_{i,j,\ell}$ play a direct role in these conditions, the routing probabilities $\mathbf{r}_{i,j,\ell}$ only indirectly affect these conditions by determining the wait times $W_{i,j,\ell}$. We will explore this connection more later.

Furthermore, Individual Rationality (IR) is satisfied if the following constraints are satisfied at equilibrium:

$$\begin{aligned} P_{i,j,\ell} &= R_i^\ell - v_i W_{i,j,\ell}, \text{ if } 0 < \lambda_{i,j,\ell} < \Lambda_{i,j,\ell} \\ P_{i,j,\ell} &< R_i^\ell - v_i W_{i,j,\ell}, \text{ if } \lambda_{i,j,\ell} = \Lambda_{i,j,\ell} \\ P_{i,j,\ell} &> R_i^\ell - v_i W_{i,j,\ell}, \text{ if } \lambda_{i,j,\ell} = 0. \end{aligned} \tag{2.2.7}$$

That is if any user of type (i, j, ℓ) joins the system, their utility from joining the system must be non-negative. Next, we study the structure of NE under any IC policies under two assumptions about rewards R_i^ℓ.

Assumption 1 *For customers with different traveling preferences, the rewards R_i^ℓ satisfy the following:*

$$\begin{aligned} &\forall i \in \mathcal{V}, \forall \ell, m \in \mathcal{B}: \\ &\textit{if } |\mathcal{G}_\ell| > |\mathcal{G}_m| \textit{ then } R_i^\ell < R_i^m, \\ &\textit{if } |\mathcal{G}_\ell| = |\mathcal{G}_m| \textit{ then } R_i^\ell = R_i^m. \end{aligned} \tag{2.2.8}$$

This means that users with a more limited set of charging options get a higher reward for receiving service.

Assumption 2 *For customers with the same traveling preference ℓ, the ratios $\frac{R_i^\ell}{v_i}$ satisfy the following:*

$$\frac{R_1^\ell}{v_1} < \frac{R_2^\ell}{v_2} < \ldots < \frac{R_V^\ell}{v_V}. \tag{2.2.9}$$

A similar structure was assumed in [28] and other past work for service differentiation through pricing-routing policies in a single server service facility with Poisson arrivals and exponential service time M/M/1.

The next lemma shows that under an IC pricing-routing policy, waiting time is a non-increasing function of VoT for users with the same traveling preference and energy demand.

Lemma 2.2.1 *Under an incentive-compatible pricing-routing policy, for any users of types $(i+1, j, \ell)$ and (i, j, ℓ) who have purchased charging services, we must have:*

$$W_{i+1,j,\ell} \leq W_{i,j,\ell}. \tag{2.2.10}$$

Proof: From vertical IC constraints (2.2.3) and (2.2.4) for customers of type (i, j, ℓ) and $(i+1, j, \ell)$, we can write:

$$(v_{i+1} - v_i)W_{i+1,j,\ell} \leq (v_{i+1} - v_i)W_{i,j,\ell},$$

and the fact that $v_{i+1} - v_i > 0$, would lead to the result. ■

The next lemma shows that it suffices to only check IC conditions for neighboring service options, e.g., the options with one level higher value in VoT or energy.

Lemma 2.2.2 *(Local IC) The IC constraints (2.2.3)-(2.2.6) are satisfied if and only if:*

$$\begin{aligned}
&\forall i \in \{1,\dots,V-1\}, \forall j \in \mathcal{E}, \forall \ell \in \mathcal{B}:\\
&\qquad P_{i+1,j,\ell} + v_{i+1} W_{i+1,j,\ell} \le P_{i,j,\ell} + v_{i+1} W_{i,j,\ell},\\
&\qquad P_{i,j,\ell} + v_{i} W_{i,j,\ell} \le P_{i+1,j,\ell} + v_{i} W_{i+1,j,\ell},\\
&\forall i \in \mathcal{V}, \forall j \in \{1,\dots,E-1\}, \forall \ell \in \mathcal{B}:\\
&\qquad P_{i,j,\ell} + v_{i} W_{i,j,\ell} \le P_{i,j+1,\ell} + v_{i} W_{i,j+1,\ell},\\
&\forall i \in \mathcal{V}, \forall j \in \mathcal{E}, \forall \ell \in \mathcal{B}, \forall t \in \mathcal{T}_\ell:\\
&\qquad P_{i,j,k} + v_{i} W_{i,j,k} \le P_{i,j,t} + v_{i} W_{i,j,t}, \qquad (2.2.11)
\end{aligned}$$

where $\mathcal{T}_\ell$ denotes the set of all travel preferences $t \in \mathcal{B}_\ell$ such that $|\mathcal{G}_t| = |\mathcal{G}_\ell| - 1$.

In the following lemma, we highlight a special structure of users' arrival pattern $\boldsymbol{\lambda}$ at equilibrium under an IC policy.

Lemma 2.2.3 *If customers of type (i,j,ℓ) have partially entered the system (i.e., $0 < \lambda_{i,j,\ell} < \Lambda_{i,j,\ell}$), under an IC policy, the effective arrival rates satisfy:*

1. *(Vertical solution structure) $\lambda_{k,j,\ell} = \Lambda_{k,j,\ell}, \forall k > i$, and $\lambda_{k,j,\ell} = 0, \forall k < i$, i.e., customers with higher VoTs and similar energy demand and similar traveling preference enter the system in full, and customers with lower VoTs do not enter the system.*

2. *(Horizontal solution structure) $\lambda_{i,k,\ell} = \Lambda_{i,k,\ell}, \forall k < j$, and $\lambda_{i,k,\ell} = 0, \forall k > j$, i.e., customers with lower energy demand and same VoT and same traveling preference enter in full, and customers with higher energy demand and same VoT and same traveling preference do not enter at all.*

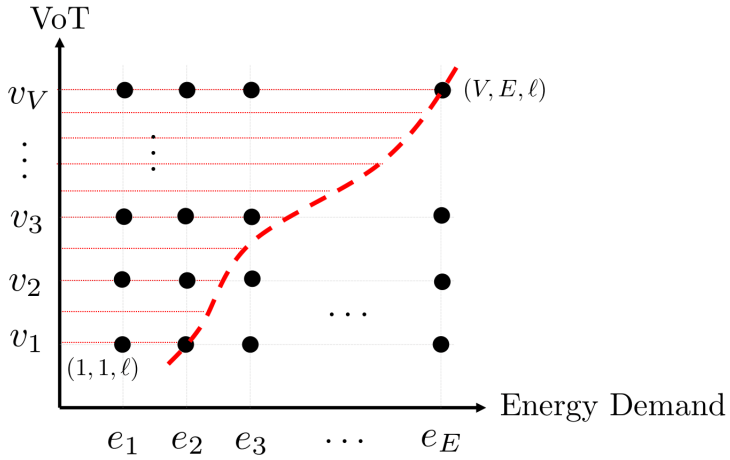

Figure 2.1: The solution structure for an IC policy.

The proof follows from combining, IR and IC conditions, as well as Assumption 1. We omit it due to brevity.

Therefore, at the Nash equilibrium, due to IC constraints, the solution structure of the effective arrival rates are similar to Fig. 2.1. The red borderline shows which user types should partially enter the system, i.e., where $0 < \lambda_{i,j,\ell} < \Lambda_{i,j,\ell}$. This means that not all users of type (i, j, ℓ) will join the system. Hence, from Lemma 2.2.3, we know that customers to the left of the line will enter the system in full, and customers to the right will not enter the system. Next, we study the design of a socially-optimal IC pricing-routing policy.

2.3 Socially-Optimal Policy

Our charging stations are located at heterogeneous distances from the users' path and have different locational marginal prices and capacities. In the socially optimal policy, the CNO's goal is to choose a routing policy that maximizes the social welfare of all EV

users with access to the network, which we can write as:

$$\max_{\substack{\mathbf{r}_{i,j,\ell}\geq 0\\ 0\leq\lambda_{i,j,\ell}\leq\Lambda_{i,j,\ell}}} \sum_{l=1}^{B}\sum_{i=1}^{V}\sum_{j=1}^{E}\left[R_i^{\ell}\lambda_{i,j,\ell} - v_i\lambda_{i,j,\ell}W_{i,j,\ell}(\boldsymbol{\lambda},\boldsymbol{R}) - \boldsymbol{\theta}^T\mathbf{r}_{i,j,\ell}e_j\lambda_{i,j,\ell}\right] \tag{2.3.1}$$

$$\text{s.t.} \quad \mathbf{1}^T\text{diag}(\mathbf{y}_\ell)\mathbf{r}_{i,j,\ell} = 1, \quad \forall i\in\mathcal{V}, j\in\mathcal{E}, \ell\in\mathcal{B}, \tag{2.3.2}$$

$$\sum_{l=1}^{B}\sum_{i=1}^{V}\sum_{j=1}^{E}\lambda_{i,j,\ell}e_j r_{i,j,\ell}^{(q)} \leq C_q, \ \forall q\in\{1,\ldots,Q\}, \tag{2.3.3}$$

where $\boldsymbol{\theta} = [\theta_q]_{q=1,\ldots,Q}$ denotes the vector of locational marginal prices of electricity at each charging station q, $\mathbf{r}_{i,j,\ell} = [r_{i,j,\ell}^q]_{q=1,\ldots,Q}$ is a column vector of routing probabilities for service option (i,j,ℓ) to each charging station q, $\mathbf{R} = [\mathbf{r}_{i,j,\ell}]_{\forall i,j,\ell}$ is the matrix of routing probabilities for all service types, with the $\big[(\ell-1)\times E\times v + E(i-1)+j\big]$-th column dedicated to type (i,j,ℓ), C_q is the capacity of charging station q, and $\boldsymbol{\lambda} = [\lambda_{i,j,\ell}]_{\forall i,j,\ell}$ is the vector of effective arrival rates. The objective function is the sum of the reward received by admitted users to the system minus waiting and electricity costs, (2.3.2) ensures that the routing probabilities sum up to one overall charging station allowed for traveling preference ℓ, and (2.3.3) is the capacity constraint for each charging station. The waiting time function $W_{i,j,\ell}(\boldsymbol{\lambda},\boldsymbol{R})$ maps the effective expected arrival rate in each station into an expected waiting time (e.g., queueing models can be appropriate here).

Can the CNO design an IC pricing policy that enforces the socially optimal routing solution (2.3.1) as an equilibrium? Next, we propose such a price. The first order necessary condition for the problem (2.3.1) is as follows:

$$\begin{aligned} R_i^{\ell} - v_i W_{i,j,\ell}(\boldsymbol{\lambda},\boldsymbol{R}) - \sum_{t,h,z}\left(\lambda_{t,h,z}v_t\frac{\partial W_{t,h,z}(\boldsymbol{\lambda},\boldsymbol{R})}{\partial\lambda_{i,j,\ell}}\right) \\ - \boldsymbol{\theta}^T\mathbf{r}_{i,j,\ell}e_j - \mathbf{x}^T\mathbf{r}_{i,j,\ell}e_j + \gamma_{i,j,\ell} - \mu_{i,j,\ell} = 0, \end{aligned} \tag{2.3.4}$$

with $\gamma_{i,j,\ell}\geq 0, \mu_{i,j,\ell}\geq 0$, $\gamma_{i,j,\ell}\lambda_{i,j,\ell} = 0$, $\mu_{i,j,\ell}(\lambda_{i,j,\ell}-\Lambda_{i,j,\ell}) = 0$, and $\mathbf{x} = [x_q]_{q=1,\ldots,Q}$ as the

Lagrange multiplier of the capacity constraint (2.3.3). We can observe that the following prices will satisfy the IR constraints (2.2.7):

$$P_{i,j,\ell} = \sum_{t=1}^{V}\sum_{h=1}^{E}\sum_{z=1}^{B}\left(\frac{\partial W_{t,h,z}(\boldsymbol{\lambda},\boldsymbol{R})}{\partial \lambda_{i,j,\ell}}\lambda_{t,h,z}v_t\right) + (\boldsymbol{\theta}+\mathbf{x})^T\mathbf{r}_{i,j,\ell}e_j. \tag{2.3.5}$$

Next, we show that the prices in (2.3.5) also satisfy IC constraints (2.2.3)-(2.2.6).

Proposition 2.3.1 *With the prices defined in (2.3.5), the solution of socially optimal problem (2.3.1) defines an incentive-compatible routing and pricing policy.*

Proof: The proof is inspired by that of Theorem 1 in [29]. To prove incentive compatibility, we need to choose two arbitrary service options and show that with the prices given by (2.3.5), customers from the first type are better off choosing their own option over the other. We first consider vertical IC constraints (2.2.3)-(2.2.4). Suppose, we have the globally optimal solution of (2.3.1). Assume customers of class (i,j,ℓ) enter the system and pretend to be of type (m,j,ℓ). We will increase the effective arrival rate of customers of type (i,j,ℓ) by an infinitesimal amount δ and treat them as customers of type (m,j,ℓ). Hence, because we were at the globally optimal solution of (2.3.1), we can write:

$$\frac{\partial}{\partial\delta}\Bigg[R_i^\ell\delta - \sum_{(t,h,z)\neq(i,j,\ell)} v_t\lambda_{t,h,z}W_{t,h,z}(\boldsymbol{\lambda}+\boldsymbol{\delta}_{m,j,\ell},\boldsymbol{R}) - v_i\lambda_{i,j,\ell}W_{i,j,\ell}(\boldsymbol{\lambda}+\boldsymbol{\delta}_{m,j,\ell},\boldsymbol{R})$$
$$- \delta v_i W_{m,j,\ell}(\boldsymbol{\lambda}+\boldsymbol{\delta}_{m,j,\ell},\boldsymbol{R}) - \delta\boldsymbol{\theta}^T\mathbf{r}_{m,j,\ell}e_j - \delta\mathbf{x}^T\mathbf{r}_{m,j,\ell}e_j\Bigg]_{\delta=0} \leq 0.$$

Hence, we can write:

$$R_i^\ell - \sum_{t,h,z}\left(\lambda_{t,h,z}v_t\frac{\partial W_{t,h,z}(\boldsymbol{\lambda},\boldsymbol{R})}{\partial\lambda_{m,j,\ell}}\right) - v_i W_{m,j,\ell}(\boldsymbol{\lambda},\boldsymbol{R}) - \boldsymbol{\theta}^T\mathbf{r}_{m,j,\ell}e_j - \delta\mathbf{x}^T\mathbf{r}_{m,j,\ell}e_j \leq 0.$$

Using the price in (2.3.5), this leads to:

$$R_i^\ell \leq v_i W_{m,j,\ell}(\boldsymbol{\lambda}, \boldsymbol{R}) + P_{m,j,\ell}$$

and from IR constraints (2.2.7), we know that if $\lambda_{i,j,\ell} > 0$, we need to have $R_i^\ell \geq v_i W_{i,j,\ell}(\boldsymbol{\lambda}, \boldsymbol{R}) + P_{i,j,\ell}$. Therefore,

$$v_i W_{i,j,\ell}(\boldsymbol{\lambda}, \boldsymbol{R}) + P_{i,j,\ell} \leq v_i W_{m,j,\ell}(\boldsymbol{\lambda}, \boldsymbol{R}) + P_{m,j,\ell},$$

which proves that vertical IC constraints hold. The proof for (2.2.5)-(2.2.6) is similar and we remove it due to brevity. ■

Our results up to this point are in their most general form. The expected waiting time $W_{i,j,\ell}(\boldsymbol{\lambda}, \boldsymbol{R})$ associated with each type (i, j, ℓ) can be defined using queueing theory as a weighted sum of wait times for the different charging stations or can have any other general form that arises in reality. However, we would like to note that the problem (2.3.1) is not convex in general, and hence finding the solution is not straightforward in all cases. While this is not devastating as this problem only has to be solved for planning, we will study the problem in the special case of hard capacity constraints next. This allows us to exploit the special structure highlighted in Lemma 2.2.3 to characterize the optimal routing policy through solving linear programs. This is especially useful for our numerical experiments.

2.3.1 Additional modeling factors: distribution network constraints and behind-the-meter solar

We would like to note that as opposed to residential and workplace charging, where temporal load shifting is possible for grid support, fast charging stations do not pro-

vide such opportunities (unless battery swapping methods are employed). Our proposed method allows the CNO to consider the following elements when optimizing pricing-routing decisions for charging stations: 1) the locational electricity prices for each charging station (already included in (2.3.1)); 2) behind the meter RES supply availability (such as solar generation) at each station; 3) distribution network information and constraints. We will elaborate on the latter two additions in this section.

In order to additionally consider network constraints such as line loading limits (defined below as the total line capacities excluding the loadings induced by conventional demands) the CNO can consider adding the following constraint to the CNO's optimization problem (2.3.1):

$$\sum_{q=1}^{Q} DE_q \left(\sum_{\ell=1}^{B} \sum_{i=1}^{V} \sum_{j=1}^{E} \lambda_{i,j,\ell} e_j r_{i,j,\ell}^{(q)} \right) \leq f_t, \ \forall t. \tag{2.3.6}$$

The constraint is similar to those adopted in [30, 31] for temporal load shifting of EV load in distribution networks. The reader should note that if this constraint is added to (2.3.1), the Lagrange multiplier of this constraint should be added to the prices we defined in (2.3.5).

Second, we would like to note that behind-the-meter solar energy available at stations can be easily accommodated by our model by adding in virtual stations with an electricity price of 0, traveling time equal to the station which is equipped by solar generation, and capacity equal to the currently available solar generation. In this case, the CNO is able to observe the available behind-the-meter solar integration in real-time, and design pricing-routing schemes in order to efficiently use real-time solar generation. This addition will help us better highlight the differences between the routing solutions of the social-welfare maximizing and profit-maximizing policies that we will discuss in our numerical results in Section 2.5.

2.3.2 The Special Case of Hard Capacity Constraints

In this special case, we assume that station queuing time (i.e., $\varrho_q = 0, \forall q = 1, \ldots, Q$) will be equal to zero as long as the station is operated below capacity. Furthermore, we assume that the travel time from the main corridor to reach each charging station k is a known and constant parameter $d_q, q = 1, \ldots, Q$. Therefore, the expected wait time for customers of type (i, j, ℓ) is:

$$W_{i,j,\ell} = \sum_{q=1}^{Q} d_q r_{i,j,\ell}^{(q)}. \tag{2.3.7}$$

Without loss of generality, we assume that stations are ordered such that $d_1 < d_2 < \ldots < d_Q$. We can now rewrite the socially-optimal problem (2.3.2) as:

$$\max_{\substack{\mathbf{r}_{i,j,\ell} \geq 0 \\ 0 \leq \lambda_{i,j,\ell} \leq \Lambda_{i,j,\ell}}} \sum_{l=1}^{B} \sum_{i=1}^{V} \sum_{j=1}^{E} \omega_{i,j,\ell}, \tag{2.3.8}$$

where

$$\omega_{i,j,\ell} = \lambda_{i,j,\ell} \left[R_i^{\ell} - \left(\sum_{q=1}^{Q-1} (v_i(d_q - d_Q) + e_j(\theta_q - \theta_Q)) r_{i,j,\ell}^{(q)} \right) - (v_i d_Q + e_j \theta_Q) \right]. \tag{2.3.9}$$

We assume that the furthest charging station Q is accessible to all customers with each traveling preference and that $\theta_Q \leq \theta_i, \forall i = 1, \ldots, Q-1$. This could represent an inconvenient outside option available to all customers. Additionally, for each charging station $k = 1, \ldots, Q$, we calculate $o_s = \big(v_1(d_s - d_Q) + e_E(\theta_s - \theta_Q)\big)$. Then, we label the charging stations with the set $\mathbf{s} = [s_i]_{i=1,\ldots,Q}$ such that $o_{s_1} \leq o_{s_2} \leq \ldots \leq o_{s_Q}$. The next lemma characterizes the specific order in which customers are assigned to these stations.

Lemma 2.3.2 *The optimal solution of (2.3.8) satisfies the following two properties:*

1. *If customers of type (i, j, ℓ) are assigned to station s_k, customers of type (n, j, ℓ) with $v_n < v_i$ are only assigned to stations $s_m, m \geq k$.*

2. *If customers of type* (i, j, ℓ) *are assigned to station* s_k*, customers of type* (i, n, ℓ) *with* $e_n > e_j$ *are only assigned to stations* $s_m, m \geq k$*.*

Proof: We prove both statements by contradiction. Consider the first statement. Suppose there is another optimal solution in which for the customers of type (n, j, ℓ) there is a positive probability $r^{(m)}_{n,j,\ell}$ of assignment to station s_m while customers with type (i, j, ℓ) have been assigned to a less desirable station s_k with $k > m$. However, we can have another set of routing probabilities such that $r^{(m)'}_{n,j,\ell} = \big(r^{(m)}_{n,j,\ell} - \varepsilon * \lambda_{i,j,\ell}/\lambda_{n,j,\ell}\big)$, $r^{(m)'}_{i,j,\ell} = \varepsilon * \lambda_{i,j,\ell}/\lambda_{n,j,\ell}$, and $r^{(k)'}_{i,j,\ell} = \big(r^{(k)}_{i,j,\ell} - \varepsilon * \lambda_{i,j,\ell}/\lambda_{n,j,\ell}\big)$, which lead to another feasible solution that increases the objective function of (2.3.8). Therefore, it is contradictory to the assumption of optimality of the first solution. The proof of the second statement is similar, and we remove it for brevity. ■

Lemma 2.3.3 *In the optimal solution of problem* (2.3.8)*, if charging stations* s_n *is not used in full capacity, then charging stations* s_m *with* $m > n$ *will be empty.*

The proof is provided in Appendix A-Section A.1.

The takeaway is that in this special case, 1) customers with a higher value of time and lower energy demand receive higher priority in joining stations with the lower value of o_s; 2) stations are filled in order. This special structure allows us to find the globally optimal solution of non-convex quadratic problem (2.3.8) by admitting customers with higher priority to charging stations with the lower value of o_S until they are full. Each station is then associated with a borderline similar to that of Fig. 2.1. User types that fall between the border lines of charging stations s_{k-1} and s_k will be routed to charging station s_k, whereas user types that fall on the borderline of station s_k will be partially routed to station s_k. User types that fall on the right side of the borderline of charging station s_k will not be routed to station s_k.

We consider the non-trivial case where all the customers receive positive utility from joining all the charging stations in their traveling preference (otherwise that station will be removed from the preference set). Hence, the CNO will assign customers to the charging stations until either the stations are full or all customers have been admitted. This means that we can assume that the set of available charging stations is:

$$\mathcal{X} = \{s_i : v_V(d_i - d_Q) + e_1(\theta_i - \theta_Q) \leq 0\}, \tag{2.3.10}$$

and the set of potential admittable customers is:

$$\mathcal{Y} = \{(i, j, \ell) : R_i^\ell - \left(v_i d_Q + e_j \theta_Q\right) \geq 0\}. \tag{2.3.11}$$

Exploiting the special solution structure highlighted in Lemmas 2.3.2 and 2.3.3, Algorithm 1 determines the optimal solution of problem (2.3.8). This is done by adding an extra virtual charging station, s_{Q+1}, without any capacity constraint such that:

$$s_{Q+1} \in \mathcal{G}_\ell, \forall \ell \in \mathcal{B}, \tag{2.3.12}$$

$$\left(\max_{\ell \in \mathcal{B}} R_V^\ell\right) < v_1 d_{Q+1} + e_1 \theta_{Q+1}. \tag{2.3.13}$$

Therefore, assigning customers to the charging station s_{Q+1} has a negative effect on social welfare. In step 2, it admits all types of customers in full, i.e., $\lambda_{i,j,\ell} = \Lambda_{i,j,\ell}, \forall (i, j, \ell)$. After fixing the variable $\lambda_{i,j,\ell} = \Lambda_{i,j,\ell}$, the resulting linear program (LP) of problem (2.3.8) is referred to as the Border-based Decision Problem (BDP), and its solution determines the temporary allocation (routing probabilities), denoted by $\mathbf{h}_{i,j,\ell} = [h_{i,j,\ell}^{(q)}]_{q=1,\ldots,Q+1}$, of admitted customers. It removes the partition of customers that join the virtual charging station as it is shown in step 3.

Algorithm 1: Optimal Admission and Routing

1 Add virtual station s_{Q+1} without capacity constraint
2 Set $\lambda_{i,j,\ell} = \Lambda_{i,j,\ell}$, $\mathbf{r}_{i,j,\ell} = \mathbf{0}$ $(\forall i, j, \ell)$
3 Solve BDP (temporary routing probabilities), and set:

$$r^{(q)}_{i,j,\ell} = h^{(q)}_{i,j,\ell} \text{ for } q = 1, \ldots, Q$$
$$\lambda_{i,j,\ell} = \Lambda_{i,j,\ell}(1 - h^{(Q+1)}_{i,j,\ell})$$

4 Report the optimal solution :

$$(\mathbf{R}^\star, \boldsymbol{\lambda}^\star) = \begin{cases} [r^{(q)\star}_{i,j,\ell}]_{q=1,\ldots,Q} = [r^{(q)}_{i,j,\ell}]_{q=1,\ldots,Q} \\ \lambda^\star_{i,j,\ell} = \lambda_{i,j,\ell} \end{cases}$$

Theorem 2.3.4 *Algorithm 1 will find the globally optimal solution (i.e., the globally optimal effective arrival rates and routing probabilities) for problem* (2.3.8).

The proof is provided in Appendix A-Section A.2.

Next, we consider the case of designing IC pricing-routing policies for a profit-maximizing CNO.

2.4 Profit-Maximizing Policy

In this section, we study the design of incentive-compatible pricing-routing policies with the goal of maximizing the profit earned by the CNO. Consider the following prob-

lem:

$$\max_{\substack{\mathbf{r}_{i,j,\ell}\geq 0,\\ 0\leq\lambda_{i,j,\ell}\leq\Lambda_{i,j,\ell}\\ P_{i,j,\ell}}} \sum_{\ell=1}^{B}\sum_{i=1}^{V}\sum_{j=1}^{E}\left[P_{i,j,\ell}\lambda_{i,j,\ell} - \boldsymbol{\theta}^T\mathbf{r}_{i,j,\ell}e_j\lambda_{i,j,\ell}\right].$$

$$\text{s.t.} \quad \forall i\in\mathcal{V},\forall j\in\mathcal{E},\ell\in\mathcal{B},\forall m\in\mathcal{B}_\ell: \tag{2.4.1}$$

$$\sum_{l=1}^{B}\sum_{i=1}^{V}\sum_{j=1}^{E}\lambda_{i,j,\ell}e_j r_{i,j,\ell}^{(q)}\leq C_q,\ \forall q\in\{1,\ldots,Q\}, \tag{2.4.2}$$

$$\mathbf{1}^T\mathbf{r}_{i,j,\ell}=1, \tag{2.4.3}$$

$$W_{V,j,\ell}\leq W_{V-1,j,\ell}\leq\ldots\leq W_{1,j,\ell}\leq\frac{R_1^\ell}{v_1}, \tag{2.4.4}$$

$$\sum_{t=1}^{i}(v_{t+1}-v_t)(W_{t,j,\ell}-W_{t,j,m})\leq R_1^m-R_1^\ell, \tag{2.4.5}$$

IC and IR Constraints (2.2.3)-(2.2.6) and (2.2.7).

The CNO's profit is not affected by the average wait times users experience. Instead, the objective function simply considers the revenue from services sold minus the electricity costs. The first and second constraints ensure that station capacity constraints are not violated and routing probabilities sum up to 1. The third (e.g., 2.4.4) and fourth (e.g., 2.4.5) constraints ensure that the wait times that result from the choice of $\lambda_{i,j,\ell}$ and $\mathbf{r}_{i,j,\ell}$ do not violate the requirements imposed on wait times in an IC pricing-routing policy. Note that the connection between the prices $P_{i,j,\ell}$ and the admission rate and routing probabilities $\boldsymbol{\lambda}$ and $\boldsymbol{R}$ are only through the IR and IC constraints. Accordingly, for a given set of feasible values of $\boldsymbol{\lambda}$ and $\boldsymbol{R}$, and hence $W_{i,j,\ell}(\boldsymbol{\lambda},\boldsymbol{R})$, one may maximize the prices independently to maximize revenue, as long as IR and IC constraints are not

violated. Consider the following prices:

$$\forall j \in \{1,\dots,E-1\}, \forall i \in \{1,\dots,V-1\}, \forall \ell \in \{1,\dots,B\}:$$

$$P_{i+1,j,\ell} = P_{i,j,\ell} + v_{i+1}W_{i,j,\ell} - v_{i+1}W_{i+1,j,\ell}, \tag{2.4.6}$$

$$P_{i,j+1,\ell} = P_{i,j,\ell} + v_{i}W_{i,j,\ell} - v_{i}W_{i,j+1,\ell}, \tag{2.4.7}$$

$$P_{1,1,\ell} = R_1^{\ell} - v_1 W_{1,1,\ell}. \tag{2.4.8}$$

The reader can verify that these prices are as high at horizontal IC constraints allow them to be, and hence, if they are valid, they will be revenue-maximizing. Next, we show that this is indeed the case, i.e., the prices are IC.

Proposition 2.4.1 *The prices defined in* (2.4.6)-(2.4.8) *are Incentive Compatible and Individually Rational.*

The proof is provided in Appendix A-Section A.3.

Accordingly, to find the optimal pricing-routing policy, we can simply substitute the prices from (2.4.6)-(2.4.8) in (2.4.1), allowing us to rewrite the problem with fewer decision variables and constraints:

$$\max_{\substack{\mathbf{r}_{i,j,\ell} \\ 0\le\lambda_{i,j,\ell}\le\Lambda_{i,j,\ell}}} \sum_{l=1}^{B}\sum_{j=1}^{E}\Bigg[\sum_{i=1}^{V}\Big(R_1^{\ell}\lambda_{i,j,\ell} - v_i W_{i,j,\ell}(\boldsymbol{\lambda},\boldsymbol{R})\lambda_{i,j,\ell} - \boldsymbol{\theta}^T\mathbf{r}_{i,j,\ell}e_j\lambda_{i,j,\ell}\Big) - \sum_{i=1}^{V-1}\Big((v_i - v_{i+1})(\sum_{m=i+1}^{V}\lambda_{m,j})W_{i,1,\ell}(\boldsymbol{\lambda},\boldsymbol{R})\Big)\Bigg].$$

$$\text{s.t.} \qquad \text{Constraints (2.4.2) - (2.4.5).} \tag{2.4.9}$$

The profit maximization problem (2.4.9) has a similar structure to that of (2.3.2), which we know is non-convex in general. However, we can still uniquely characterize the globally optimal solution in the special case of hard capacity constraints on charging stations,

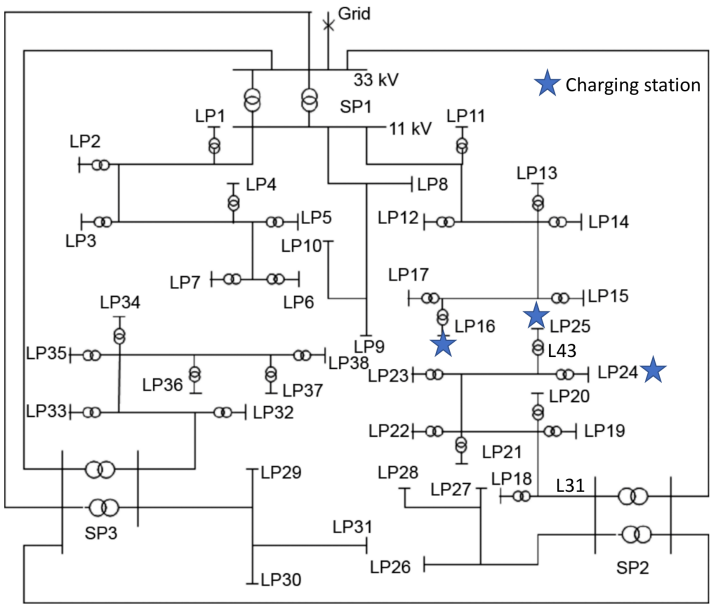

Figure 2.2: Single line diagram of bus 4 distribution system of RBTS

which is especially helpful in our numerical experiments.

2.4.1 The Special Case of Hard Capacity Constraints

In the special case of hard capacity constraints, where (2.4.9) can be rewritten as:

$$\max_{\substack{\mathbf{r}_{i,j}\geq 0 \\ 0\leq\lambda_{i,j}\leq\Lambda_{i,j}}} \sum_{q=1}^{Q-1}\sum_{l=1}^{B}\sum_{i=1}^{V}\sum_{j=1}^{E}\Bigg[R_1^{\ell}\lambda_{i,j,\ell} - \Big(\lambda_{i,j,\ell}\big[v_i(d_q-d_Q)+e_j(\theta_q-\theta_Q)\big]r_{i,j,\ell}^{(q)} + v_i d_Q + e_j\theta_Q\Big) - \left((v_i-v_{i+1})(d_q-d_Q)\left(\sum_{m=i+1}^{V}\lambda_{m,j,\ell}\right)r_{i,1,\ell}^{(q)}\right)\Bigg]. \tag{2.4.10}$$

We can show that (2.4.10) can be similarly solved through BDP linear programs. We remove the details for brevity.

Value of Time ($/h)	Energy Demand (kWh)	Traveling Preferences
$v_1 = 20$	$e_1 = 30$	$b_1 = \{s_1, s_2\}$
$v_2 = 30$	$e_2 = 40$	$b_2 = \{s_3, s_4\}$
$v_3 = 40$	$e_3 = 50$	$b_3 = \{s_5, s_6\}$
$v_4 = 50$	$e_4 = 60$	$b_4 = \{s_2, s_3\}$
$v_5 = 60$	$e_5 = 70$	$b_5 = \{s_4, s_5\}$

Table 2.1: Customers' types

2.5 Numerical Results

2.5.1 Grid Structure

To study the effect of distribution system constraints on the pricing/routing solutions, we use bus 4 of the Roy Billinton Test System (RBTS) [32]. Fig. 2.2 shows the single-line diagram of Bus 4 distribution networks. Line limit details are shown in Table 2.5.3. In the case study, we include 6 charging stations with parameters shown in Table 2.5.3. The first three stations are load points LP6, LP7, and LP15 in bus 2 of RBTS, and the rest of the charging stations are in bus 4 of RBTS as shown in Fig. 2.2. We assume that each load point with a charging station also has a commercial conventional loading with an average of 415 kW and a peak of 671.4 KW. Furthermore, for each bus, we use the locational marginal electricity prices data from [33].

2.5.2 EV Arrivals

In our case study, we assume each customer belongs to one of the 125 user types considering 5 different values of times, 5 different energy demand, and 5 different traveling preferences as it is shown in Table 2.5.1. We note that the dimension of the type grid is not a major issue and it can be further expanded if needed. We consider 24-time slots with varying potential arrival rates for each day (note that at each time slot, we solve a static problem as we have assumed that the dynamics of charging, which takes around

20 minutes, is faster than the dynamics of the variations of arrival rates). We use the Danish driving pattern in [34] to model EVs arrival rates (see Fig. 2.3).

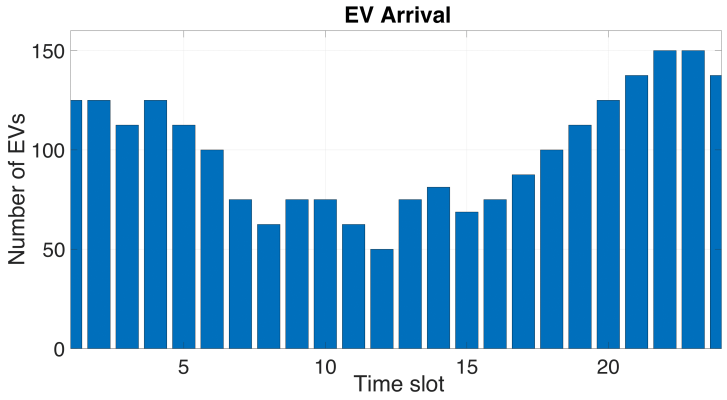

Figure 2.3: EVs arrival to the system at each time step.

We focus specifically on the special case of stations with hard capacity constraints, where our proposed Algorithm 1 can determine the globally optimal pricing-routing policy. Then we study both socially optimal and profit-maximizing scenarios. We highlight the results of our algorithm by considering both charging stations equipped with behind-the-meter solar generation and without any solar generation.

2.5.3 Experiment Results

In a socially optimal scenario, it can be seen from Fig. 2.4 that line loadings reach but do not exceed the limit at hours 14, 23, and 24, which means the distribution network constraints are active for station 6. Hence, the CNO can design an incentive-compatible pricing and routing scheme while considering the impact of EV charging in the power distribution system (in Fig. 2.4, it is shown that in the absence of distribution system constraints, the optimal pricing/routing strategy would violate network constraints).

Now, let us assume that a charging station of 6, which is the farthest charging station from customers' routes (i.e., the least desirable assignment for them in terms of traveling

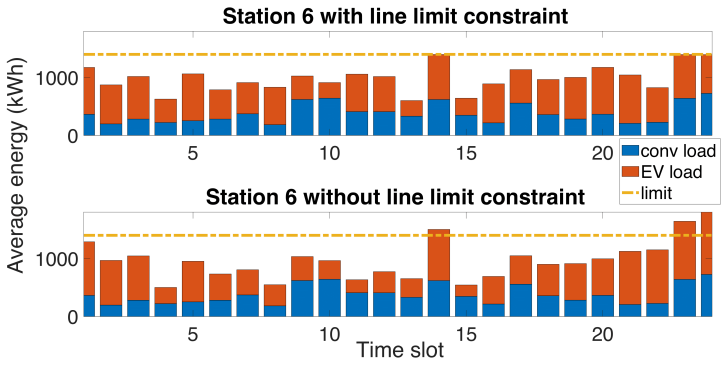

Figure 2.4: Line loading of the socially optimal problem for station 6.

Line	L31	L43
limit (kWh)	7000	1400

Table 2.2: Line loading limit

distance), can potentially be equipped with a behind-the-meter large-scale (500kW) solar system (this will require 1500m^2 of roof space to install). For the random generation profiles, we use solar data from [35] for June 2019 (one realization shown in Fig. 2.5).

The first result we highlight is the energy consumption profile of station 6 under the social-welfare maximizing scenario with available solar capacity. Essentially, by comparing energy demand with no solar generation, i.e., Fig. 2.4 and with solar generation, i.e., Fig. 2.5, we see that the availability of free solar energy makes the farthest charging station have higher levels of demand in order to maximize welfare, and so customers have to drive further on average. We will highlight this trade-off more thoroughly next.

Time travel distance (hour)	Capacity (MWh)
$d_1 = 0.03$	$c_1 = 0.6$
$d_2 = 0.06$	$c_2 = 0.7$
$d_3 = 0.09$	$c_3 = 0.8$
$d_4 = 0.12$	$c_4 = 0.6$
$d_5 = 0.15$	$c_5 = 0.8$
$d_6 = 0.18$	$c_6 = 1$

Table 2.3: Charging stations' values

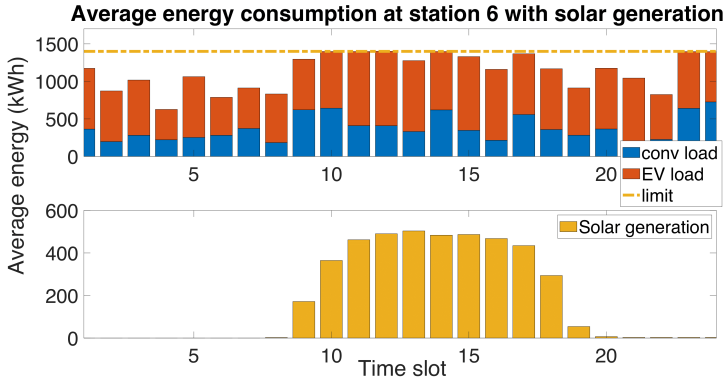

Figure 2.5: Energy demand for charging station 6 with behind-the-meter solar generation capacity.

	Socially optimal	Profit maximizing
With solar generation	9460 ($)	9320 ($)
Without solar generation	8280 ($)	8440 ($)

Table 2.4: Cost of traveling of all customers over a day

Specifically, Table 2.5.3 shows the cost of traveling from the main corridor to reach charging stations for all types of customers with vehicle arrivals shown in Fig. 2.3. We calculate $\sum_{l=1}^{B}\sum_{i=1}^{V}\sum_{j=1}^{E} v_i \lambda_{i,j,\ell} W_{i,j,\ell}$ as the cost of traveling in both socially optimal and profit-maximizing scenarios over a day. Without solar generation, for both cases in which the objective is to maximize social welfare and maximize profits, customers with a higher VoT and lower energy demand have priority in joining the closer charging stations. With solar generation, in the socially optimal case, customers with higher energy demand are assigned to the furthest charging station even to get cheaper electricity, and the traveling cost is larger. However, for the profit-maximizing case, customers with a higher value of time (and hence higher willingness to pay) are still assigned to the closer charging stations (and are charged more), and the overall cost of traveling is less than when the objective is to maximize social welfare, and larger than not having solar generation.

We would like to note that the concept of incentive compatibility as highlighted in our work only applies to each individual's incentive for incorrectly reporting their type to

Energy demand (kWh)	50	60	40
Value of time ($/h)	20	30	40
Reward ($)	440	635	845
Locational marginal price ($/kWh)	0.5	0.4	0.3
Time travel distance (h)	0.3	0.6	0.9

Table 2.5: Parameters

the CNO under the differentiated service program. The algorithm provides no guarantee that every individual is better off under the differentiated SO policy than they would be under a Nash Equilibrium with no centralized routing, hence incentivizing them to request the existence of the differentiated service program. This is considered normal since any type of congestion pricing mechanism (including locational marginal pricing) to maximize welfare could lead to cost increases for some individuals but overall improve welfare for society.

2.5.4 Bench-marking with status-quo

The goal of this experiment is to highlight the benefits of a mobility-aware differentiated service mechanism as opposed to self-routing by customers to stations, which can be considered the status-quo. We have compared the performance of our proposed solution to the equilibrium load and wait time pattern at the stations in the scenario where users self-route. We assume that in the self-routing scenario, customers will be charged at locational marginal prices for energy (which can vary across stations). For the experiment, we assume 3 different user types, and 3 charging stations (this is clearly not a realistic choice of the parameters, but computing all the equilibria is computationally challenging in bigger cases). The values we used for the numerical experiment are shown in Table 2.5.4:

Then, we let the customers selfishly choose the charging station they want to charge at in order to maximize their utility. We need to note that multiple Nash equilibria may

exist for this game. In our setup, there exist 4 different equilibria, and the values of social welfare are $7290.9\$, 7302.1\$, 7312.1\$, 7328.1\$$. Observe that they are all less than the value of social welfare achieved using our proposed solution based on differentiated services, which is $7398.9\$$. We can argue that this is a natural observation given the lack of appropriate congestion pricing schemes that can deter users from the most popular choice of stations. We note that congestion pricing to guide users towards a socially-optimal charge footprint while considering station capacities is not straightforward to apply in this case for reasons explained in the Introduction.

2.6 Conclusions

We studied the decision problem of a CNO for managing EVs in a public charging station network through differentiated services. In this case, EV users cannot directly choose which charging station they will charge at. Instead, they choose their energy demand and their priority level, as well as their traveling preferences (which stations they are willing to visit) from among a menu of service options that are offered to them, and the CNO then assigns them to the charging stations directly to control station wait times and electricity costs. This is reminiscent of incentive-based direct load control algorithms that are very popular in demand response. We propose incentive-compatible pricing and routing policies for maximizing the social welfare or the profit of the CNO. We proposed an algorithm that finds the globally optimal solution for the non-convex optimizations that appear in our work in the special case of hard capacity constraints in both social welfare and profit maximization scenarios and highlighted the benefits of our algorithms towards behind-the-meter solar integration at the station level. [36, 37, 22] are the the results of this work.

Chapter 3

Stochastic Linear Bandit: An Overview

3.1 Introduction

In this chapter, we present a summary on the stochastic linear bandit (LB) problem. Stochastic bandit optimization is a sequential decision-making problem that has long found applications in many fields where some characteristics of the users' response are not known and can only be learned through a limited number of noisy observations, including recommendation engines, advertisement placement, personalized medicine, etc. The learner's objective for the overall learning task consists of maximizing the cumulative reward gained during T rounds of interaction with the user. The expected reward gained at each round t is a function $f(x_t)$ of the feature vector x_t associated with each action that the learner chooses to play, and f is not known to the learner. There is rich literature covering parametric or non-parametric characterizations of f, as well as finite or continuous action sets.

Two popular algorithms have been studied in order to capture the trade-off between

exploration and exploitation in sequential decision-making problems: 1) Upper confidence bound (UCB), which consists of choosing the optimal action according to the upper-confidence bounds on the true parameter (i.e., $\theta_\star$) [38]; 2) Thompson Sampling (TS), which samples the true parameter from a prior distribution, and selects the optimal action with respect to the sampled parameter [39]. [40] formalized the linear bandit problem which is the extension of the Multi-armed bandit problem introduced by [41]. In [40], the arms are associated with vectors in $\mathbb{R}^d$, and the reward is a noisy and unknown linear function of the arms, and they derived an *optimistic* algorithm that relies on the least-square estimation of the unknown reward parameter. Further, [42] introduced new concentration inequalities for the least square estimates which allows them to improve the regret bound of [40]. Similarly, a Thompson Sampling algorithm that has shown a good empirical performance can be derived for the LB problem. [39] provided the first regret analysis for the LB problem, and later on, [43] proposed the linear Thompson Sampling Algorithm with least square concentration inequalities. In this chapter, we first present the LB setting, and then we present the OFUL and Linear Thompson Sampling algorithms with their regret guarantees.

3.2 Problem Setting

In LB, the learner is given a convex and compact set of actions $\mathcal{X} \subset \mathbb{R}^d$. At each round t, playing an action $x_t \in \mathcal{X}$ results in observing reward:

$$r_t := x_t^\top \theta_* + \xi_t \tag{3.2.1}$$

where $\theta_* \in \mathbb{R}^d$ is a fixed but unknown vector that describes the users' characteristics, and ξ_t is a zero-mean noise. Here, the expected reward is linear in the action x_t, i.e.,

$f(x_t) = x_t^\top \theta_*$.

We denote the optimal arm as:

$$x_* = \underset{x \in \mathcal{X}}{\operatorname{argmax}}\, x^\top \theta_* \tag{3.2.2}$$

At each step t, the learner selects an arm $x_t \in \mathcal{X}$ based on the past observations (and possibly additional randomization), it observes the reward $r_t := x_t^\top \theta_* + \xi_t$ and it suffers a regret equal to the difference in expected reward between the optimal arm x_* and the arm x_t. All the information observed up to time t is encoded in the filtration $\mathcal{F}_t = (\mathcal{F}_1, \sigma(x_1, \ldots, x_t, \xi_1, \ldots, \xi_t,))$, where $\mathcal{F}_1$ contains any prior knowledge. The objective of the learner is to minimize the *cumulative pseudo-regret* up to round T:

$$R(T) = \sum_{t=1}^{T} x_*^\top \theta_* - x_t^\top \theta_*. \tag{3.2.3}$$

We make the following standard assumptions on the noise distribution, the reward parameter, and the actions.

Assumption 3 *For all t, ξ_t is conditionally zero-mean R-sub-Gaussian noise variables, i.e., $\mathbb{E}[\xi_t | \mathcal{F}_{t-1}] = 0$, and $\mathbb{E}[e^{\lambda \xi_t} | \mathcal{F}_{t-1}] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right), \forall \lambda \in \mathbb{R}^d$.*

Assumption 4 *There exists a positive constant S such that $\|\theta_\star\|_2 \leq S$.*

Assumption 5 *The action set $\mathcal{X}$ is a compact and convex subset of $\mathbb{R}^d$ that contains the unit ball. We assume that $\|x\|_2 \leq L, \forall x \in \mathcal{X}$. Also, we assume $\langle x, \theta_\star \rangle \leq 1, \forall x \in \mathcal{X}$.*

3.3 Regularized Least-Square Estimation

The LB problem is characterized by bandit feedback, i.e., the learner only observes the rewards without any additional information on θ_*. However, an estimate $\hat{\theta}_t$ can be

computed using the standard least square procedure. Let $(x_1, \ldots, x_t)$ be the sequence of arms chosen up to round t, and $(r_1, \ldots, r_t)$ be their corresponding rewards. Then, θ_* can be estimated by regularized least-square (RLS). For any parameter, $\lambda > 0$, the design matrix and the RLS estimates are:

$$V_t = \lambda I + \sum_{s=1}^{t-1} x_s x_s^\top, \ \hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t-1} x_s r_{s+1}. \tag{3.3.1}$$

Leveraging the theory of self-normalized processes, [42] derived a new concentration inequality for the RLS estimate.

Proposition 3.3.1 *Let Assumptions 3, 4, and 5 hold. For any fixed $\delta \in (0, 1)$, and*

$$\beta_t(\delta) = R\sqrt{d \log\left(\frac{1 + \frac{tL^2}{\lambda}}{\delta}\right)} + \sqrt{\lambda} S \tag{3.3.2}$$

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

$$\left\|\hat{\theta}_t - \theta_\star\right\|_{V_t} \leq \beta_t(\delta). \tag{3.3.3}$$

3.4 Optimism in Face of Uncertainty for Linear Bandit

In this section, we present the Optimism in Face of Uncertainty for the Linear bandit (OFUL) algorithm of [42]. The summary of the OFUL algorithm is presented in Algorithm 2.

Algorithm 2: OFUL algorithm

Input: $\delta, T, \lambda, V_1 = \lambda I$.
for $t = 1, \ldots, T$ **do**
  Build the confidence region: $\mathcal{C}_t(\delta) = \{\theta \in \mathbb{R}^d : \left\|\theta - \hat{\theta}_t\right\|_{V_t} \leq \beta_t(\delta)\}$
  Select the action-parameter pair: $(x_t, \tilde{\theta}_t) = \text{argmax}_{x \in \mathcal{X}, \theta \in \mathcal{C}_t(\delta)} x^\top \theta$
  Play action x_t, and observe reward $r_t = x_t^\top \theta_* + \xi_t$
  Update the RLS estimate $\hat{\theta}_{t+1}$ and design matrix V_{t+1} using (3.3.1).
end for

At each time step t, for a fixed $\delta \in (0, 1)$, OFUL constructs a confidence set $\mathcal{C}_t(\delta)$ as:

$$\mathcal{C}_t(\delta) = \{\theta \in \mathbb{R}^d : \left\|\theta - \hat{\theta}_t\right\|_{V_t} \leq \beta_t(\delta)\}. \tag{3.4.1}$$

Proposition 3.3.1 guarantees that $\theta_* \in \mathcal{C}_t$ with probability at least $1 - \delta$. Then at each time step, OFUL selects the optimistic action-parameter pair as

$$(x_t, \tilde{\theta}_t) = \underset{x \in \mathcal{X}, \theta \in \mathcal{C}_t(\delta)}{\text{argmax}} \; x^\top \theta. \tag{3.4.2}$$

Under the Assumptions 3, 4, and 5, [42] proved the following result on the regret bound of OFUL algorithm:

Theorem 3.4.1 *For any $\delta \in (0, 1)$, with probability at least $1-\delta$, the regret of the OFUL algorithm is bounded as:*

$$R(T) \leq 2\beta_T \sqrt{2dT \log(1 + \frac{TL^2}{\lambda d})}, \tag{3.4.3}$$

where $\beta_t = R\sqrt{d \log\left(\frac{1 + \frac{tL^2}{\lambda}}{\delta}\right)} + \sqrt{\lambda} S$.

3.5 Thompson Sampling Algorithm for Linear Bandit

In this section, we present Linear Thompson Sampling (LTS) algorithm from [43]. First, [44] define Thompson Sampling (TS) for linear bandit as a Bayesian algorithm where a Gaussian prior over θ_* is updated according to the observed reward, a random sample is drawn from the posterior, and the corresponding optimal arm is selected at each step. Then, [43] show that LTS can be defined as a generic randomized algorithm constructed on the RLS-estimate rather than sampling from a Bayesian posterior. The summary of LTS is presented in Algorithm 3.

Algorithm 3: LTS algorithm

Input: $\delta, T, \lambda, V_1 = \lambda I$.
Set $\delta' = \delta/(4T)$
for $t = 1, \ldots, T$ **do**
  Sample $\eta_t \sim \mathcal{D}^{\text{TS}}$, and compute $\tilde{\theta}_t = \hat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta_t$
  Select the action : $x_t = \text{argmax}_{x \in \mathcal{X}}\, x^\top \tilde{\theta}_t$
  Play action x_t, and observe reward $r_t = x_t^\top \theta_* + \xi_t$
  Update the RLS estimate $\hat{\theta}_{t+1}$ and design matrix V_{t+1} using (3.3.1).
end for

At each time step t, given RLS-estimate $\hat{\theta}_t$ and the design matrix V_t, LTS samples a perturbed parameter $\tilde{\theta}_t$ as:

$$\tilde{\theta}_t = \hat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta_t, \tag{3.5.1}$$

where η_t is a random sample drawn i.i.d. from a suitable multivariate distribution $\mathcal{D}^{\text{TS}}$. Then, the optimal arm associated with the perturbed parameter is chosen and the corresponding reward is observed. LTS requires solving only a linear optimization problem in contrast with OFUL which requires solving a bilinear optimization problem in (3.4.2).

The key idea of [44, 43] on how to select the random perturbation $\eta_t \in \mathbb{R}^d$ to guarantee good regret performance is as follows. On the one hand, $\tilde{\theta}_t$ must stay close enough to the RLS-estimate $\hat{\theta}_t$ so that $x_t^\top \tilde{\theta}_t$ is a good proxy for the true (but unknown) reward $x_t^\top \theta_\star$. Thus, η_t must satisfy an appropriate *concentration* property. On the other hand, $\tilde{\theta}_t$ must also favor exploration in a sense that it leads –often enough– to actions x_t that are *optimistic*, i.e., they satisfy

$$x_t^\top \tilde{\theta}_t \geq x_\star^\top \theta_\star \tag{3.5.2}$$

Thus, η_t must satisfy an appropriate *anti-concentration* property. This translates into the following conditions on $\mathcal{D}^{\mathrm{TS}}$.

Definition 3.5.1 *$\mathcal{D}^{TS}$ is a multivariate distribution on $\mathbb{R}^d$ absolutely continuous with respect to the Lebesgue measure which satisfies the following properties:*

- *Anti-concentration: there exists a constant $p > 0$ such that for any $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$:*

$$\mathbb{P}(u^\top \eta_t \geq 1) \leq p. \tag{3.5.3}$$

- *Concentration: There exists constants $c, c' > 0$, such that $\forall \delta \in (0, 1)$:*

$$\mathbb{P}\left(\|\eta_t\|_2 \leq \sqrt{cd \log(\frac{c'd}{\delta})} \right) \geq 1 - \delta. \tag{3.5.4}$$

[43] interpreted that the definition of $\mathcal{D}^{\mathrm{TS}}$ requires LTS to explore far enough from $\hat{\theta}_t$ (anti-concentration) but not too much (concentration). This implies that LTS performs "useful" exploration with enough frequency (notably it performs optimistic steps), but without selecting arms with too large regret.

[43] also prove a regret bound for the LTS algorithm. To do that, they first use the following standard decomposition of the cumulative regret:

$$R(T) = \sum_{t=1}^{T} \big(\underbrace{x_\star^\top \theta_\star - x_t^\top \tilde{\theta}_t}_{\text{Term I}} \big) + \sum_{t=1}^{T} \big(\underbrace{x_t^\top \tilde{\theta}_t - x_t^\top \theta_\star}_{\text{Term II}} \big), \tag{3.5.5}$$

where Term I depends on the randomization of LTS, and Term II mostly depends on the properties of RLS-estimation. Regarding Term II, the concentration property of $\mathcal{H}^{TS}$ guarantees that $\tilde{\theta}_t$ is close to $\hat{\theta}_t$, and consequently, close to $\theta_\star$ thanks to Proposition 3.3.1. Therefore, controlling Term II can be done similarly to previous works e.g., [42].

Theorem 3.5.2 (Regret of LTS) *Let $\lambda \geq 1$ and Assumptions 3, 4, and 5 hold. Fix $\delta \in (0,1)$. Then, with probability at least $1-\delta$, the regret of LTS is upper bounded as follows:*

$$R(T) \leq \big(\beta_T(\delta') + \gamma_T(\delta')(1+\frac{4}{p})\big)\sqrt{2Td\log\big(1+\frac{TL^2}{\lambda}\big)} + \frac{4\gamma_T(\delta')}{p}\sqrt{\frac{8TL^2}{\lambda}\log\frac{4}{\delta}}, \tag{3.5.6}$$

where $\delta' = \frac{\delta}{4T}$, $\beta_t(\delta')$ as in (3.3.2) and, $\gamma_t(\delta') = \beta_t(\delta')\big(1+\frac{2}{C}LS\big)\sqrt{cd\log\big(\frac{c'd}{\delta'}\big)}$.

In summary, the two efficient approaches for the LB problem:

- linear UCB (LUCB): [42] provides a regret bound of order $\mathcal{O}(dT^{1/2}\log T)$.
- linear Thompson Sampling (LTS): [43] adopt a frequentist view and show a regret bound of order $\mathcal{O}(d^{3/2}\log^{1/2} dT^{1/2}\log^{3/2} T)$.

Chapter 4

Safety-constrained Bandit Algorithms with Applications to Human-Cyber-Physical Systems

4.1 Introduction

The application of stochastic bandit optimization algorithms to safety-critical systems has received significant attention in the past few years. In such cases, the learner repeatedly interacts with a system with an uncertain reward function and operational constraints. In spite of this uncertainty, the learner needs to ensure that her actions do not violate the operational constraints *at any round of the learning process*. This shares a similar challenge to the one in H-CPS due to the involvement of humans in the control loop. For example, in electricity pricing for societal-scale infrastructure systems such as power grids or transportation networks where minimizing the operational costs with a limited number of user interactions. In this case, it is required that the operational constraints of the power grid are not violated in response to our actions. However, the

existing bandit algorithms might not be directly applicable to these cases due to the existence of infrastructural safety constraints that have to be met at each round of user interactions. Especially in the earlier rounds, there is a need to choose actions with caution, while at the same time making sure that the chosen action provides sufficient learning opportunities about the set of safe actions. This uncertainty about safety and the resulting conservative behavior means the learner could experience additional costs in such constrained environments.

In this chapter we focus on linear stochastic bandits (LB) (see Chapter 3 for details) where each action is associated with a feature vector x, and the expected reward of playing each action is equal to the inner product of its feature vector and an unknown parameter vector $\theta^\star$. Two efficient approaches have been developed: *linear UCB* (LUCB) and *linear Thompson Sampling* (LTS). For LUCB, [42] provides a regret bound of order $\mathcal{O}(d\cdot T^{1/2}\log T)$. For LTS [44, 43] adopt a frequentist view and show regret $\mathcal{O}(d^{3/2}\log^{1/2} d\cdot T^{1/2}\log^{3/2} T)$. Here we provide an LTS algorithm that respects *linear safety constraints* and study its performance.

4.2 Problem Setting

Reward function. The learner is given a convex and compact set of actions $\mathcal{D}_0 \subset \mathbb{R}^d$. At each round t, playing an action $x_t \in \mathcal{D}_0$ results in observing reward $r_t := x_t^\top \theta_\star + \xi_t$, where $\theta_\star \in \mathbb{R}^d$ is a fixed, but *unknown*, parameter and ξ_t is a zero-mean additive noise.

Safety constraint. We further assume that the environment is subject to a linear constraint:

$$x_t^\top \mu_\star \leq C, \tag{4.2.1}$$

which needs to be satisfied by the action x_t at every round t, to guarantee the safe operation of the system. Here, C is a *positive* constant that is *known* to the learner, while $\mu_\star$ is fixed, but *unknown* vector parameter. Let us denote the set of "safe actions" that satisfy the constraint (4.2.1) as follows:

$$\mathcal{D}_0^s(\mu_\star) := \{x \in \mathcal{D}_0 : x^\top \mu_\star \leq C\}. \tag{4.2.2}$$

By having $C > 0$ and further assuming that $0 \in \mathcal{D}_0$, we know that the action 0 is always as safe action. However, beyond that $\mathcal{D}_0^s(\mu_\star)$ is unknown to the learner, since $\mu_\star$ is itself unknown. We consider a bandit-feedback setting in which, at every round t, the learner receives *side information* about the safety set via noisy measurements:

$$w_t = x_t^\top \mu_\star + \zeta_t, \tag{4.2.3}$$

where ζ_t is zero-mean additive noise. During the learning process, the learner needs a mechanism that allows her to use the side measurements in (4.2.3) for determining the safe set $\mathcal{D}_0^s(\mu_\star)$. This is critical since it is required (at least with high probability) that $x_t \in \mathcal{D}_0^s(\mu_\star)$ for all rounds t.

Regret. The *cumulative pseudo-regret* for T rounds is $R(T) = \sum_{t=1}^{T} x_\star^\top \theta_\star - x_t^\top \theta_\star$, where $x_\star = \arg\max_{x \in \mathcal{D}_0^s(\mu^*)} x^\top \theta_\star$ is the optimal *safe* action that maximizes the expected reward over $D_0^s(\mu_\star)$.

Learning goal. The learner's objective is to control the growth of the pseudo-regret. Moreover, we require that the chosen actions $x_t, t \in [T]$ are safe (i.e., they belong to $\mathcal{D}_0^s(\mu_\star)$ in (4.2.2)), with high probability over T rounds. As is common, we use regret to refer to the pseudo-regret $R(T)$.

Example. As a concrete motivation example of our setting, consider medical trials,

a problem traditionally advocated as an application area for linear bandits, where the effect of different therapies is unknown a-priori to the doctors and can only be determined through clinical trials. Free exploration is not possible, since it may lead to actions that cause harm to the patient, an outcome to be avoided at all times. To model this, we pick the unknown parameter $\mu_\star$ so as to represent the patients' response, and the known parameter C so as to represent a safety threshold that doctors need to account for. The hazard-threshold C can be assumed known as it is the same for all patients (and can be estimated from existing data). In this example, actions x_t represent selected therapies at time t (e.g., drug dosage) and we assume that a (conservative) safe seed set of harmless (but, plausibly not efficient) therapies is known to the doctor. Overall, while doctors try to select therapies (x_t) with high rewards (which could be a signal that shows improvement in a patient's health condition), they should not violate the safety constraint $x_t^\top \mu_\star \leq C$ at any time.

4.2.1 Contribution

We do believe that albeit simple, linear models for safety constraints could be directly relevant in traditionally advocated applications of bandit problems such as medical trials applications [45], recommendation systems [46], and ad placement [47]. Even in more complex settings where linear models are not directly applicable, we still believe that this is the appropriate first step toward a principled study of the performance of safe algorithms.

The contribution of our work in this chapter are the following:

- We provide the first *safe* Linear Thompson Sampling (Safe-LTS) algorithm with provable regret guarantees for the linear bandit problem with linear safety constraints.
- Our analysis shows that Safe-LTS achieves the *same* order $\mathcal{O}(d^{3/2}\log^{1/2} d{\cdot}T^{1/2}\log^{3/2} T)$

of regret as the original LTS (without safety constraints) [43]. Hence, the dependence of our regret bound on the time horizon T *cannot* be improved modulo logarithmic factors (see lower bounds in [40, 48]).

- We compare Safe-LTS to existing safe versions of LUCB. We show that our algorithm has: better regret in the worst-case, fewer parameters to tune, and superior empirical performance.
- We propose a heuristic modification to our Safe-LTS algorithm that adapts a *dynamic noise-distribution scheme* and is shown empirically to outperform the latter. This idea might also be relevant in the unconstrained linear bandit setting.

On a technical level, to derive Safe-LTS and its regret bound, need to properly account for the fact that the optimal safe action $x_\star$ is *not* necessarily in the estimated safe decision set (see Eqn. (4.3.3) for formal definition) at each round t. This is because, at each time step, we only have an estimate of the unknown parameter μ_*, thus the estimated set is only a conservative inner approximation of the actual safe set in (4.2.2). Consequently, we need to design an action selection rule that is simultaneous: (i) *Frequently optimistic in spite of limitations on actions imposed because of safety.* Here, we achieve this by appropriately tuning the randomization of Thompson Sampling. Specifically, through careful analysis, essentially controlling the distance of the optimal action $x_\star$ from the estimated safe set, we find that the appropriate tuning involves scaling with a simple function of the problem parameters including the safety constant C. (ii) Guarantees a *proper expansion of the estimated safe set* so as to not exclude *good* actions for a long time, leading to large *regret of safety.* Here, we show that it is the **randomized** nature of LTS that achieves this second goal, and this is exactly where the LUCB action selection rule seems to fail to provide the same guarantees.

4.2.2 Other Related Work

Multi-armed Bandits (MAB) - Two popular algorithms have been studied in MAB in order to capture the trade-off between exploration and exploitation in sequential decision-making problems: 1) Upper confidence bound (UCB), which consists of choosing the optimal action according to the upper-confidence bounds on the true parameter (i.e., $\theta_\star$) [38]; 2) Thompson Sampling (TS), which samples the true parameter from a prior distribution, and selects the optimal action with respect to the sampled parameter [39]. Moreover, [49] considers a new approach to the MAB problem based on Deterministic Sequencing of Exploration and Exploitation (DSEE). In particular, they divide the time horizon into the pure exploration phase and the pure exploitation phase. In the former, the player plays all arms in a round-robin fashion. In the latter, the player plays the arm with the largest sample mean. [50, 51] study the MAB problem in multiplayer settings where a team of agents cooperates on a network in order to maximize their collective reward. In [52, 53], they study the multi-objective MAB problem where the components of the reward signal correspond to different objectives. They evaluate the performance of their algorithm with notions of 2-D regret and Pareto regret. Other lines of work have studied the best-arm identification problem in MAB that aims to identify the arm with the largest expected regret [54] as well as cascading bandits where the goal is to learn arms in order to rank them based on the users' preferences such as recommendation systems [55]. In [56, 57], they study the MAB problem given adversarial attacks, where the adversary can change the action selected by the learner, and they propose a robust algorithm for the case that the total attack cost is given. Also, [58] studies the MAB problem in the case that the statistical rewards of different arms may be correlated. In particular, they study the regional bandits problem where the arms belong to different groups such the expected reward of the arms in the same group is a function of the

common parameter, and the parameters are independent across different groups. Another line of work focuses on the design of risk-sensitive algorithms [59, 60, 61]. In particular, in economic and finance applications, the learner may be more interested in reducing the uncertainty (i.e., risk) in the outcome, rather than achieving the highest cumulative reward [62, 63].

Safety - A diverse body of related works on stochastic optimization and control have considered the effect of safety constraints that need to be met during the run of the algorithm [64, 65] and references therein. Closely related to our work, [66, 67] study *non-linear* bandit optimization with *nonlinear* safety constraints using Gaussian processes (GPs) as non-parametric models for both the reward and the constraint functions. Their algorithms have shown great promise in robotics applications [68, 69]. Without the GP assumption, [45] proposes and analyzes a safe variant of the Frank-Wolfe algorithm to solve a smooth optimization problem with an unknown convex objective function and unknown *linear* constraints (with side information, similar to our setting). All the above algorithms come with provable convergence guarantees, but *no* regret bounds. To the best of our knowledge, the first work that derived an algorithm with provable regret guarantees for bandit optimization with stage-wise safety constraints, as the ones imposed on the aforementioned works, is [70]. While [70] restricts attention to a *linear* setting, their results reveal that the presence of the safety constraint –even though linear– can have a non-trivial effect on the performance of LUCB-type algorithms. Specifically, the proposed Safe-LUCB algorithm comes with a problem-dependent regret bound that depends critically on the location of the optimal action in the safe action set – increasingly so in problem instances for which the safety constraint is active. In [70], the linear constraint function involves the same unknown vector (say, $\theta_\star$) as the one that specifies the linear reward. Instead, in Section 4.2 we allow the constraint to depend on a new parameter vector (say, $\mu_\star$) to which the learner gets access via side-information measurements (4.2.3).

This latter setting is the direct *linear* analog to that of [66, 67, 45] and we demonstrate that an appropriate Safe-LTS algorithm enjoys regret guarantees of the same order as the original LTS *without* safety constraints. A more elaborate comparison to [70] is provided in Section 4.5.3. In contrast to the previously mentioned references, another recent work [46] defines safety as the requirement of ensuring that the *cumulative* (linear) reward up to each round stays *above* a given percentage of the performance of a known *baseline* policy. A "stage-wise" variant of this type of constraint was recently studied in another interesting work [71]. Compared to [46], the setting of [71] is closer to ours, but there are still some key differences. Most notably, in contrast, to [71], the constraint studied here is such that the optimal action $x_\star$ is not guaranteed to be in the estimated safe-set (especially at early rounds t). Because of this, the analysis of [71] is not directly applicable here. On a technical side, [71] proves a bound on the expected reward (but they restrict actions to an ellipsoidal). Instead, we present a high-probability bound on the regret similar to [46, 70]. Moreover, [72] also considers more relaxed safety constraints with respect to ours (from high probability to expectation) in the bandit setting. They propose an optimism-pessimism algorithm for both linear bandit and MAB problems.

Also, it is worth mentioning that the algorithms presented in [70, 71] require distinct rounds of randomization that are dedicated to learning the unknown constraints. Instead, our analysis shows that the inherent randomization of the TS action selection rule suffices for this purpose. As a closing remark, [70, 46, 45, 71, 73, 74, 75] show that simple linear models for safety constraints might be directly relevant to several applications such as medical trials, recommendation systems, managing the customers' demand in power-grid systems. Moreover, even in more complex settings where linear models do not directly apply (e.g., [68, 69]), we believe that this simplification is an appropriate first step towards a principled study of regret behavior of safe algorithms in sequential decision settings.

Thompson Sampling - Even though TS-based algorithms [76] are computationally

easier to implement than UCB-based algorithms and have shown great empirical performance, they were largely ignored by the academic community until a few years ago, when a series of papers [77, 43, 39, 78] showed that TS achieves optimal performance in both frequentist and Bayesian settings. Most of the literature focused on the analysis of the Bayesian regret of TS for general settings such as linear bandits or reinforcement learning (see e.g., [79]). More recently, [80, 81, 82] provided an information-theoretic analysis of TS. Additionally, [83] provides regret guarantees for TS in the finite and infinite MDP setting. Another notable paper is [84], which studies the stochastic MAB problem in complex action settings providing a regret bound that scales logarithmically in time with improved constants. None of these papers study the performance of TS for LB with safety constraints.

4.3 Safe Linear Thompson Sampling

Our proposed algorithm is a safe variant of Linear Thompson Sampling (LTS). At any round t, given a regularized least-squares (RLS) estimate $\hat{\theta}_t$, the algorithm samples a perturbed parameter $\tilde{\theta}_t$ that is appropriately distributed to guarantee sufficient exploration. Considering this sampled $\tilde{\theta}_t$ as the true environment, the algorithm chooses the action with the highest possible reward while making sure that the safety constraint (4.2.1) holds. The presence of the safety constraint complicates the learner's choice of actions. In order to ensure that actions remain safe at all rounds, the algorithm uses the side-information (4.2.3) to construct a confidence region $\mathcal{C}_t$, which contains the unknown parameter $\mu_\star$ with high probability. With this, it forms an *inner* approximation $\mathcal{D}_t^s$ of the safe set, which is composed of all actions x_t that satisfy the safety constraint *for all* $v \in \mathcal{C}_t$. The summary is presented in Algorithm 4 and a detailed description follows.

Algorithm 4: Safe Linear Thompson Sampling (Safe-LTS)

Input: δ, T, λ. Set $\delta' = \frac{\delta}{6T}$
for $t = 1, \ldots, T$ **do**
  Sample $\eta_t \sim \mathcal{H}^{\mathrm{TS}}$
  Set $V_t = \lambda I + \sum_{s=1}^{t-1} x_s x_s^\top$ and compute RLS-estimates $\hat{\theta}_t$ and $\hat{\mu}_t$
  Set: $\tilde{\theta}_t = \hat{\theta}_t + \beta_t(\delta') V_t^{-\frac{1}{2}} \eta_t$
  Build the confidence region: $\mathcal{C}_t(\delta') = \{v \in \mathbb{R} : \|v - \hat{\mu}_t\|_{V_t} \leq \beta_t(\delta')\}$
  Compute the estimated safe set: $\mathcal{D}_t^s = \{x \in \mathcal{D}_0 : x^\top v \leq C, \forall v \in \mathcal{C}_t(\delta')\}$
  Play the following action: $x_t = \arg\max_{x \in \mathcal{D}_t^s} x^\top \tilde{\theta}_t$
  Observe reward r_t and measurement w_t
end for

4.3.1 Model Assumptions

Notation. $[n]$ denotes the set $\{1, 2, \ldots, n\}$. The Euclidean norm of a vector x is denoted by $\|x\|_2$. Its weighted ℓ_2-norm with respect to a positive semidefinite matrix V is denoted by $\|x\|_V = \sqrt{x^\top V x}$. We also use the standard $\widetilde{\mathcal{O}}$ notation that ignores poly-logarithmic factors. Finally, for ease of notation, from now on-wards we refer to the safe set in (4.2.2) by $\mathcal{D}_0^s$ and drop the dependence on $\mu_\star$. Let $\mathcal{F}_t = (\mathcal{F}_1, \sigma(x_1, \ldots, x_t, \xi_1, \ldots, \xi_t, \zeta_1, \ldots, \zeta_t))$ denote the filtration representing the accumulated information up to round t. We also introduce standard assumptions on the problem as follows.

Assumption 6 *For all t, ξ_t and ζ_t are conditionally zero-mean, R-sub-Gaussian noise variables, i.e., $\mathbb{E}[\xi_t|\mathcal{F}_{t-1}] = \mathbb{E}[\zeta_t|\mathcal{F}_{t-1}] = 0$, $\mathbb{E}[e^{\alpha\xi_t}|\mathcal{F}_{t-1}] \leq \exp\left(\frac{\alpha^2 R^2}{2}\right)$, $\mathbb{E}[e^{\alpha\zeta_t}|\mathcal{F}_{t-1}] \leq \exp\left(\frac{\alpha^2 R^2}{2}\right), \forall \alpha \in \mathbb{R}$.*

Assumption 7 *There exists a positive constant S such that $\|\theta_\star\|_2 \leq S$ and $\|\mu_\star\|_2 \leq S$.*

Assumption 8 *The action set $\mathcal{D}_0$ is a star-convex subset of $\mathbb{R}^d$ and contains the origin. We assume $\|x\|_2 \leq L$, $\forall x \in \mathcal{D}_0$.*

It is straightforward to generalize our results when the sub-Gaussian constants of ξ_t and ζ_t and/or the upper bounds on $\|\theta_\star\|_2$ and $\|\mu_\star\|_2$ are different. Throughout, we assume they are equal, for brevity.

4.3.2 Algorithm description and discussion

Let $\{x_i\}_{i\in[t]}$ be the sequence of actions and $\{r_i\}_{i\in[t]}$, $\{w_i\}_{i\in[t]}$ be the corresponding rewards and side-information measurements. For any $\lambda > 0$, the RLS-estimates $\hat{\theta}_t$ of $\theta_\star$ and $\hat{\mu}_t$ of $\mu_\star$ are $\hat{\theta}_t = V_t^{-1}\sum_{s=1}^{t-1} r_s x_s$, $\hat{\mu}_t = V_t^{-1}\sum_{s=1}^{t-1} w_s x_s$, where $V_t = \lambda I + \sum_{s=1}^{t-1} x_s x_s^\top$. Based on $\hat{\theta}_t$ and $\hat{\mu}_t$, we construct two confidence regions $\mathcal{E}_t := \mathcal{E}_t(\delta')$ and $\mathcal{C}_t := \mathcal{C}_t(\delta')$ as follows:

$$\mathcal{E}_t := \{\theta \in \mathbb{R}^d : \left\|\theta - \hat{\theta}_t\right\|_{V_t} \leq \beta_t(\delta')\}, \tag{4.3.1}$$

$$\mathcal{C}_t := \{v \in \mathbb{R}^d : \|v - \hat{\mu}_t\|_{V_t} \leq \beta_t(\delta')\}. \tag{4.3.2}$$

Both $\mathcal{E}_t$ and $\mathcal{C}_t$ depend on δ', but we will often suppress notation for simplicity. The ellipsoid radius β_t is properly chosen as in [42] in order to guarantee that $\theta_\star \in \mathcal{E}_t$ and $\mu_\star \in \mathcal{C}_t$ with high probability.

Theorem 4.3.1 *Let Assumptions 6-7 hold. For $\delta \in (0,1)$, and $\beta_t(\delta) = R\sqrt{d\log\left(\frac{1+\frac{tL^2}{\lambda}}{\delta}\right)} + \sqrt{\lambda}S$, with probability at least $1-\delta$, it holds that $\theta_\star \in \mathcal{E}_t(\delta)$ and $\mu_\star \in \mathcal{C}_t(\delta)$, $\forall t \geq 1$.*

For Background on the frequentist view of LTS, please see Chapter 3-Section 3.5.

Addressing challenges in the safe setting

Compared to the classical linear bandit setting [44, 43], the presence of the safety constraint raises the following two questions: *(i)* How to guarantee actions played at each round are safe? *(ii)* In the face of the safety restrictions, how can optimism (cf.

(3.5.2)) be maintained? In the rest of this section, we explain the mechanisms that Safe-LTS employs to address both of these challenges.

Safety - First, the chosen action x_t at each round need not only maximize $x_t^\top \tilde{\theta}_t$, but also, it needs to be safe. Since the learner does not know the safe action set $\mathcal{D}_0^s$, Algorithm 4 performs conservatively and guarantees safety as follows. After creating the confidence region $\mathcal{C}_t$ around the RLS-estimate $\hat{\mu}_t$, it forms the so-called *safe decision set at round t* denoted as $\mathcal{D}_t^s$:

$$\mathcal{D}_t^s = \{x \in \mathcal{D}_0 : x^\top v \leq C, \forall v \in \mathcal{C}_t\}. \tag{4.3.3}$$

Then, the chosen action is optimized over only the subset $\mathcal{D}_t^s$, i.e.,

$$x_t = \arg\max_{x \in \mathcal{D}_t^s} x^\top \tilde{\theta}_t. \tag{4.3.4}$$

We make the following two remarks about $\mathcal{D}_t^s$. On a positive note, $\mathcal{D}_t^s$ is easy to compute:

$$\mathcal{D}_t^s := \{x \in \mathcal{D}_0 : x^\top v \leq C, \forall v \in \mathcal{C}_t\} \tag{4.3.5}$$

$$= \{x \in \mathcal{D}_0 : \max_{v \in \mathcal{C}_t} x^\top v \leq C\} \tag{4.3.6}$$

$$= \{x \in \mathcal{D}_0 : x^\top \hat{\mu}_t + \beta_t(\delta') \|x\|_{V_t^{-1}} \leq C\}. \tag{4.3.7}$$

Indeed, the optimization in (4.3.4) is an efficient convex quadratic program. Yet, the challenge remains that $\mathcal{D}_t^s$ contains actions that are safe with respect to *all* the parameters in $\mathcal{C}_t$, and not only $\mu_\star$. As such, it is only an *inner* approximation of the true safe set $\mathcal{D}_0^s$. As we will see next, this fact complicates the requirement for optimism.

Optimism in the face of safety - The fact that $\mathcal{D}_t^s$ is only an inner approximation of $\mathcal{D}_0^s$ makes it harder to maintain optimism of x_t as defined in (3.5.2). To see this,

note that in the classical setting, the algorithm of [43] would choose x_t as the action that maximizes $\tilde{\theta}_t$ over the *entire* set $\mathcal{D}_0$. In turn, this would imply that $x_t^\top \tilde{\theta}_t \geq x_\star^\top \tilde{\theta}_t$ because $x_\star$ belongs to the feasible set $\mathcal{D}_0$. This observation is the critical first argument in proving that x_t is optimistic often enough, i.e., (3.5.2) holds with fixed probability $p > 0$. Unfortunately, in the presence of safety constraints, x_t is a maximizer over only the subset $\mathcal{D}_t^s$. Since $x_\star$ may *not* lie within $\mathcal{D}_t^s$, there is no guarantee that $x_t^\top \tilde{\theta}_t \geq x_\star^\top \tilde{\theta}_t$ as before. So, how does then one guarantee optimism?

Intuitively, at the first rounds, the estimated safe set $\mathcal{D}_t^s$ is only a small subset of the true $\mathcal{D}_0^s$. Thus, $x_t \in \mathcal{D}_t^s$ is a vector of the small norm compared to that of $x_\star \in \mathcal{D}_0^s$. Thus, for (3.5.2) to hold, it must be that $\tilde{\theta}_t$ is not only in the direction of $\theta_\star$, but it also has a larger norm than that. To satisfy this latter requirement, the random vector η_t must be large; hence, it will "anti-concentrate more". As the algorithm progresses, and –thanks to side-information measurements– the set $\mathcal{D}_t^s$ becomes an increasingly better approximation of $\mathcal{D}_0^s$, the requirements on anti-concentration of η_t become the same as if no safety constraints were present. Overall, at least intuitively, we might hope that optimism is possible in the face of safety but only provided that η_t is set to satisfy a stronger (at least at the first rounds) anti-concentration property than that required by [43] in the classical setting.

At the heart of Algorithm 4 and its proof of regret lies an analytic argument that materializes the intuition described above. Specifically, we will prove that optimism is possible in the presence of safety at the cost of a stricter anti-concentration property compared to that specified in [43]. While the proof of this fact is deferred to Section 4.4.1, we now summarize the appropriate distributional properties that provably guarantee good regret performance of Algorithm 4 in the safe setting.

Definition 4.3.2 *In Algorithm 4, the random vector η_t is sampled IID at each t from a*

distribution $\mathcal{H}^{TS}$ *on* $\mathbb{R}^d$ *that is absolutely continuous with respect to the Lebesgue measure and satisfies:*

Anti-concentration: *There exists constant* $p > 0$ *such that for any* $u \in \mathbb{R}^d$ *with* $\|u\|_2 = 1$,

$$\mathbb{P}\big(u^\top \eta_t \geq 1 + \frac{2}{C} LS\big) \geq p. \tag{4.3.8}$$

Concentration: *There exists positive constants* $c, c' > 0$ *such that* $\forall \delta \in (0, 1)$,

$$\mathbb{P}\big(\|\eta_t\|_2 \leq \big(1 + \frac{2}{C} LS\big) \sqrt{cd \log\big(\frac{c'd}{\delta}\big)} \big) \geq 1 - \delta. \tag{4.3.9}$$

In particular, the difference to the distributional assumptions required by [43] in the classical setting is the extra term $\frac{2}{C} LS$ in (4.3.8) (naturally, the same term affects the concentration property (4.3.9)). Our proof of regret in Section 4.4 shows that this extra term captures an appropriate notion of the distance between the approximation $\mathcal{D}_t^s$ (where x_t lives) and the true safe set $\mathcal{D}_0^s$ (where $x_\star$ lives), and provides enough exploration for the sampled parameter $\tilde{\theta}_t$ so that actions in $\mathcal{D}_t^s$ can be optimistic. While this intuition can possibly explain the need for an additive term in Definition 4.3.2, it is insufficient when it comes to determining its "correct" value. This is determined by our analytic treatment in Section 4.4.1.

Finally, we remark that it is not hard to construct distributions that simultaneously satisfy the two conditions in (4.3.8) and (4.3.9). For example, a multivariate zero-mean IID Gaussian distribution with all entries having a (possibly time-dependent) variance $(1 + \frac{2}{C} LS)^2$ satisfies the Definition 4.3.2 and can be chosen to sample η_t in Algorithm 4 from it.

4.4 Regret Analysis

Here, we present a tight regret bound for Safe-LTS by proving that its action selection rule is simultaneously: 1) frequently optimistic, and, 2) guarantees a proper expansion of the estimated safe set. Our main result Theorem 4.4.1 is perhaps surprising: in spite of the additional safety constraints, Safe-LTS has regret $\mathcal{O}(d^{3/2}\log^{1/2} d \cdot T^{1/2}\log^{3/2} T)$ that is order-wise the same as that in the classical setting [44, 43].

Theorem 4.4.1 (Regret of Safe-LTS) *Let $\lambda \geq 1$ and Assumptions 6, 7, 8 hold. Fix $\delta \in (0,1)$. Then, with probability at least $1-\delta$, Safe-LTS is safe and its regret is upper bounded as follows:*

$$R(T) \leq \big(\beta_T(\delta') + \gamma_T(\delta')(1+\frac{4}{p})\big)\sqrt{2Td\log\big(1+\frac{TL^2}{\lambda}\big)} + \frac{4\gamma_T(\delta')}{p}\sqrt{\frac{8TL^2}{\lambda}\log\frac{4}{\delta}}, \tag{4.4.1}$$

where $\delta' = \frac{\delta}{6T}$, $\beta_t(\delta')$ as in Theorem 4.3.1 and, $\gamma_t(\delta') = \beta_t(\delta')\big(1+\frac{2}{C}LS\big)\sqrt{cd\log\big(\frac{c'd}{\delta'}\big)}$.

The theorem above provides guarantees both on the safety of the actions chosen by Safe-LTS Algorithm 4, as well as, on its regret.

First, we comment on the safety of the actions, which is ensured by the construction of the algorithm as discussed in Section 4.3.2. Formally, fix a desired δ and set $\delta' = \frac{\delta}{6T}$. Consider any time $t \in [T]$. On the one hand, from Theorem 4.3.1, it holds that $\mathbb{P}(\mu_\star \in \mathcal{C}_t(\delta')) \geq 1-\delta'$. On the other hand, by construction (lines 7-8, Algorithm 4), Safe-LTS guarantees that x_t at time t belongs to $\mathcal{D}_t^s$, i.e., $x_t^\top v \leq C, \forall v \in C_t(\delta')$. Putting these two together shows that $\mathbb{P}(x_t^\top \mu_\star \leq C) \geq 1-\delta'$. Then, a union bound (see Lemma B.1.5) over all time steps from 1 to T proves that $\mathbb{P}(\forall t \in [T] : x_t^\top \mu_\star \leq C) \geq 1 - T\delta' \geq 1 - \frac{\delta}{6}$, i.e., Safe-LTS is with high probability at least $1-\delta$ safe at all rounds.

Next, we discuss the regret bound of Theorem 4.4.1, which requires careful analysis. The detailed proof is in given in the Appendix B-Section B.2. In the rest of the section, we highlight the key changes compared to [44, 43] that occur due to the safety constraint. To begin, let us consider the following standard decomposition of the cumulative regret

$$R(T) = \sum_{t=1}^{T} \big(\underbrace{x_\star^\top \theta_\star - x_t^\top \tilde{\theta}_t}_{\text{Term I}} \big) + \sum_{t=1}^{T} \big(\underbrace{x_t^\top \tilde{\theta}_t - x_t^\top \theta_\star}_{\text{Term II}} \big). \tag{4.4.2}$$

Regarding Term II, the concentration property of $\mathcal{H}^{TS}$ guarantees that $\tilde{\theta}_t$ is close to $\hat{\theta}_t$, and consequently, close to $\theta_\star$ thanks to Theorem 4.3.1. Therefore, controlling Term II can be done similarly to previous works e.g., [42, 43]; see App. B.2.2 for more details. Next, we focus on Term I.

To see how the safety constraints affect the proofs let us first review the treatment of Term I in the classical setting. For UCB-type algorithms, Term I is always non-positive since the pair $(\tilde{\theta}_t, x_t)$ is optimistic at each round t by design [40, 48, 42]. For LTS, Term I can be positive; that is, (3.5.2) may not hold at every round t. However, [44, 43] proved that thanks to the anti-concentration property of η_t, this optimistic property occurs often enough.

Our main technical contribution, detailed in the next section, is to show that the properly modified anti-concentration property in Definition 4.3.2 together with the construction of approximated safe sets as in (4.3.7) can yield frequently optimistic actions even in the face of safety. Specifically, it is the extra term $\frac{2}{C}LS$ in (4.3.8) that allows enough exploration to the sampled parameter $\tilde{\theta}_t$ in order to compensate for safety limitations on the chosen actions, and because of that we are able to show Safe-LTS obtains the same order of regret as that of [43]. After that, in Section 4.4.2, we show that we can bound the overall regret of Term I with the V_τ norm of the optimistic actions.

As a closing remark, we note that our proof of optimism in the face of safety directly

applies as is above to a scenario where the constraint and the reward function are parameterized by the same vector $\theta_\star$, i.e., the constraint is of the form $x_t^\top \theta_\star \leq C$. In this case, obviously, no side information is required and we can show the same order of regret as in Theorem 4.4.1. Please see Section 4.5.3 for a discussion on how this result improves upon that of [70] who studied constraints parameterized by $\theta_\star$.

4.4.1 Proof sketch: Optimism despite safety constraints

We prove that $\tilde{\theta}_t$ is optimistic with constant probability (see Appendix B-Section B.1 for a formal statement and proof).

Lemma 4.4.2 *(Optimism in the face of safety; Informal) For any $t \geq 1$, Safe-LTS samples parameter $\tilde{\theta}_t$ and chooses action x_t such that the pair $(\tilde{\theta}_t, x_t)$ is optimistic frequently enough, i.e., $\mathbb{P}\left(x_t^\top \tilde{\theta}_t \geq x_\star^\top \theta_\star\right) \geq p$, where $p > 0$ is the probability of the anti-concentration property* (4.3.8).

The challenge in the proof is that x_t is chosen from $\mathcal{D}_t^s$, which does not necessarily contain all feasible actions and hence, *may not contain* $x_\star$. Thus, we need a mechanism to control the distance of $x_\star$ from the optimistic actions that can only lie within the subset $\mathcal{D}_t^s$ (*distance* is defined here in terms of an inner product with the optimistic parameters $\tilde{\theta}_t$). Unfortunately, we do not have direct control over this distance term and so at the heart of the proof lies the idea of identifying a "good" feasible action $\tilde{x}_t \in \mathcal{D}_t^s$ whose distance to $x_\star$ is easier to control. To be concrete, we show that it suffices to choose the good feasible point in the direction of $x_\star$, i.e., $\tilde{x}_t = \alpha_t x_\star$, where the key parameter $\alpha_t \in (0, 1]$ must be set to satisfy $\tilde{x}_t \in \mathcal{D}_t^s$. Naturally, the value of α_t is determined by the approximated safe set $\mathcal{D}_t^s$ as defined in (4.3.7). The challenge though is that we do not know how the value of $x_*^\top \hat{\mu}_t$ compares to the constant C. We circumvent this issue by introducing an enlarged confidence region centered at $\mu_\star$ as $\tilde{\mathcal{C}}_t := \{v \in \mathbb{R}^d : \|v - \mu_\star\|_{V_t} \leq 2\beta_t(\delta')\}$, and

the corresponding shrunk safe decision set as

$$\begin{aligned}\tilde{\mathcal{D}}_t^s &:= \{x \in \mathcal{D}_0 : x^\top v \le C, \forall v \in \tilde{\mathcal{C}}_t\} \\ &= \{x \in \mathcal{D}_0 : x^\top \mu_\star + 2\beta_t(\delta') \|x\|_{V_t^{-1}} \le C\} \subseteq \mathcal{D}_t^s. \end{aligned} \tag{4.4.3}$$

$\tilde{\mathcal{D}}_t^s$ is defined with respect to an ellipsoid centered at $\mu_\star$ (rather than at $\hat{\mu}_t$). This is convenient since $x_\star^\top \mu_\star \le C$. Using this, it can be easily checked that $\alpha_t = \big(1 + \frac{2}{C}\beta_t(\delta') \|x_\star\|_{V_t^{-1}}\big)^{-1}$ ensures $\alpha_t x_\star \in \tilde{\mathcal{D}}_t^s \subseteq \mathcal{D}_t^s$. From this, and optimality of $x_t = \arg\max_{x \in \mathcal{D}_t^s} x^\top \tilde{\theta}_t$ we have that

$$x_t^\top \tilde{\theta}_t \ge \alpha_t x_\star^\top \tilde{\theta}_t. \tag{4.4.4}$$

Using (4.4.4), it suffices to prove that $p \le \mathbb{P}\big(\alpha_t x_\star^\top \tilde{\theta}_t \ge x_\star^\top \theta_\star\big) = \mathbb{P}\big(x_\star^\top \tilde{\theta}_t \ge x_\star^\top \theta_\star + \frac{2}{C}\beta_t(\delta') \|x_\star\|_{V_t^{-1}} x_\star^\top \theta_\star\big)$, where, the equality follows by definition of α_t. To continue, recall that $\tilde{\theta}_t = \hat{\theta}_t + \beta_t V_t^{-\frac{1}{2}} \eta_t$. Thus, the probability we want to lower bound can be equivalently rewritten as

$$\mathbb{P}\big(\beta_t(\delta') x_\star^\top V_t^{-\frac{1}{2}} \eta_t \ge x_\star^\top(\theta_\star - \hat{\theta}_t) + \frac{2}{C}\beta_t(\delta') \|x_\star\|_{V_t^{-1}} x_\star^\top \theta_\star\big).$$

To simplify the above, we use (i) $|x_\star^\top \theta_\star| \le \|x_\star\|_2 \|\theta_\star\|_2 \le LS$; (ii) $x_\star^\top(\theta_\star - \hat{\theta}_t) \le \|x_\star\|_{V_t^{-1}} \|\theta_\star - \hat{\theta}_t\|_{V_t} \le \beta_t(\delta') \|x_\star\|_{V_t^{-1}}$, because of Cauchy-Schwartz and Theorem 4.3.1. Put together, we need that $p \le \mathbb{P}\big(\beta_t(\delta') x_\star V_t^{-\frac{1}{2}} \eta_t \ge \beta_t(\delta') \|x_\star\|_{V_t^{-1}} + \frac{2}{C} LS \beta_t(\delta') \|x_\star\|_{V_t^{-1}}\big)$, or equivalently,

$$p \le \mathbb{P}\big(u_t^\top \eta_t \ge 1 + (2/C)LS\big), \tag{4.4.5}$$

where we have defined $u_t = V_t^{-\frac{1}{2}} x_\star / \|x_\star\|_{V_t^{-1}}$. By definition of u_t, note that $\|u_t\|_2 = 1$. Hence, the desired (4.4.5) holds due to the anti-concentration property of the $\mathcal{H}^{\text{TS}}$

distribution in (4.3.8).

The key differences to the proof of optimism in the classical setting in [43, Lemma 3] are as follows. First, we present an algebraic version of the basic machinery introduced in [43, Sec. 5] that we show is convenient to extend to the safe setting. Second, we employ the idea of relating x_t to a "better" feasible point $\alpha_t x_\star$ and show optimism for the latter. Third, even after introducing α_t, the fact that $1/\alpha_t - 1$ is proportional to $\|x_\star\|_{V_t^{-1}}$ is critical for the seemingly simple algebraic steps that follow (4.4.4). In particular, in deducing (4.4.5) from the expression above, note that we have divided both sides in the probability term by $\|x_\star\|_{V_t-1}$. It is only thanks to the proportionality observation that we made above that the term $\|x_\star\|_{V_t-1}$ cancels throughout and we can conclude with (4.4.5) without a need to lower bound the minimum eigenvalue of the Gram matrix V_t (which is known to be hard).

4.4.2 Proof sketch: Why frequent optimism is enough to bound Term I

As discussed in Section 4.4, the presence of the safety constraints complicates the requirement for optimism. We show in Section 4.4.1 that Safe-LTS is optimistic with constant probability in spite of safety constraints. Based on this, we complete the sketch of the proof here by showing that we can bound the overall regret of Term I in (4.4.2) with the V_τ-norm of optimistic (and in our case, safe) actions. Let us first define the set of the optimistic parameters as

$$\Theta_t^{\text{opt}}(\delta') = \{\theta \in \mathbb{R}^d : \max_{x \in \mathcal{D}_t^s} x^\top \theta \geq x_\star^\top \theta_\star\}. \tag{4.4.6}$$

In Section 4.4.1, we show that Safe-LTS samples from this set i.e., $\tilde{\theta}_t \in \Theta_t^{\text{opt}}$, with constant probability. Note that, if at round t Safe-LTS samples from the set of optimistic parameters, Term I at that round is non-positive. In the following, we show that selecting the optimal arm corresponding to any optimistic parameter can control the overall regret of Term I. The argument below is adapted from [43] with required modifications.

For the purpose of this proof sketch, we assume that at each round t, the safe decision set contains the previous safe action that the algorithm played, i.e., $x_{t-1} \in \mathcal{D}_t^s$. However, for the formal proof in App. B.2.1, we do not need such an assumption. Let τ be a time such that $\tilde{\theta}_\tau \in \Theta_t^{\text{opt}}$, i.e., $x_\tau^\top \tilde{\theta}_\tau \geq x_\star^\top \theta_\star$. Then, for any $t \geq \tau$ we have

$$\begin{aligned} \text{Term I} := R_t^{\text{TS}} &= x_\star^\top \theta_\star - x_t^\top \tilde{\theta}_t \\ &\leq x_\tau^\top \tilde{\theta}_\tau - x_t^\top \tilde{\theta}_t \leq x_\tau^\top \left(\tilde{\theta}_\tau - \tilde{\theta}_t\right). \end{aligned} \tag{4.4.7}$$

The last inequality comes from the assumption that at each round t, the safe decision set contains the previous played safe actions for rounds $s \leq t$; hence, $x_\tau^\top \tilde{\theta}_t \leq x_t^\top \tilde{\theta}_t$. To continue from (4.4.7), we use Cauchy-Schwarz, and obtain

$$\begin{aligned} R_t^{\text{TS}} &\leq \|x_\tau\|_{V_\tau^{-1}} \left\|\tilde{\theta}_\tau - \tilde{\theta}_t\right\|_{V_\tau} \\ &\leq \left(\left\|\tilde{\theta}_\tau - \theta_\star\right\|_{V_\tau} + \left\|\theta_\star - \tilde{\theta}_t\right\|_{V_\tau}\right) \|x_\tau\|_{V_\tau^{-1}} \\ &\leq \left(\left\|\tilde{\theta}_\tau - \theta_\star\right\|_{V_\tau} + \left\|\theta_\star - \tilde{\theta}_t\right\|_{V_t}\right) \|x_\tau\|_{V_\tau^{-1}}. \end{aligned} \tag{4.4.8}$$

The last inequality comes from the fact that the Gram matrices construct a non-decreasing sequence ($V_\tau \preceq V_t$, $\forall t \geq \tau$). Then, we define the ellipsoid $\mathcal{E}_t^{\text{TS}}(\delta')$ such that

$$\mathcal{E}_t^{\text{TS}}(\delta') := \{\theta \in \mathbb{R}^d : \left\|\theta - \hat{\theta}_t\right\|_{V_t} \leq \gamma_t(\delta')\}, \tag{4.4.9}$$

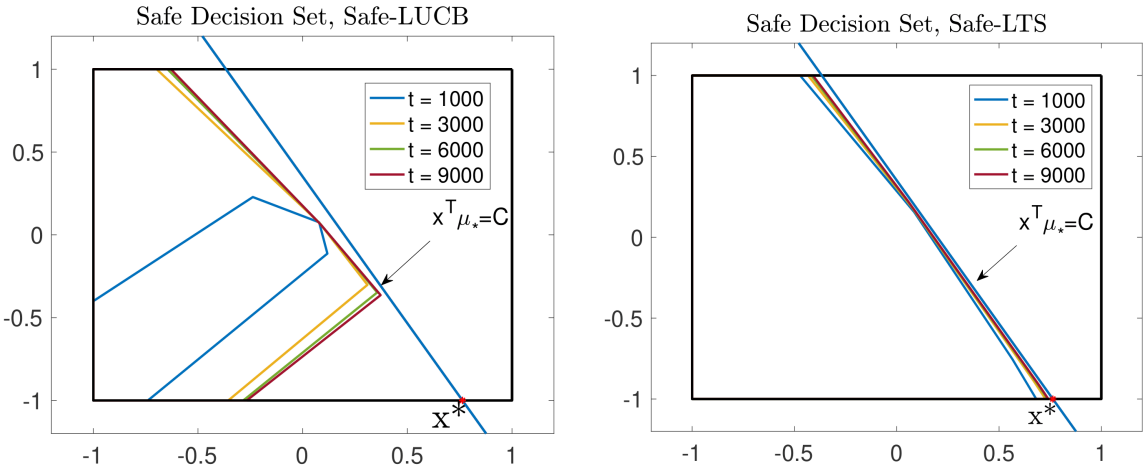

Figure 4.1: Comparison of expansion of a safe decision sets for Safe-LUCB and Safe-LTS, for a single problem instance.

where

$$\gamma_t(\delta') = \beta_t(\delta')\big(1 + \frac{2}{C}LS\big)\sqrt{cd\log\big(\frac{c'd}{\delta}\big)}. \tag{4.4.10}$$

It is not hard to see by combining Theorem 4.3.1 and the concentration property that $\tilde{\theta}_t \in \mathcal{E}_t^{\text{TS}}(\delta')$ with high probability. Hence, we can bound (4.4.8) using triangular inequality such that:

$$R_t^{\text{TS}} \leq \Big(\gamma_\tau(\delta') + \beta_\tau(\delta') + \gamma_t(\delta') + \beta_t(\delta')\Big)\,\|x_\tau\|_{V_\tau^{-1}} \tag{4.4.11}$$

$$\leq 2\Big(\gamma_T(\delta') + \beta_T(\delta')\Big)\,\|x_\tau\|_{V_\tau^{-1}} \tag{4.4.12}$$

The last inequality comes from the fact that $\beta_t(\delta')$ and $\gamma_t(\delta')$ are non-decreasing in t by construction. Therefore, following the intuition of [43], we can upper bound Term I with respect to the V_τ-norm of the optimal safe action at time τ (see Appendix B-Section B.2.1 for formal proof). Bounding the term $\|x_\tau\|_{V_\tau^{-1}}$ is standard based on the analysis provided in [42] (see Proposition B.1.1 in Appendix B-Section B.1).

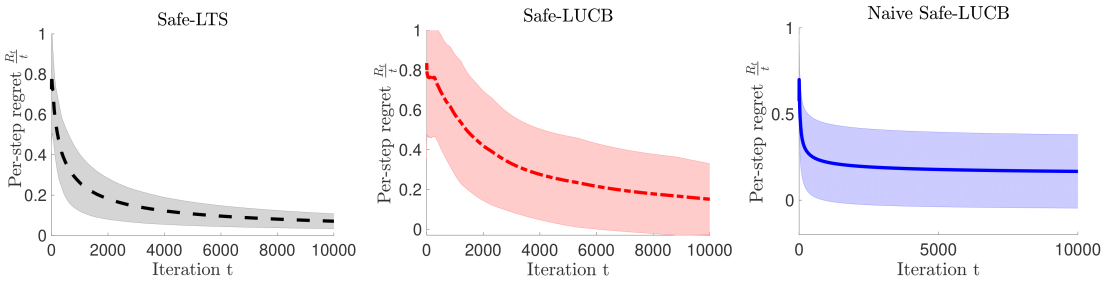

Figure 4.2: Comparison of mean per-step regret for Safe-LTS, Safe-LUCB, and Naive Safe-LUCB. The shaded regions show one standard deviation around the mean. The results are averages over 30 problem realizations.

4.5 Numerical Results and Comparison to State of the Art

We present details of our numerical experiments on synthetic data. First, we show how the presence of safety constraints affects the performance of LTS in terms of regret. Next, we evaluate Safe-LTS by comparing it against safe versions of LUCB. Then, we compare Safe-LTS to [70]'s Safe-LUCB. In all the implementations, we used: $T = 10000, \delta = 1/4T$, $R = 0.1$ and $\mathcal{D}_0 = [-1, 1]^2$. Unless otherwise specified, the reward and constraint parameters $\theta_\star$ and $\mu_\star$ are drawn from $\mathcal{N}(0, I_2)$ each; C is drawn uniformly from $[0, 1]$. Throughout, we have implemented a modified version of Safe-LUCB which uses ℓ_1-norms instead of ℓ_2-norms, due to computational considerations (e.g., [40, 70]). This highlights a well-known benefit associated with TS-based algorithms, namely that they are easier to implement and more computationally-efficient than UCB-based algorithms. In particular, the action selection rule in UCB-based algorithms involves solving optimization problems with bilinear objective functions, whereas, for TS-based algorithms, it would lead to linear objectives (see [43]).

4.5.1 The effect of safety constraints on LTS

In Fig. 4.3(left), we compare the average cumulative regret of Safe-LTS to the standard LTS with *oracle access* to the true safe set $\mathcal{D}_0^s$. The results are averages over 20 problem realizations. As shown, even though Safe-LTS requires that chosen actions belong to the conservative inner-approximation set $\mathcal{D}_t^s$, it still achieves a regret of the same order as the oracle reaffirming the prediction of Theorem 4.4.1. Also, the comparison to the oracle reveals that the action selection rule of Safe-LTS is indeed such that it guarantees fast safe-set expansion so as to not exclude optimistic actions for a long time. Fig. 4.3(left) also shows the performance Safe-LTS with dynamic noise distribution. In order for Safe-LTS to be frequently optimistic, our theory requires that the random perturbation η_t satisfies (4.3.8) *for all rounds*. Specifically, we need the extra $\frac{2}{C}LS$ factor compared to [43] in order to ensure safe set expansion. While this result is already sufficient for the tight regret guarantees of Theorem 4.4.1, it does not fully capture our intuition (see also Sec. 4.3.2) that as the algorithm progresses and $\mathcal{D}_t^s$ gets closer to $\mathcal{D}_0^s$, exploration (and thus, the requirement on anti-concentration) does not need to be so aggressive. Based on this intuition, we propose the following heuristic modification, in which Safe-LTS uses a perturbation with the following *dynamic* property:

$$\mathbb{P}_{\eta\sim\mathcal{H}^{\text{TS}}}\left(u^\top\eta \geq k(t)\right) \geq p, \tag{4.5.1}$$

for $k(t)$ a linearly-decreasing function $k(t) = (1+\frac{2}{C}LS)^2(1-t/T)$. In particular, this can be implemented by sampling each entry of $\eta_t, t \in [T]$ i.i.d from $\mathcal{N}(0, k(t))$. Fig. 4.3(left) shows empirical evidence of the superiority of the heuristic.

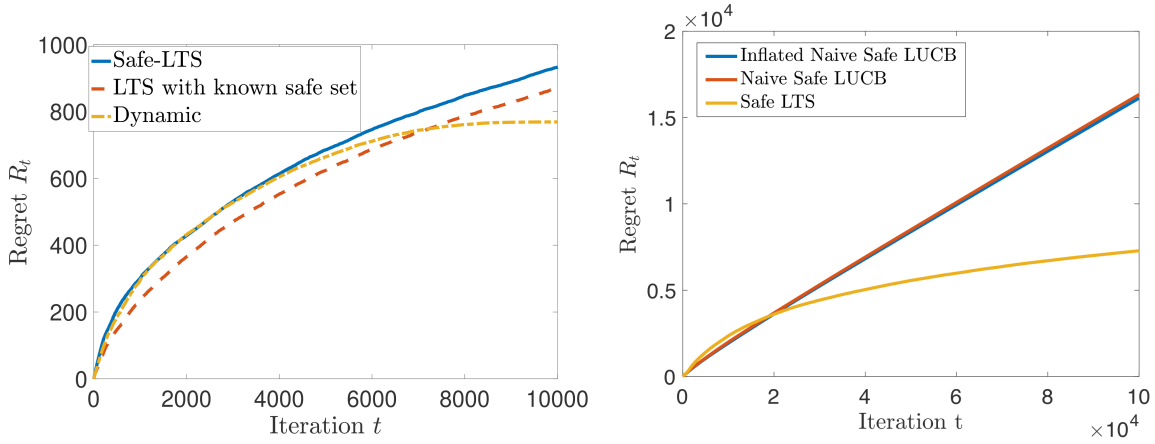

Figure 4.3: Left: Average cumulative regret of Safe-LTS vs standard LTS with oracle access to the safe set and Safe-LTS with a dynamic noise distribution described in Section 4.5.1. Right: Cumulative regret of Safe-LTS, Naive Safe-LUCB and Inflated Naive Safe-LUCB for a specific problem instance.

4.5.2 Comparison to the safe version of LUCB

Here, we compare the performance of our algorithm with the safe version of LUCB, as follows.

We implement a natural extension of the classical LUCB algorithm in [40], which we call "Naive Safe-LUCB" and which respects safety constraints by choosing actions from the estimated safe set in (4.3.3). We consider an improved version, which we call "Inflated Naive Safe-LUCB" and which is motivated by our analysis of Safe-LTS. Specifically, in light of Lemma 4.4.2, we implement the improved LUCB algorithm with an inflated confidence ellipsoid by a fraction $1+\frac{2}{C}LS$ in order to favor optimistic exploration. In Fig. 4.3(right), we employ these two algorithms for a specific problem instance showing that both fail to provide the $\widetilde{\mathcal{O}}(\sqrt{T})$ regret of Safe-LTS, in general. Specifically, we chooose $\theta_* = \begin{bmatrix} 0.5766 \\ -0.1899 \end{bmatrix}$, $\mu_* = \begin{bmatrix} 0.2138 \\ -0.0020 \end{bmatrix}$, and $C = 0.0615$. Further numerical simulations suggest that while Safe-LTS always outperforms Naive Safe-LUCB, the Inflated Naive Safe-LUCB can have superior performance to Safe-LTS in many problem instances (see

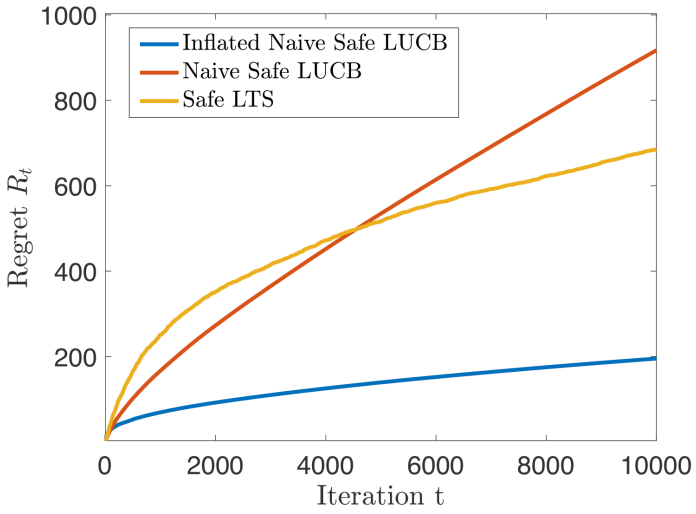

Figure 4.4: Comparison of the cumulative regret of Safe-LTS and Naive Safe-LUCB and Inflated Naive Safe-LUCB algorithms over randomly generated instances.

Fig. 4.4). Unfortunately, not only is this not always the case (cf. Fig. 4.3(right)), but also we are not aware of an appropriate modification to our proofs to show this problem-dependent performance. Further investigations in this direction might be of interest.

4.5.3 Comparison to Safe-LUCB

We compare our algorithm to the Safe-LUCB algorithm of [70]. In [70], the linear safety constraint involves the *same* unknown parameter vector $\theta_\star$ of the linear reward function and –in our notation– it takes the form $x^\top B\theta_\star \leq C$, for some *known* matrix B. As such, *no* side-information measurements are needed.

First, while our proof does not show a regret of $\widetilde{\mathcal{O}}(\sqrt{T})$ for the setting of [70] in the general case, it does so for special cases. For example, it is not hard to see that our proofs readily extend to their setting when $B = I$. This already improves upon the $\widetilde{\mathcal{O}}(T^{2/3})$ guarantee provided by [70]. Indeed, for $B = I$, there are non-trivial instances where $C - x_*^\top \theta_* = 0$ (i.e., the safety constraint is active), in which Safe-LUCB suffers

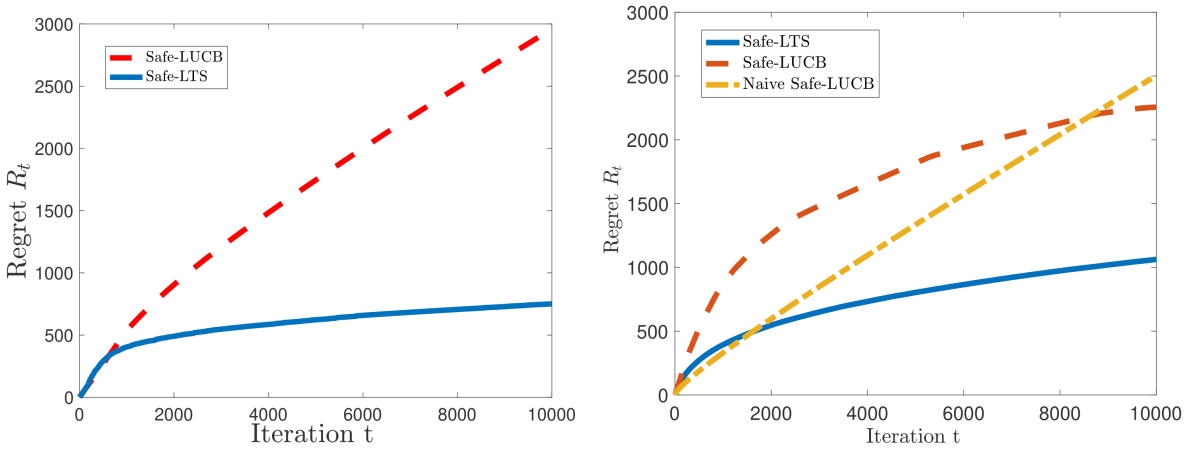

Figure 4.5: Left: Regret of Safe-LUCB vs Safe-LTS, for a single problem instance with active safety constraint. Right: Average cumulative regret of Safe-LTS vs two safe LUCB algorithms.

from a $\widetilde{\mathcal{O}}(T^{2/3})$ bound [70]. Second, while our proof adapts to a special case of [70]'s setting, the other way around is *not* true, i.e., it is not obvious how one would modify the proof of [70] to obtain a $\widetilde{\mathcal{O}}(\sqrt{T})$ guarantee even in the presence of side information. This point is highlighted by Fig. 4.5(left) that numerically compares the two algorithms for a specific problem instance with side information: $\theta_* = [0.9, 0.23]^\top$, $\mu_* = [0.55, 0.31]^T$, and $C = 0.11$ (note that the constraint is active at the optimal). Also, see Section 4.5.5 for a numerical comparison of the estimated safe-sets' expansion for the two algorithms. Fig. 4.5(right) compares Safe-LTS against Safe-LUCB and Naive Safe-LUCB over 30 problem realizations. As already pointed out in [70], Naive Safe-LUCB generally leads to poor regret, since the LUCB action selection rule alone does not provide sufficient exploration towards safe set expansion. In contrast, Safe-LUCB is equipped with a pure exploration phase over a given seed safe set, which is shown to lead to proper safe set expansion. Our paper reveals that the inherent randomized nature of Safe-LTS is alone capable to properly expand the safe set without the need for an explicit initialization phase (during which regret grows linearly).

4.5.4 Standard deviations

Figure 4.2 shows the sample standard deviation of regret around the average per-step regret for each one of the curves depicted in Figure 4.5(right). We remark on the strong dependency of the performance of LUCB-based algorithms on the specific problem instance, whereas the performance of Safe-LTS does not vary significantly under the same instances.

4.5.5 Safe-set expansion

We also plot the expansion of the estimated safe set $\mathcal{D}_t^s$ in time for different problem instances for Saf-LTS and "Inflated Naieve Safe-LUCB" and Safe-LUCB in [70]. In particular, Fig. 4.1 highlights the gradual expansion of the safe decision set for Safe-LUCB in [70] and Safe-LTS for a problem instance in which the safety constraint is active for parameters $\theta_* = \begin{bmatrix} 0.9 \\ 0.23 \end{bmatrix}$, $\mu_* = \begin{bmatrix} 0.55 \\ -0.31 \end{bmatrix}$, and $C = 0.11$. Similarly, Fig. 4.6 illustrates the expansion of the safe decision set for "Inflated Naive Safe-LUCB" and Safe-LTS for a problem instance with parameters $\theta_* = \begin{bmatrix} 0.5766 \\ -0.1899 \end{bmatrix}$, $\mu_* = \begin{bmatrix} 0.2138 \\ -0.0020 \end{bmatrix}$, and $C = 0.0615$ in which the former provides poor (almost linear) regret. These empirical experiments reinforce the main message of our paper that the inherent randomized nature of TS is crucial for properly expanding the safe action set.

Next, we comment on the dependence of the regret of Safe-LTS on the size of the safe set. Note that the size of the safe action set depends on the safety constant C as well as on the unknown parameter $\mu_\star$. Recall that S is an upper bound on the norm of $\mu_\star$, and, also $\|x\|_2 \leq L$ for any action vector $x \in \mathcal{D}_0$. Since the constraint is of the form $x^\top \mu_\star \leq C$, the size of the set of safe actions depends on the values L, S, C. We will also assume that $LS > C$, since otherwise, it follows by Cauchy-Schwartz that all

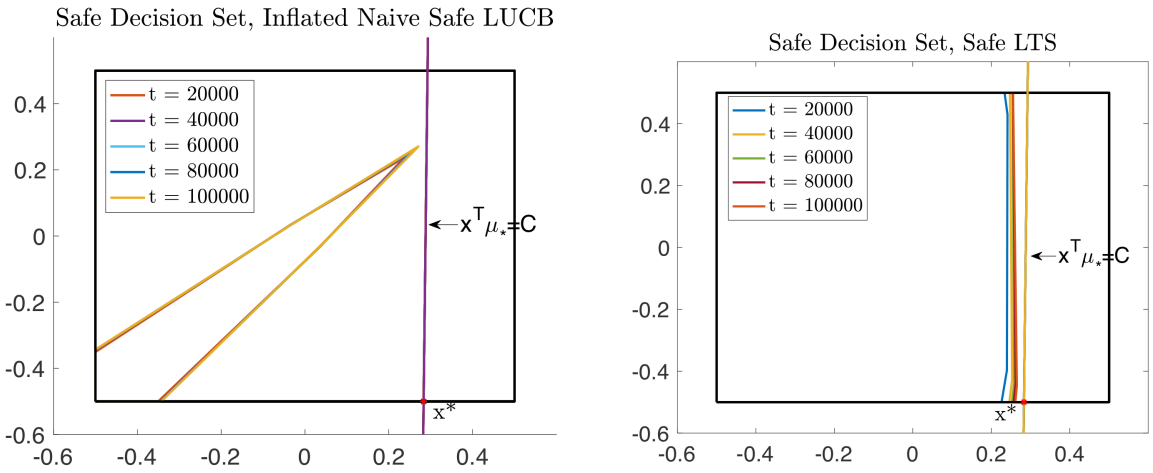

Figure 4.6: Comparison of expansion of safe decision sets for Safe-LTS, and Inflated Naive Safe-LUCB.

actions in $\mathcal{D}_0$ are safe and the regret is no different compared to the unconstrained case. Intuitively, for smaller values of C (compared to LS), the "smaller" the safe set around zero. This means that the algorithm can only take actions in a very conservative manner to guarantee that actions remain safe. At an intuitive level, we would then expect an increase on regret. This intuition is in fact captured by our regret bound in Theorem 4.4.1 showing that the bound increases with increasing values of the ratio $\frac{LS}{C}$. Thus, the smaller C, the larger our regret bound. In Figure 4.7 we showcase the effect of decreasing C on regret. Specifically, we have chosen $\theta_\star = [0.3; 0.8]$, $\mu_\star = [0.2; 0.7]$, $\mathcal{D}_0 = [-1, 1]^2$, $S = \sqrt{2}$ and $L = \sqrt{2}$ and we have plotted the regret of Safe-LTS for different values of $C = 0.7, 0.8, 0.9$ and 1. We see that the regret increases for smaller values of C as suggested by our bound of Theorem 4.4.1. As a closing remark, while we make no claim that our bound captures sharply the effect of the size of the safe set (perhaps measured in terms of some geometric quantity such as volume), we showed that our bound captures the effect of the size in the summary term LS/C, which also appears to agree with the empirical results of Figure 4.7.

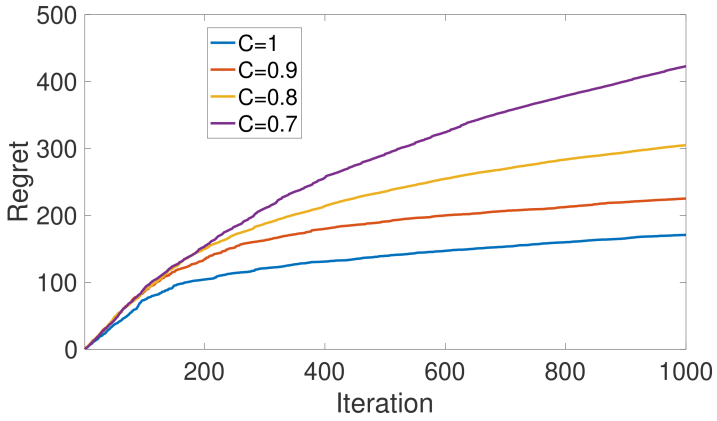

Figure 4.7: Comparison of the cumulative regret of Safe-LTS for different values of the safety constant C.

4.6 Conclusion

In this paper, we study a linear stochastic bandit (LB) problem in which the environment is subject to unknown linear safety constraints that need to be satisfied at each round. As such, the learner must make necessary modifications to ensure that the chosen actions belong to the unknown safe set. We propose Safe-LTS, which to the best of our knowledge, is the first safe linear TS algorithm with provable regret guarantees for this problem. Moreover, we show that the Safe-LTS achieves the same frequentist regret of order $\mathcal{O}(d^{3/2}\log^{1/2} d \cdot T^{1/2}\log^{3/2} T)$ as the original LTS problem studied in [43]. We also compare Safe-LTS with several UCB-type safe algorithms. We show that our algorithm has: better regret in the worst-case ($\widetilde{\mathcal{O}}(T^{1/2})$ vs. $\widetilde{\mathcal{O}}(T^{2/3})$), fewer parameters to tune and often superior empirical performance. Interesting directions for future work include gaining a theoretical understanding of the regret of the algorithm when the TS distribution satisfies the dynamic property in (4.5.1), which empirically leads to regret of smaller order as well as, investigating TS-based alternatives to the GP-UCB-type algorithms of [66, 67]. Additionally, it is interesting to study extensions of our theory on linear constraints to the more general setting in which constraints are modeled as Gaus-

sian Processes. This would also allow more complex settings in which the safe regions may even be disconnected. [85, 86, 87, 88, 89] are the results of this chapter.

Chapter 5

Stage-wise Conservative Stochastic Linear Bandits

5.1 Introduction

Machine Learning algorithms have found an increasingly widespread deployment in healthcare [90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 90], communications [87, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111], robotics [112, 113, 114, 115, 116], quantum research [117, 118, 119], etc. With the growing range of applications of bandit algorithms for safety-critical real-world systems, the demand for safe learning is receiving increasing attention [120]. In this paper, we investigate the effect of stage-wise safety constraints on the linear stochastic bandit problem. Inspired by the earlier work of [46, 121], the type of safety constraint we consider in this paper was first introduced by [71]. As with the classic linear stochastic bandit problem, the learner wishes to choose a sequence of actions x_t that maximize the expected reward over the horizon. However, here the learner is also given a baseline policy that suggests an action with a guaranteed level of expected reward at each stage of the algorithm. This could be based on historical data, e.g., historical ad

placement or medical treatment policies with known success rates. The safety constraint imposed on the learner requires her to ensure that the expected reward of her chosen action at every single round is no less than a predetermined fraction of the expected reward of the action suggested by the baseline policy. An example that might benefit from the design of stage-wise conservative learning algorithms arises in recommender systems, where the recommender might wish to avoid recommendations that are extremely disliked by the users at any single round. Our proposed stage-wise conservative constraints ensure that at no round would the recommendation system cause severe dissatisfaction for the user, and the reward of action employed by the learning algorithm, if not better, should be close to that of baseline policy. Another example is in clinical trials where the effects of different therapies on patients' health are initially unknown. We can consider the baseline policy to be treatments that have been historically employed, with known effectiveness. The proposed stage-wise conservative constraint guarantees that at each stage, the learning algorithm suggests an action (a therapy) that achieves the expected reward close to that of the baseline treatment, and as such, this experimentation does not cause harm to *any single patient's health*. To tackle this problem, [71] proposed a greedy algorithm called SEGE. They use the decomposition of the regret first proposed in [46], and show an upper bound of order $\mathcal{O}(\sqrt{T})$ over the number of times that the learning algorithm plays the baseline actions, overall resulting in an expected regret of $\mathcal{O}(\sqrt{T}\log T)$. For this problem, we present two algorithms, SCLTS and SCLUCB, and we provide regret bounds of order $\mathcal{O}(\sqrt{T}\log^{3/2} T)$ and $\mathcal{O}(\sqrt{T}\log T)$, respectively. As it is explained in detail in Section 5.4, we improve the result of [71], i.e., we show our proposed algorithms play the (non-optimal) baseline actions at most $\mathcal{O}(\log T)$ times, while also relaxing a number of assumptions made in [71]. Moreover, we show that our proposed algorithms are adaptable with minor modifications to other safety-constrained variations of this problem. This includes the case where the constraint has a different unknown

parameter than the reward function with bandit feedback (Section 5.4.1), as well as the setting where the reward of baseline action is unknown to the learner in advance (Section 5.5).

5.2 Problem Setting

Linear Bandit. The learner is given a convex and compact set of actions $\mathcal{X} \subset \mathbb{R}^d$. At each round t, she chooses an action x_t and observes a random reward

$$y_t = \langle x_t, \theta_\star \rangle + \xi_t, \tag{5.2.1}$$

where $\theta_\star \in \mathbb{R}^d$ is *unknown* but fixed reward parameter and ξ_t is zero-mean additive noise. We let r_t be the expected reward of action x_t at round t, i.e., $r_t := \mathbb{E}[y_t] = \langle x_t, \theta_\star \rangle$.

Baseline actions and stage-wise constraint. We assume that the learner is given a baseline policy such that selecting the baseline action x_{b_t} at round t, she would receive an expected reward $r_{b_t} := \langle x_{b_t}, \theta_\star \rangle$. We assume that the learner knows the expected reward of the actions chosen by the baseline policy. We further assume that the learner's action selection rule is subject to a stage-wise conservative constraint of the form[1]

$$r_t = \langle x_t, \theta_\star \rangle \geq (1 - \alpha) r_{b_t}, \tag{5.2.2}$$

that needs to be satisfied at each round t. In particular, constraint (5.2.2) guarantees that at each round t, the expected reward of the action chosen by the learner stays above the predefined fraction $1-\alpha \in (0, 1)$ of the baseline policy. The parameter α, controlling the conservatism level of the learning process, is assumed known to the learner similar

[1] In Section 5.4.1, we show that our results also extend to constraints of the form $\langle x_t, \mu_\star \rangle \geq (1-\alpha) q_{b_t}$, where $\mu_\star$ is an additional unknown parameter. In this case, we assume the learner receives additional bandit feedback on the constraint after each round.

to [46, 121]. At each round t, an action is called *safe* if its expected reward is above the predetermined fraction of the baseline policy, i.e., $(1-\alpha)r_{b_t}$.

Remark 5.2.1 *It is reasonable to assume that the learner has an accurate estimate of the expected reward of the actions chosen by baseline policy [46]. However, in Section 5.5, we relax this assumption and propose an algorithm for the case where the expected rewards of the actions chosen by baseline policy are unknown to the learner in advance.*

Regret. The *cumulative pseudo-regret* of the learner up to round T is defined as $R(T) = \sum_{t=1}^{T}\langle x_\star, \theta_\star\rangle - \langle x_t, \theta_\star\rangle$,

$$R(T) = \sum_{t=1}^{T}\langle x_\star, \theta_\star\rangle - \langle x_t, \theta_\star\rangle, \tag{5.2.3}$$

where $x_\star$ is the optimal safe action that maximizes the expected reward,

$$x_\star = \arg\max_{x\in\mathcal{X}}\langle x, \theta_\star\rangle. \tag{5.2.4}$$

The learner's objective is to minimize the pseudo-regret while respecting the stage-wise conservative constraint in (5.2.2). For the rest of the paper, we use regret to refer to the pseudo-regret $R(T)$.

5.2.1 Previous work

The baseline model adopted in this paper was first proposed in [46, 121] in the case of *cumulative constraints* on the reward. In [46, 121], an action is considered feasible/safe at round t as long as it keeps the cumulative reward up to round t above a given fraction of a given baseline policy. This differs from our setting, which is focused on stage-wise constraints, where we want the expected reward of the *every single action* to exceed a given fraction of the baseline reward at each time t. This is a tighter constraint

than that of [46, 121]. The setting considered in this paper was first studied in [71], which proposed an algorithm called SEGE to guarantee the satisfaction of the safety constraint at each stage of the algorithm. While our paper is motivated by [71], there are a few key differences: 1) We prove an upper bound of order $\mathcal{O}(\log T)$ for the number of times that the learning algorithm plays the conservative actions which is an order-wise improvement with respect to that of [71], which shows an upper bound of order $\mathcal{O}(\sqrt{T})$; 2) In our setting, the action set is assumed to be a general convex and compact set in $\mathbb{R}^d$. However, in [71], the proof relies on the action set being a specific ellipsoid; 3) In Section 5.5, we provide a regret guarantee for the learning algorithm for the case where the baseline reward is unknown. However, the results of [71] have not been extended to this case; 4) In Section 5.4.1, we also modify our proposed algorithm and provide a regret guarantee for the case where the constraint has a different unknown parameter than the one in the reward function. However, this is not discussed in [71]. Another difference between the two works is on the type of performance guarantees. In [71], the authors bound the *expected* regret. Towards this goal, they manage to quantify the effect of the risk level δ on regret and constraint satisfaction. However, it appears that the analysis in [71] is limited to ellipsoidal action sets. Instead, in this paper, we present a bound on the regret that holds with high (constant) probability (parameterized by δ) over *all* T rounds of the algorithm. This type of result is very common in the bandit literature, e.g. [42, 40], and in the emerging safe-bandit literature [46, 70, 67].

Another variant of safety w.r.t a baseline policy has also been studied in [122, 123] in the multi-armed bandits framework. Moreover, there has been increasing attention on studying the effect of safety constraints in the Gaussian process (GP) optimization literature. For example, [66, 67] study the problem of *nonlinear* bandit optimization with nonlinear constraints using GPs (as non-parametric models). The algorithms in [66, 67] come with convergence guarantees but no regret bound. Moreover, [68, 69] study

safety-constrained optimization using GPs in robotics applications. A large body of work has considered safety in the context of model-predictive control, see, e.g., [64, 65] and references therein. Focusing specifically on linear stochastic bandits, an extension of UCB-type algorithms to provide safety guarantees with provable regret bounds was considered recently in [70]. This work considers the effect of a linear constraint of the form $x^\top B\theta_\star \leq C$, where B and C are respectively a known matrix and positive constant and provides a problem-dependent regret bound for a safety-constrained version of LUCB that depends on the location of the optimal action in the safe action set. Notice that this setting requires the linear function $x^\top B\theta_\star$ to remain below a threshold C, as opposed to our setting which considers a lower bound on the reward. We note that the algorithm and proof technique in [70] does not extend to our setting and would only work for inequalities of the given form; however, we discuss how our algorithm can be modified to provide a regret bound of order $\mathcal{O}(\sqrt{T}\log T)$ for the setting of [70] in Appendix C-Section C.8. A TS variant of this setting has been studied in [124, 125].

5.2.2 Model Assumptions

Notation. The weighted ℓ_2-norm with respect to a positive semi-definite matrix V is denoted by $\|x\|_V = \sqrt{x^\top V x}$. The minimum of two numbers a, b is denoted $a \wedge b$. Let $\mathcal{F}_t = (\mathcal{F}_1, \sigma(x_1, \xi_1, \ldots, x_t, \xi_t))$ be the filtration (σ-algebra) that represents the information up to round t.

Assumption 9 *For all t, ξ_t is conditionally zero-mean R-sub-Gaussian noise variables, i.e., $\mathbb{E}[\xi_t|\mathcal{F}_{t-1}] = 0$, and $\mathbb{E}[e^{\lambda\xi_t}|\mathcal{F}_{t-1}] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right), \forall\lambda \in \mathbb{R}^d$.*

Assumption 10 *There exists a positive constant S such that $\|\theta_\star\|_2 \leq S$.*

Assumption 11 *The action set $\mathcal{X}$ is a compact and convex subset of $\mathbb{R}^d$ that contains the unit ball. We assume that $\|x\|_2 \leq L, \forall x \in \mathcal{X}$. Also, we assume $\langle x, \theta_\star\rangle \leq 1, \forall x \in \mathcal{X}$.*

Let $\kappa_{b_t} = \langle x_\star, \theta_\star \rangle - r_{b_t}$ be the difference between expected reward of the optimal and baseline actions at round t. As in [46], we assume the following.

Assumption 12 *There exist* $0 \leq \kappa_l \leq \kappa_h$ *and* $0 < r_l \leq r_h$ *such that, at each round* t

$$\kappa_l \leq \kappa_{b_t} \leq \kappa_h \text{ and } r_l \leq r_{b_t} \leq r_h. \tag{5.2.5}$$

We note that since these parameters are associated with the baseline policy, it can be reasonably assumed that they can be estimated accurately from data. This is because we think of the baseline policy as a "past strategy", implemented before bandit optimization, thus producing a large amount of data. The lower bound $0 < r_l \leq r_{b_t}$ on the baseline reward ensures a minimum level of performance at each round. κ_h and r_h could be at most 1, due to Assumption 11. For simplicity, we assume the lower bound κ_l on the sub-optimality gap κ_{b_t} is known. If not, we can always choose $\kappa_l = 0$ by optimality of $x_\star$.

5.3 Stage-wise Conservative Linear Thompson Sampling (SCLTS) Algorithm

In this section, we propose a TS variant algorithm in a frequentist setting referred to as *Stage-wise Conservative Linear Thompson Sampling* (SCLTS) for the problem setting in Section 5.2. Our adoption of TS is due to its well-known computational efficiency over UCB-based algorithms, since action selection via the latter involves solving optimization problems with bilinear objective functions, whereas the former would lead to linear objectives. However, this choice does not fundamentally affect our approach. In fact, in Appendix C-Section C.7, we propose a Stage-wise Conservative Linear UCB (SCLUCB) algorithm, and we provide the regret guarantee for it. In particular, we show a regret

of order $\mathcal{O}\left(d\sqrt{T}\log(\frac{TL^2}{\lambda\delta})\right)$ for SCLUCB, which has the same order as the lower bound proposed for LB in [40, 48].

At each round t, given a regularized least-square (RLS) estimate of $\hat{\theta}_t$, SCLTS samples a perturbed parameter $\tilde{\theta}_t$ with an appropriate distributional property. Then, it searches for the action that maximizes the expected reward considering the parameter $\tilde{\theta}_t$ as the true parameter while respecting the safety constraint (5.2.2). If any such action exists, it is played under certain conditions; else, the algorithm resorts to playing a perturbed version of the baseline action that satisfies the safety constraint. In order to guarantee constraint satisfaction (a.k.a safety of actions), the algorithm builds a confidence region $\mathcal{E}_t$ that contains the unknown parameter $\theta_\star$ with high probability. Then, it constructs an *estimated safe* set $\mathcal{X}_t^s$ such that all actions $x_t \in \mathcal{X}_t^s$ satisfy the safety constraint for all $v \in \mathcal{E}_t$. The summary of the SCLTS is presented in Algorithm 5, and a detailed explanation follows.

Algorithm 5: Stage-wise Conservative Linear Thompson Sampling (SCLTS)

Input: $\delta, T, \lambda, \rho_1$
Set $\delta' = \frac{\delta}{4T}$
for $t = 1, \dots, T$ **do**
  Sample $\eta_t \sim \mathcal{H}^{\text{TS}}$
  Compute RLS-estimate $\hat{\theta}_t$ and V_t according to (5.3.1)
  Set $\tilde{\theta}_t = \hat{\theta}_t + \beta_t V_t^{-1/2}\eta_t$
  Build the confidence region $\mathcal{E}_t(\delta')$ in (5.3.2)
  Compute the estimated safe set $\mathcal{X}_t^s$ in (5.3.3)
  if the following optimization is feasible: $x(\tilde{\theta}_t) = \text{argmax}_{x\in\mathcal{X}_t^s}\langle x, \tilde{\theta}_t\rangle$, **then**
  Set $F = 1$, **else** $F = 0$
  if $F = 1$ **and** $\lambda_{\min}(V_t) \geq \left(\frac{2L\beta_t}{\kappa_l + \alpha r_{b_l}}\right)^2$, **then** Play $x_t = x(\tilde{\theta}_t)$
  else Play $x_t = (1-\rho_1)x_{b_t} + \rho_1\zeta_t$
  Observe reward y_t in (5.2.1)
end for

5.3.1 Algorithm description

Let $x_1, \dots, x_t$ be the sequence of the actions and $r_1, \dots, r_t$ be their corresponding rewards. For any $\lambda > 0$, we can obtain a regularized least-squares (RLS) estimate $\hat{\theta}_t$ of $\theta_\star$ as follows

$$\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t-1} y_s x_s, \text{ where } V_t = \lambda I + \sum_{s=1}^{t-1} x_s x_s^\top. \tag{5.3.1}$$

Algorithm 5 construct a confidence region

$$\mathcal{E}_t(\delta') = \mathcal{E}_t := \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|_{V_t} \leq \beta_t(\delta')\}, \tag{5.3.2}$$

where the ellipsoid radius β_t is chosen according to the Proposition 5.3.1 in [42] (restated below for completeness) in order to guarantee that $\theta_\star \in \mathcal{E}_t$ with high probability.

Proposition 5.3.1 *Let Assumptions 9, 10, and 11 hold. For a fixed $\delta \in (0, 1)$, and*

$$\beta_t(\delta) = R\sqrt{d \log\left(\frac{1 + \frac{tL^2}{\lambda}}{\delta}\right)} + \sqrt{\lambda} S$$

with probability at least $1 - \delta$, it holds that $\theta_\star \in \mathcal{E}_t$.

The estimated safe action set

Since $\theta_\star$ is unknown to the learner, she does not know whether an action $x \in \mathcal{X}$ is safe or not. Thus, she builds an estimated safe set such that each action $x_t \in \mathcal{X}_t^s$ satisfies

the safety constraint for all $v \in \mathcal{E}_t$, i.e.,

$$\mathcal{X}_t^s := \{x \in \mathcal{X} : \langle x, v\rangle \geq (1-\alpha) r_{b_t}, \forall v \in \mathcal{E}_t\} = \{x \in \mathcal{X} : \min_{v \in \mathcal{E}_t} \langle x, v\rangle \geq (1-\alpha) r_{b_t}\} \quad (5.3.3)$$

$$= \{x \in \mathcal{X} : \langle x, \hat{\theta}_t\rangle - \beta_t(\delta') \|x\|_{V_t^{-1}} \geq (1-\alpha) r_{b_t}\}. \quad (5.3.4)$$

Note that $\mathcal{X}_t^s$ is easy to compute since (5.3.4) involves a convex quadratic program. In order to guarantee safety, at each round t, the learner chooses her actions only from this estimated safe set in order to maximize the reward given the sampled parameter $\tilde{\theta}_t$, i.e.,

$$x(\tilde{\theta}_t) = \arg\max_{x \in \mathcal{X}_t^s} \langle x, \tilde{\theta}_t\rangle, \quad (5.3.5)$$

where $\tilde{\theta}_t = \hat{\theta}_t + \beta_t V_t^{-1/2} \eta_t$, and η_t is a random IID sample from a distribution $\mathcal{H}^{\text{TS}}$ that satisfies certain distributional properties (see [43] or Defnition C.3.1 in Appendix C-Section C.3 for more details). The challenge with $\mathcal{X}_t^s$ is that it contains actions that are safe with respect to all the parameters in $\mathcal{E}_t$, and not only $\theta_\star$. Hence, there may exist some rounds that $\mathcal{X}_t^s$ is empty. In order to face this problem, the algorithm proceeds as follows. At round t, if the estimated action set $\mathcal{X}_t^s$ is not empty, SCLTS plays the safe action $x(\tilde{\theta}_t)$ in (5.3.5) only if the minimum eigenvalue of the Gram matrix V_t is greater than $k_t^1 = \left(\frac{2L\beta_t}{\kappa_l + \alpha r_{b_l}}\right)^2$, i.e., $\lambda_{\min}(V_t) \geq k_t^1$, where k_t^1 is of order $\mathcal{O}(\log t)$. Otherwise, it plays the conservative action that is presented next. We show in Appendix C-Section C.3 that $\lambda_{\min}(V_t) \geq k_t^1$ ensures that for the rounds that SCLTS plays the action $x(\tilde{\theta}_t)$ in (5.3.5), the optimal action $x_\star$ belongs to the estimated safe set $\mathcal{X}_t^s$, from which we can bound the regret of Term I in (5.4.1).

Conservative actions

In our setting, we assume that the learner is given a baseline policy that at each round t suggests a baseline action x_{b_t}. We employ the idea proposed in [71], which is merging the baseline actions with random exploration actions under stage-wise safety constraints. In particular, at each round t, SCLTS constructs a conservative action x_t^{cb} as a convex combination of the baseline action x_{b_t} and a random vector ζ_t as follows:

$$x_t^{\text{cb}} = (1 - \rho_1)x_{b_t} + \rho_1\zeta_t, \tag{5.3.6}$$

where ζ_t is assumed to be a sequence of independent, zero-mean and bounded random vectors. Moreover, we assume that $\|\zeta_t\|_2 = 1$ almost surely and $\sigma_\zeta^2 = \lambda_{\min}(\text{Cov}(\zeta_t)) > 0$. The parameters σ_ζ and ρ_1 control the exploration level of the conservative actions. In order to ensure that the conservative actions are safe, in Lemma 5.3.2, we establish an upper bound on ρ_1 such that for all $\rho_1 \in (0, \bar{\rho})$, the conservative action $x_t^{\text{cb}} = (1-\rho_1)x_{b_t} + \rho_1\zeta_t$ is guaranteed to be safe.

Lemma 5.3.2 *At each round t, given the fraction α, for any $\rho \in (0, \bar{\rho})$, where $\bar{\rho} = \frac{\alpha r_l}{S+r_h}$, the conservative action $x_t^{cb} = (1 - \rho)x_{b_t} + \rho\zeta_t$ is guaranteed to be safe almost surely.*

For the ease of notation, in the rest of this paper, we simply assume that $\rho_1 = \frac{r_l}{S+r_h}\alpha$.

At round t, SCLTS plays the conservative action x_t^{cb} if the two conditions defined in Section 5.3.1 do not hold, i.e., either the estimated safe set $\mathcal{X}_t^s$ is empty or $\lambda_{\min}(V_t) < k_t^1$.

5.4 Regret Analysis

In this section, we provide a tight regret bound for SCLTS. In Proposition 5.4.1, we show that the regret of SCLTS can be decomposed into regret caused by choosing safe

Thompson Sampling actions plus that of playing conservative actions. Then, we bound both terms separately. Let N_{t-1} be the set of rounds $i < t$ at which SCLTS plays the action in (5.3.5). Similarly, $N_{t-1}^c = \{1, \dots, t-1\} - N_{t-1}$ is the set of rounds $j < t$ at which SCLTS plays the conservative actions.

Proposition 5.4.1 *The regret of SCLTS can be decomposed into two terms as follows:*

$$R(T) \leq \underbrace{\sum_{t \in N_T} \left(\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle \right)}_{\textit{Term I}} + \underbrace{|N_T^c| \left(\kappa_h + \rho_1 (r_h + S) \right)}_{\textit{Term II}} \tag{5.4.1}$$

The idea of bounding Term I is inspired by [43]: we wish to show that LTS has a constant probability of being "optimistic", in spite of the need to be conservative. In Theorem 5.4.2, we provide an upper bound on the regret of Term I which is of order $\mathcal{O}(d^{3/2} \log^{1/2} d \; T^{1/2} \log^{3/2} T)$.

Theorem 5.4.2 *Let $\lambda, L \geq 1$. On event $\{\theta_\star \in \mathcal{E}_t, \forall t \in [T]\}$, and under Assumption 12, we can bound Term I in (5.4.1) as:*

$$\textit{Term I} \leq (\beta_T(\delta') + \gamma_T(\delta')(1 + \frac{4}{p})) \sqrt{2Td \log \left(1 + \frac{TL^2}{\lambda}\right)} + \frac{4\gamma_T(\delta')}{p} \sqrt{\frac{8TL^2}{\lambda} \log \frac{4}{\delta}},$$

where $\delta' = \frac{\delta}{6T}$, and $\gamma_t(\delta) = \beta_t(\delta') \left(1 + \frac{2}{C}\right) \sqrt{cd \log \left(\frac{c'd}{\delta}\right)}$

We note that the regret of Term I has the same bound as that of [43] in spite of the additional safety constraints imposed on the problem. As the next step, in order to bound Term II in (5.4.1), we need to find an upper bound on the number of times $|N_T^c|$ that SCLTS plays the conservative actions up to time T. We prove an upper bound on $|N_T^c|$ in Theorem 5.4.3.

Theorem 5.4.3 *Let $\lambda, L \geq 1$. On event $\{\theta_\star \in \mathcal{E}_t, \forall t \in [T]\}$, and under Assumption 12, it holds that*

$$|N_T^c| \leq \left(\frac{2L\beta_T}{\rho_1\sigma_\zeta(\kappa_l + \alpha r_l)}\right)^2 + \frac{2h_1^2}{\rho_1^4\sigma_\zeta^4}\log(\frac{d}{\delta}) + \frac{2Lh_1\beta_T\sqrt{8\log(\frac{d}{\delta})}}{\rho_1^3\sigma_\zeta^3(\kappa_l + \alpha r_l)},$$

where $h_1 = 2\rho_1(1-\rho_1)L + 2\rho_1^2$ and $\rho_1 = (\frac{r_l}{S+r_h})\alpha$.

Remark 5.4.4 *The upper bound on the number of times SCLTS plays the conservative actions up to time T provided in Theorem 5.4.3 has the order $\mathcal{O}\left(\frac{L^2 d\log(\frac{T}{\delta})\log(\frac{d}{\delta})}{\alpha^4(r_l^2 \wedge r_l^4)\kappa_l(\sigma_\zeta^2 \wedge \sigma_\zeta^4)}\right)$.*

The first idea of the proof is based on the intuition that if a baseline action is played at round τ, then the algorithm does not yet have a good estimate of the unknown parameter $\theta_\star$ and the safe actions played thus far have not yet expanded properly in all directions. Formally, this translates to small $\lambda_{\min}(V_\tau)$ and the upper bound $O(\log \tau) \geq \lambda_{\min}(V_\tau)$. The second key idea is to exploit the randomized nature of the conservative actions (cf. (11)) to lower bound $\lambda_{\min}(V_\tau)$ by the number of times (N_τ^c) that SCLTS plays the baseline actions up to that round (cf. Lemma C.4.1 in Appendix C). Putting these together leads to the advertised upper bound $O(\log T)$ on the total number of times (N_T^c) the algorithm plays the baseline actions.

5.4.1 Additional Side Constraint with Bandit Feedback

We also consider the setting where the constraint depends on an unknown parameter that is different than the one in reward function. In particular, we assume the constraint of the form

$$\langle x_t, \mu_\star \rangle \geq (1-\alpha)q_{b_t}, \tag{5.4.2}$$

which needs to be satisfied by the action x_t at every round t. In (5.4.2), $\mu_\star$ is a fixed, but unknown and the positive constants $q_{b_t} = \langle x_{b_t}, \mu_\star \rangle$ are known to the learner. In Section 5.5, we relax this assumption and consider the case where the learner does not know the value of q_{b_t}. Let $\nu_{b_t} = \langle x_\star, \mu_\star \rangle - \langle x_{b_t}, \mu_\star \rangle$. Similar to Assumption 12, we assume there exist constants $0 \leq \nu_l \leq \nu_h$ and $0 < q_l \leq q_h$ such that $\nu_l \leq \nu_{b_t} \leq \nu_h$ and $q_l \leq q_{b_t} \leq q_h$.

We assume that with playing an action x_t, the learner observes the following bandit feedback:

$$w_t = \langle x_t, \mu_\star \rangle + \chi_t, \tag{5.4.3}$$

where χ_t is assumed to be a zero-mean R-sub-Gaussian noise. In order to handle this case, we show how SCLTS should be modified, and we propose a new algorithm called SCLTS-BF. The details on SCLTS-BF are presented in Appendix C.5. In the following, we only mention the difference between SCLTS-BF with SCLTS, and show an upper bound on its regret.

The main difference is that SCLTS-BF constructs two confidence regions $\mathcal{E}_t$ in (5.3.2) and $\mathcal{C}_t$ based on the bandit feedback such that $\theta_\star \in \mathcal{E}_t$ and $\mu_\star \in \mathcal{C}_t$ with high probability. Then, based on $\mathcal{C}_t$, it constructs the estimated safe decision set denoted $\mathcal{P}_t^s = \{x \in \mathcal{X} : \langle x, v \rangle \geq (1-\alpha) q_{b_t}, \forall v \in \mathcal{C}_t\}$. We note that SCLTS-BF only plays the actions from $\mathcal{P}_t^s$ that are safe with respect to all the parameters in $\mathcal{C}_t$.

We report the details on proving the regret bound for SCLTS-BF in Appendix C-Section C.5. We use the decomposition in Proposition 5.4.1, and we upper bound Term I similar to the Theorem 5.4.2. Then, we show an upper bound of order $\mathcal{O}\left(\frac{L^2 d \log(\frac{T}{\delta}) \log(\frac{d}{\delta})}{\alpha^4 (q_l^2 \wedge q_l^4) \nu_l (\sigma_\zeta^2 \wedge \sigma_\zeta^4)}\right)$ over the number of times that SCLTS-BF plays the conservative actions.

5.5 Unknown Baseline Reward

Inspired by [46], which studies this problem in the presence of *safety constraints on the cumulative rewards*, we consider the case where the expected reward of the action chosen by baseline policy, i.e., r_{b_t} is unknown to the learner. However, we assume that the learner knows the value of r_l in (5.2.5). We describe the required modifications on SCLTS to handle this case and present a new algorithm called SCLTS2. Then, we prove the regret bound for SCLTS2, which has the same order as SCLTS. Therefore, the lack of information about the reward of the baseline policy does not cause any harm to our algorithm in terms of the order of the regret.

Here, the learner does not know the value of r_{b_t}; however, she knows that the unknown parameter $\theta_\star$ falls in the confidence region $\mathcal{E}_t$ with high probability. Hence, we can upper bound the RHS of (5.2.2) with $\max_{v\in\mathcal{E}_t}\langle x_{b_t}, v\rangle \geq r_{b_t}$. Therefore, any action x that satisfies

$$\min_{v\in\mathcal{E}_t}\langle x(\tilde{\theta}_t), v\rangle \geq (1-\alpha)\max_{v\in\mathcal{E}_t}\langle x_{b_t}, v\rangle,$$

is safe with high probability. In order to ensure safety, SCLTS2 only plays the safe actions from the estimated safe actions set $\mathcal{Z}_t^s = \{x \in \mathcal{X} : \min_{v\in\mathcal{E}_t}\langle x, v\rangle \geq (1-\alpha)\max_{v\in\mathcal{E}_t}\langle x_{b_t}, v\rangle\}$. We report the details on SCLTS2 in Appendix C-Section C.6.

Next, we provide an upper bound on the regret of SCLTS2. To do so, we first use the decomposition in Proposition 5.4.1. The regret of Term I is similar to that of SCLTS (Theorem 5.4.2), and in Theorem 5.5.1, we prove an upper bound on the number of time SCLTS2 plays the conservative actions. Note that similar steps can be generalized to the setting of additional side constraints with bandit feedback.

Theorem 5.5.1 *Let $\lambda, L \geq 1$. On event $\{\theta_\star \in \mathcal{E}_t, \forall t \in [T]\}$, and under Assumption 12, we can upper bound the number of times SCLTS2 plays the conservative actions, i.e.,*

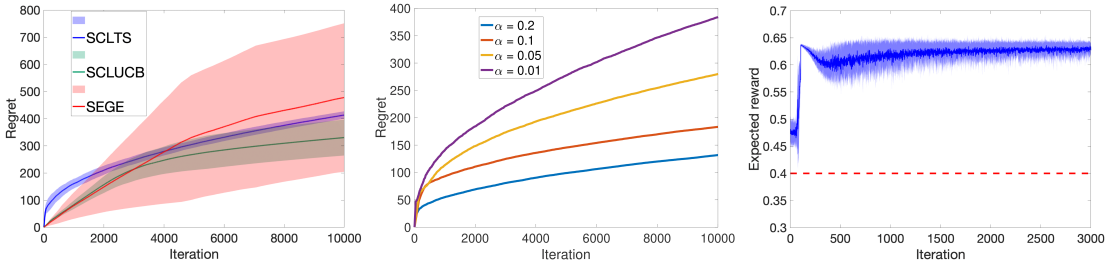

Figure 5.1: Left: comparison of the cumulative regret of SCLTS and SCLUCB versus SEGE algorithm in [71]. Middle: average regret (over 100 runs) of SCLTS algorithm for different values of α. Right: expected reward under SCLTS algorithm in the first 3000 rounds for $\alpha = 0.2$.

$|N_T^c|$ *as:*

$$|N_T^c| \leq \left(\frac{2L\beta_T(2-\alpha)}{\rho_3\sigma_\zeta(\kappa_l+\alpha r_l)}\right)^2 + \frac{2h_3^2}{\rho_3^4\sigma_\zeta^4}\log(\frac{d}{\delta}) + \frac{2Lh_3\beta_T(2-\alpha)}{\rho_3^3\sigma_\zeta^3(\kappa_l+\alpha r_l)}\sqrt{8\log(\frac{d}{\delta})}, \tag{5.5.1}$$

where $h_3 = 2\rho_3(1-\rho_3)L + 2\rho_3^2$ *and* $\rho_3 = (\frac{r_l}{S+1})\alpha$.

Remark 5.5.2 *The regret of SCLTS2 has order of* $\mathcal{O}\left(\frac{L^2 d\log(\frac{T}{\delta})\log(\frac{d}{\delta})(2-\alpha)^2}{\alpha^4(r_l^2\wedge r_l^4)\kappa_l(\sigma_\zeta^2\wedge\sigma_\zeta^4)}\right)$, *which has the same rate as that of SCLTS. Therefore, the lack of information about the reward function only hurt the regret with a constant* $(2-\alpha)^2$.

5.6 Numerical Results

In this section, we investigate the numerical performance of SCLTS and SCLUCB on synthetic data, and compare it with SEGE algorithm introduced by [71]. In all the implementations, we used the following parameters: $R = 0.1, S = 1, \lambda = 1, d = 2$. We consider the action set $\mathcal{X}$ to be a unit ball centered on the origin. The reward parameter $\theta_\star$ is drawn from $\mathcal{N}(0, I_4)$. We generate the sequence $\{\zeta_t\}_{t=1}^\infty$ to be IID random vectors that are uniformly distributed on the unit circle. The results are averaged over 100 realizations.

In Figure 5.1(left), we plot the cumulative regret of the SCLTS algorithm and SCLUCB and SEGE algorithm from [71] for $\alpha = 0.2$ over 100 realizations. The shaded regions show standard deviation around the mean. In view of the discussion in [40] regarding computational issues of LUCB algorithms with confidence regions specified with ℓ_2-norms, we implement a modified version of Safe-LUCB which uses ℓ_1-norms instead of ℓ_2-norms. Figure 5.1(left) shows that SEGE algorithm suffers a high variance of the regret over different problem instances which shows the strong dependency of the performance of SEGE algorithm on the specific problem instance. However, the regret of SCLTS and SCLUCB algorithms do not vary significantly under different problem instances, and has a low variance. Moreover, the regret of SEGE algorithm grows faster in the beginning steps, since it heavily relies on the baseline action in order to satisfy the safety constraint. However, the randomized nature of SCLTS leads to a natural exploration ability that is much faster in expanding the estimated safe set, and hence it plays the baseline actions less frequently than SEGE algorithm even in the initial exploration stages.

In Figure 5.1(middle), we plot the average regret of SCLTS for different values of α over a horizon $T = 10000$. Figure 5.1(middle) shows that, SCLTS has a better performance (i.e., smaller regret) for the larger value of α, since for the small value of α, SCLTS needs to be more conservative in order to satisfy the safety constraint, and hence it plays more baseline actions. Moreover, Figure 5.1(right) illustrates the expected reward of SCLTS algorithm in the first 3000 rounds. In this setting, we assume there exists one baseline action $x_b = [0.6, 0.5]$, which is available to the learner, $\theta_\star = [0.5, 0.4]$ and the safety fraction $\alpha = 0.2$. Thus, the safety threshold is $(1 - \alpha)\langle x_b, \theta_\star \rangle = 0.4$ (shown as a dashed red line), which SCLTS respects in all rounds. In particular, in initial rounds, SCTLS plays the conservative actions in order to respect the safety constraint, which as shown have an expected reward close to 0.475. Over time as the algorithm achieves a better estimate of the unknown parameter $\theta_\star$, it is able to play more optimistic actions

and as such receives higher rewards.

5.7 Conclusion

In this paper, we study the stage-wise conservative linear stochastic bandit problem. Specifically, we consider safety constraints that requires the action chosen by the learner at each individual stage to have an expected reward higher than a predefined fraction of the reward of a given baseline policy. We propose extensions of Linear Thompson Sampling and Linear UCB in order to minimize the regret of the learner while respecting safety constraint with high probability and provide regret guarantees for them. We also consider the setting of constraints with bandit feedback, where the safety constraint has a different unknown parameter than that of the reward function, and we propose the SCLTS-BF algorithm to handle this case. Third, we study the case where the rewards of the baseline actions are unknown to the learner. Lastly, our numerical experiments compare the performance of our algorithm to SEGE of [71] and showcase the value of the randomized nature of our exploration phase. In particular, we show that the randomized nature of SCLTS leads to a natural exploration ability that is faster in expanding the estimated safe set, and hence SCLTS plays the baseline actions less frequently as theoretically shown. For future work, natural extension of the problem setting to generalized linear bandits, and possibly with generalized linear constrains might be of interest. [73, 126] are the results of this chapter.

Chapter 6

Model Selection in Stochastic Linear Bandits

6.1 Introduction

Learning under bandit feedback is a class of online learning problems in which an agent interacts with the environment through a set of actions (arms), and receives rewards only from the arms that it has pulled. The goal of the agent is to maximize its expected cumulative reward without knowledge of the reward distributions of the arms. *Multi-armed bandit* (MAB) is the simplest form of this problem [127, 38, 128, 129]. *Linear bandit* [40, 130, 42] is a generalization of MAB to (possibly) infinitely many arms, each associated with a feature vector. The mean reward of each arm is assumed to be the dot product of its feature vector and an *unknown* parameter vector. This setting contains *contextual linear bandit* in which action sets and feature vectors change at every round. The main component of bandit algorithms is to balance *exploration* and *exploitation*: to decide when to *explore* and learn about the arms, and when to *exploit* and select the action with the highest estimated reward. The most common exploration strategies are

optimism in the face of uncertainty (OFU) or upper confidence bound (UCB) [38, 40, 42, 73, 126], and Thompson sampling (TS) [76, 44, 77, 43, 86, 85].

In this paper, we study *model selection* in stochastic linear bandits (LB), where the LB problem at hand is selected from a set of M models. The agent has information about the models but does not know the identity of the one(s) that the new LB problem has been selected from. The goal of the agent is to identify the true model(s) and transfer its (their) collected experience to speedup the learning of the task at hand. It is a common scenario in many application domains that the new task belongs to a family of models that are either known accurately or with misspecification. For example, it is reasonable to assume that the customers of an online marketing website, the users of an app, or the patients in a medical trial belong to a certain number of categories based on their shopping and browsing habits or their genetic signatures. It is also common these days that websites, apps, and clinics have a large amount of information from each of these categories that can be used to build a model.

Model selection is particularly challenging with bandit information. A common approach is to consider each model as a black-box that runs a bandit algorithm with its own information, and then a meta algorithm plays a form of bandit-over-bandits strategy with their outcomes. These algorithms often achieve a regret of $\widetilde{\mathcal{O}}(\sqrt{MT})$, and thus, are not desirable when the number of models M is large. In this paper, we consider two bandit model selection settings and show that it is possible to improve this rate so that the regret scales as $\sqrt{\log M}$ with the number of models. The main innovation in our algorithms is utilizing reductions from bandits to full-information problems, and performing model selection in the full-information setting for which much stronger results exist. The main reason for $\widetilde{\mathcal{O}}(\sqrt{MT})$ regret in bandit-over-bandits algorithms is that no information is shared among the models (bandit algorithms), i.e., when a bandit algorithm is used to take an action in a round, the resulting feedback is not shared with the other models. On

the other hand, model selection in the full-information setting allows the model to share information among each other, which makes the superior $\sqrt{\log M}$ regret bound possible.

The two model selection settings we consider in this paper are: *feature selection*, where the mean reward of the LB problem is in the linear span of at least one of M given feature maps (models), and *parameter selection*, where the reward parameter of the LB problem is arbitrarily selected from M models represented as (possibly) overlapping balls in $\mathbb{R}^d$. Here the models can be misspecified, i.e., only estimates of the centers and radii of the balls are given to the algorithm. We derive algorithms in these settings that use reductions from bandits to full-information. Our algorithms are computationally efficient and have regret bounds that are not worse (up to a $\sqrt{\log M}$ factor) than the case where the true model is known. We achieve this by properly instantiating existing algorithmic paradigms: SquareCB [131] and OFUL [42]. The SquareCB algorithm in its original form uses a set of static experts, but we need adaptive (learning) experts in order to have a computational efficient algorithm with the desired regret in our *feature selection* setting. Working with adaptive (time-varying) experts requires appropriate and non-trivial modifications to the proof of SquareCB.

There are mainly two types of reductions from bandits to full-information problems. The first one is the classical reduction that uses importance weighted estimates. A popular algorithm in this class is EXP3 that uses Exponentially Weighted Average forecaster as the full-information algorithm. The bandit model selection strategy of [132], known as CORRAL, also uses this type of reduction with an online mirror descent method and a carefully selected mirror map as the full-information algorithm. Given that importance weighted estimates are fed to the full-information algorithm, a $\sqrt{M}$ term is in general unavoidable in the regret of the methods that use this type of reduction. In this work, we use a different type of full-information reduction introduced by [131] and [133]. Here, the full-information algorithm has direct access to its losses without any importance weighted

estimates, and thus, allows us to obtain regrets that scales as $\sqrt{\log M}$ with the number of model.

6.2 Problem Formulation

In this section, we first provide a brief overview of stochastic linear bandits. We then describe the two model selection settings studied in the paper. We conclude by introducing a regression oracle used by our algorithms that is based on sequential prediction with expert advice and square loss.

6.2.1 Stochastic Linear Bandits

A stochastic linear bandit (LB) problem is defined by a sequence of T interactions of a learning agent with a stochastic environment. At each round $t \in [T]$, the agent is given a decision set $\mathcal{A}_t \subset \mathbb{R}^d$ from which it has to select an action a_t. Upon taking the action $a_t \in \mathcal{A}_t$, it observes a reward $y_t = \langle \phi_t(a_t), \theta_* \rangle + \eta_t$, where $\theta_* \in \mathbb{R}^d$ is the unknown reward parameter, $\phi_t(a) \in \mathbb{R}^d$ is the feature vector of action a at round t, and η_t is a zero-mean R-sub-Gaussian noise. When the features correspond to the canonical basis, this formulation reduces to *multi-armed bandit*. In case the features depend on both an action $a \in \mathcal{A}$ and a context $x \in \mathcal{X}$, i.e., $\phi_t(a_t) = \phi(x_t, a_t)$, this LB formulation is called *contextual linear bandit*. It is also common in practice that the action set is fixed and finite, i.e., $\mathcal{A} = [K]$, in which case we are in the finite K-action setting. The history H_t of a LB algorithm up to round t consists of all the contexts, actions, and rewards that it has observed from the beginning until the end of round $t-1$, i.e., $H_t = \{(x_s, a_s, y_s)\}_{s=1}^{t-1}$, or equivalently $H_t = \{(\phi_s(a_s), y_s)\}_{s=1}^{t-1}$.

The goal of the agent in LB is to maximize its expected cumulative reward in T

rounds, or equivalently to minimize its T-round (pseudo) regret, i.e.,

$$\mathcal{R}(T, \theta_*) = \sum_{t=1}^{T} \langle \phi_t(a_t^*), \theta_* \rangle - \langle \phi_t(a_t), \theta_* \rangle, \tag{6.2.1}$$

where $a_t^* = \text{argmax}_{a \in \mathcal{A}_t} \langle \phi_t(a), \theta_* \rangle$ is the optimal action in round t.

6.2.2 Feature Selection Setting

In this setting, the agent is given a set of M feature maps $\{\phi^i\}_{i=1}^M$ with dimension d. We assume that the expected reward of the LB problem belongs to the linear span of at least one of these M models (features), i.e., there exists an $i \in [M]$ and a $\theta_*^i \in \mathbb{R}^d$, such that for all rounds $t \in [T]$, contexts $x \in \mathcal{X}$, and actions $a \in \mathcal{A}$, we may write the mean rewards as $\mathbb{E}[y_t] = \langle \phi^i(x, a), \theta_*^i \rangle$.[1] We refer to such feature maps as *true* models and denote them by i_*. Note that the agent does not know the identity of the true model(s) i_*.

As a motivational example for this setting, we can consider a recommender system that has trained M models (e.g., M neural networks) to predict the score of customer-item pairs. Each model corresponds to a particular mood or type of the customer, or any other latent component of the customer's state. Each model provides an embedding for customer-item pairs and the score is linear in this embedding (think of an embedding as the one to the last layer of a trained NN). When a new customer arrives, the recommender system should find out as soon as possible which of the M models (embeddings) is the best match to the current mood/type of this customer in order to recommend her desirable items.

We make the following standard assumption on the boundedness of the reward pa-

[1] Note that we use the contextual linear bandit notation for this setting and in the corresponding sections.

rameters and features of the M models.

Assumption 13 *There are constants $L, S, G \geq 0$, such that for all $i \in [M]$, $t \in [T]$, $x \in \mathcal{X}$, and $a \in \mathcal{A}$, we have $\|\theta_*^i\| \leq S$, $\|\phi^i(x,a)\| \leq L$, and $|\langle \phi^i(x,a), \theta_*^i \rangle| \leq G$.*

Remark 6.2.1 *Note that we consider the same dimension d for all feature maps (models). The reason for that is a recent result by [134] in which they prove no algorithm can adapt to the unknown intrinsic dimension d_* and achieve the regret of $\mathcal{O}(\sqrt{d_* T})$, simultaneously for all values of d_*, in a sequence of nested linear hypothesis classes with dimensions $d_1 < d_2 < \ldots < d_M$.*

Our goal here is to design an algorithm that minimizes *transfer regret*, which in this setting we define it as

$$\mathcal{R}(T) = \sum_{t=1}^{T} \langle \phi^{i_*}(x_t, a_t^*), \theta_*^{i_*} \rangle - \langle \phi^{i_*}(x_t, a_t), \theta_*^{i_*} \rangle, \tag{6.2.2}$$

where $a_t^* = \text{argmax}_{a \in \mathcal{A}} \langle \phi^{i_*}(x_t, a), \theta_*^{i_*} \rangle$. In the results we report for this setting in Section 6.3, we make two assumptions: **1)** the feature maps are all known (no model misspecification), and **2)** the number of actions is finite, i.e., we are in the finite K-action setting described in Section 6.2.1. However, we believe that our algorithm and analysis can be extended to the case of having misspecified models and convex action sets using the results in [135].

6.2.3 Parameter Selection Setting

In this setting, unlike the classical setting in Section 6.2.1, we no longer assume that the unknown parameter θ_* can be any vector in $\mathbb{R}^d$. Rather, θ_* can be generated from M possible reward models, each defined as a ball $B(\mu_i, b_i) = \{\theta \in \mathbb{R}^d : \|\theta - \mu_i\| \leq b_i\}$, with center $\mu_i \in \mathbb{R}^d$ and radius $b_i \geq 0$. Note that the models (balls) may overlap and

do not have to be disjoint. The M models can be thought of the responses of M types (or clusters) of customers to different items in a recommender system or the reactions of patients with M genotypes to a set of drugs. The radii $\{b_i\}_{i=1}^M$ represent the variation within each cluster. The reward parameter θ_* of the new task (LB problem) is arbitrarily selected from the M models. For example, it can be adversarially selected from the union of the models, i.e., $\theta_* \in \bigcup_{i=1}^M B(\mu_i, b_i)$. In this case, we denote by $\mathcal{I}_*$, the set of indices of the balls that contain θ_*. Since the models are often computed from (finite) historical data, it is reasonable to assume that only *estimates* of their centers $\{\widehat{\mu}_i\}_{i=1}^M$ are available, together with upper-bounds on the error of these estimates $\{c_i\}_{i=1}^M$, such that $\|\mu_i - \widehat{\mu}_i\| \le c_i$, for all $i \in [M]$.

The agent has no knowledge either about θ_* or the process according to which it has been selected. The only information given to the agent are: **1)** estimates $\widehat{\mu}_i$ of the center of the models, **2)** upper-bounds c_i on the errors of these estimates, and **3)** the exact radii b_i of the models, for all $i \in [M]$. This means that although θ_* is selected from the *actual* models $B(\mu_i, b_i)$, the agent has only access to *estimated* models $B(\widehat{\mu}_i, b_i + c_i)$ that have more uncertainty (their corresponding balls are larger). For simplicity, we assume that the exact values of radii $\{b_i\}_{i=1}^M$ are known. However, our results can be easily extended to the case that instead of b_i's, their estimates $\widehat{b}_i$ and upper-bounds on their errors c_i', i.e., $\|b_i - \widehat{b}_i\| \le c_i'$, for all $i \in [M]$, are given to the agent. In this case, the agent has to use even more uncertain estimates of the models $B(\widehat{\mu}_i, b_i + c_i + c_i')$.

Our goal is to design an algorithm that can transfer knowledge from these estimated models and learn the new task with parameter θ_* more efficiently than when it is independently learned. This goal can be quantitatively stated as minimizing the *transfer regret*,

$$\mathcal{R}(T) = \sup_{\theta_* \in \bigcup_{i=1}^M B(\mu_i, b_i)} \mathcal{R}(T, \theta_*), \tag{6.2.3}$$

where $\mathcal{R}(T, \theta_*)$ is the regret defined by (6.2.1). We make the following standard assumption on the boundedness of the features and expected rewards.

Assumption 14 *There exist constants $L, G \geq 0$, such that $\forall t \in [T]$ and $\forall a \in \bigcup_{t=1}^{T} \mathcal{A}_t$, we have $\|\phi_t(a)\| \leq L$, and $\forall \theta \in \bigcup_{i=1}^{M} B(\mu_i, b_i)$, we have $|\langle \phi_t(a), \theta \rangle| \leq G$.*

6.2.4 Regression Oracle

In both model selection settings studied in the paper, our proposed algorithms use a regression oracle that is based on sequential prediction with expert advice and square loss. Following [131] and [135], we refer to this regression oracle as `SqAlg`. We can consider `SqAlg` as a meta algorithm that consists of M learning algorithms (or experts), each corresponding to one of our M models, and returns a prediction by aggregating the predictions of its experts. More precisely, in each round $t \in [T]$, `SqAlg` takes the current context-action pair (x_t, a_t), or equivalently $\phi_t(a_t)$, as input, and gives them to its M experts to predict their reward, i.e., $f_t^i(H_t) = f^i(\phi_t(a_t); H_t)$, $\forall i \in [M]$, given the current history H_t. Then, the meta algorithm `SqAlg` aggregates its experts' predictions, $\{f_t^i(H_t)\}_{i=1}^{M}$, given their current weights, and returns its own prediction $\widehat{y}_t = \texttt{SqAlg}_t(\phi_t(a_t); H_t)$. Upon observing the actual reward y_t, `SqAlg` updates the weights of its experts according to the difference between their predictions $f^i(\phi_t(a_t); H_t)$ and the actual reward y_t.

The regression oracles (`SqAlg`) used by our model selection algorithms differ in the prediction algorithm used by their experts. However, in both cases, `SqAlg` aggregates its experts' predictions using an algorithm by [136] (see Algorithm 12 in Appendix D-Section D.1). The performance of `SqAlg` is evaluated in terms of its regret $\mathcal{R}_{\texttt{Sq}}(T)$, which is defined as its accuracy (in terms of square loss) w.r.t. the accuracy of the best expert in the set, i.e.,

$$\sum_{t=1}^{T} (\widehat{y}_t - y_t)^2 - \min_{i \in [M]} \sum_{t=1}^{T} (f_t^i(H_t) - y_t)^2 \leq \mathcal{R}_{\texttt{Sq}}(T). \tag{6.2.4}$$

In each round t, we define the oracle prediction for a context x and an action a as $\widehat{y}_t(x,a) := \texttt{SqAlg}_t(x,a;H_t)$. As shown in [136], in case all observations and experts' predictions are bounded in an interval of size ℓ, this regret can be bounded as $\mathcal{R}_{\texttt{Sq}}(T) \leq \ell^2 \log M$ (see Appendix D-Section D.1 for more details). We use this regret bound in the analysis of our proposed algorithms.

6.3 Feature Selection Algorithm

In this section, we derive an algorithm for the feature selection setting described in Section 6.2.2 that is based on the SquareCB algorithm [131]. We refer to our algorithm as *feature selection SquareCB* (FS-SCB). We prove an upper-bound on the transfer regret of FS-SCB in Section 6.3.1, and provide an overview of the related work and a discussion on our results in Section 6.3.2.

Algorithm 6 contains the pseudo-code of FS-SCB. In each round $t \in [T]$, the algorithm observes a context $x_t \in \mathcal{X}$ and passes it to its regression oracle `SqAlg` to produce its reward predictions $\widehat{y}_t(x_t,a), \forall a \in [K]$. Each expert in `SqAlg` corresponds to one of the M models and is a *ridge regression* algorithm with the feature map of that model. Expert $i \in [M]$ predicts the reward of the context x_t, for each action $a \in [K]$, as $f^i(x_t,a;H_t) = \langle \phi^i(x_t,a), \widehat{\theta}^i_t \rangle$, where $\widehat{\theta}^i_t = \text{argmin}_\theta \, \|\Phi_t^{i\top}\theta - Y_t\|^2 + \lambda_i \|\theta\|^2$. We may write $\widehat{\theta}^i_t$ in closed-form as $\widehat{\theta}^i_t = (V_t^{\lambda_i})^{-1}\Phi_t^{i\top}Y_t$. In these equations, $Y_t = (y_1,\dots,y_{t-1})^\top$ is the reward vector; Φ^i_t is the feature matrix of the i^{th} model, whose rows are $\phi^i(x_1,a_1),\dots,\phi^i(x_{t-1},a_{t-1})$; λ_i is the regularization parameter of model i, which our analysis shows that it only needs to be larger than one, i.e., $\lambda_i \geq 1$; and finally $V_t^{\lambda_i} = \lambda_i I + \Phi_t^\top \Phi_t$. The meta algorithm `SqAlg` aggregates the experts' predictions $\{f^i(x_t,a;H_t)\}_{i=1}^M$ and produces its own predictions $\widehat{y}_t(x_t,a)$, $\forall a \in [K]$, using Algorithm 12 in Appendix D-SectionD.1 (see Remark 6.3.1).

The next step in FS-SCB is computing the action with the highest predicted reward,

Algorithm 6: Feature Selection Square-CB (FS-SCB)

Input: Models $\{\phi^i\}_{i=1}^M$, Confidence Parameter δ, Learning Rate α, Exploration Parameter κ

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

  Observe context x_t

  Oracle predicts:

  $\widehat{y}_t(x_t, a) = \texttt{SqAlg}_t(x_t, a; H_t), \quad \forall a \in [K]$

  Define a distribution p_t over the actions:

$$p_t(a) = \begin{cases} \frac{1}{\kappa + \alpha\left(\widehat{y}_t(x_t,a) - \widehat{y}_t(x_t,a'_t)\right)}, & a \neq a'_t, \\ 1 - \sum_{a \neq a'_t} p_t(a), & a = a'_t, \end{cases} \tag{6.3.1}$$

  where $a'_t = \operatorname{argmax}_{a \in [K]} \widehat{y}_t(x_t, a)$;

  Sample action $a_t \sim p_t(\cdot)$ and play it;

  Build the confidence region $\mathcal{E}_t(\delta')$ in (5.3.2)

  Observe reward $y_t = \langle \phi^{i_*}(x_t, a_t), \theta^{i_*}_* \rangle + \eta_t$;

  Update $\texttt{SqAlg}$ with (x_t, a_t, y_t);

end for

i.e., $a'_t = \operatorname{argmax}_{a \in [K]} \widehat{y}_t(x_t, a)$, and using it to define a distribution $p_t \in \Delta_K$ over the actions (see Eq. 6.3.1). The distribution p_t in (6.3.1) is defined similarly to the probability selection scheme of [137], and assigns a probability to every action inversely proportional to the gap between its prediction and that of a'_t. The algorithm then samples its action a_t from p_t, observes reward y_t, and feeds the tuple (x_t, a_t, y_t) to the oracle to update its weights over the experts. Our analysis in Section 6.3.1 and Appendix D-Section D.2 suggest to set the exploration parameter to $\kappa = K$ and the learning rate to $\alpha = \sqrt{KT/D_T(\delta)}$, where we define $D_T(\delta)$ in Lemma 6.3.3 and give its exact expression in Eq. D.2.25 in Appendix D-Section D.2.2.

Remark 6.3.1 (Admissible Experts) *It is important to note that in each round $t \in [T]$, FS-SCB only uses predictions by admissible experts, i.e., experts i that belong to the*

set

$$\mathcal{S}_t := \left\{ i \in \mathcal{S}_{t-1} : \langle \phi^i(x_t, a), \widehat{\theta}_t^i \rangle \leq G + RL\sqrt{d \log\left(\frac{1 + \frac{tL^2}{\lambda_i d}}{\delta}\right)} + L\sqrt{\lambda_i} S, \quad \forall a \in [K] \right\}, \tag{6.3.2}$$

with $\mathcal{S}_0 = [M]$. *This is the set of experts* i *whose predictions* $f^i(x_t, a; H_t) = \langle \phi^i(x_t, a), \widehat{\theta}_t^i \rangle$, $\forall a \in [K]$ *are within a bound defined by* (6.3.2). *When an expert was removed from the admissible set in a round* t, *it will remain out for the rest of the game. We discuss the technical reasons for defining this set in the proof of Lemma 6.3.5 in Appendix D-Section D.2.2.*

6.3.1 Regret Analysis of FS-SCB

We state a regret bound for FS-SCB followed by a proof sketch. The detailed proofs are all reported in Appendix D-Section D.2.

Theorem 6.3.2 *Let Assumption 13 hold and the regularization parameters* λ_i, *exploration parameter* κ, *and learning rate* α *set to the values described above. Then, for any* $\delta \in [0, 1/4)$, *w.p. at least* $1 - \delta$, *the regret defined by* (6.2.2) *for FS-SCB is bounded as*

$$\mathcal{R}_{\text{FS-SCB}}(T) \leq \mathcal{O}\Bigg(\sqrt{2T \log(2/\delta)} + RLG \times \sqrt{KT(1 + \log(M)) \max_{i \in [M]} \left\{ \lambda_i S^2 + 4d \log\left(\frac{1 + \frac{TL^2}{\lambda_i d}}{\delta}\right) \right\}} \Bigg).$$

Proof Sketch. The proof consists of two main steps:

Step 1. We first need to bound the prediction error of the online regression oracle.

Lemma 6.3.3 *For any* $\delta \in (0, 1/4]$, *w.p. at least* $1 - \delta$, *we can bound the prediction error*

of the regression oracle as

$$\sum_{s=1}^{t-1}\big(\widehat{y}_s(x_s,a_s)-\langle\phi^{i_*}(x_s,a_s),\theta_*^{i_*}\rangle\big)^2\leq D_t(\delta):=$$
$$\mathcal{O}\Big(\big(1+R^2L^2G^2\log(M)\big)\max_{i\in[M]}\big\{\lambda_i S^2+4d\log\big(\frac{1+\frac{TL^2}{\lambda_i d}}{\delta}\big)\big\}\Big).$$

The exact definition of $D_t(\delta)$ (see Eq. D.2.25 in Appendix D-Section D.2.3) shows its dependence on the following two terms: **1)** an upper-bound Q_t on the prediction error of the true models,

$$\max_{i_*}\sum_{s=1}^{t-1}\big(\langle\phi^{i_*}(x_s,a_s),\widehat{\theta}_s^{i_*}\rangle-\langle\phi^{i_*}(x_s,a_s),\theta_*^{i_*}\rangle\big)^2\leq Q_t, \tag{6.3.3}$$

and **2)** the regret $\mathcal{R}_{\texttt{Sq}}(t)$ of the regression oracle. Thus, the proof of Lemma 6.3.3 requires finding expressions for these quantities, which we derive them in the following lemmas.

Lemma 6.3.4 *For any $\delta\in(0,1)$, with probability at least $1-\delta$, we may write Q_t defined in (6.3.3) as (see Eq. D.2.9 in Appendix D-Section D.2.1 for the exact expression)*

$$Q_t=\mathcal{O}\Big(\max_{i\in[M]}\big\{\lambda_i S^2+4d\log\big(1+\frac{tL^2}{\lambda_i d}\big)\big\}+R^2\log(1/\delta)\Big).$$

Lemma 6.3.5 *For any $\delta\in(0,1)$, with probability at least $1-\delta$, we may write the regret of the regression oracle as (see Eq. D.2.19 in Appendix D.2.2 for the exact expression)*

$$\mathcal{R}_{\texttt{Sq}}(t)=\mathcal{O}\Big(R^2L^2\log(M)\times$$
$$\Big(G^2+\max_{i\in[M]}\big\{\lambda_i S^2+d\log\big(1+\frac{tL^2}{\lambda_i d}\big)\big\}+\log(1/\delta)\Big)\Big).$$

Step 2. We then show how the overall regret of FS-SCB is related to the prediction

error of the online regression oracle, $D_t(\delta)$, using the following lemma:

Lemma 6.3.6 *Under the same assumptions as Theorem 6.3.2, for any $\delta \in (0, 1/4]$, with probability at least $1-\delta$, the regret of the FS-SCB algorithm is bounded as*

$$\mathcal{R}_{\text{FS-SCB}}(T) \leq \sqrt{2T\log(2/\delta)} + \frac{\alpha}{4} D_T(\delta) + \sum_{t=1}^{T} \sum_{a\in[K]} p_t(a) \Big(\langle \phi^{i_*}(x_t, a), \theta_*^{i_*} \rangle - \langle \phi^{i_*}(x_t, a_t^*), \theta_*^{i_*} \rangle - \frac{\alpha}{4} \big(\widehat{y}_t(x_t, a) - \langle \phi^{i_*}(x_t, a_t), \theta_*^{i_*} \rangle \big)^2 \Big). \tag{6.3.4}$$

Finally, we conclude the proof of Theorem 6.3.2 by bounding the last term on the RHS of (6.3.4) using Lemma D.2.1 (see Appendix D-Section D.2.5 for details).

6.3.2 Related Work (Feature Selection)

The most straightforward solution to the feature selection problem described in Section 6.2.2 is to concatenate all models (feature maps) and build a $(M \times d)$-dimensional feature, and then search for the sparse reward parameter $\theta_* \in \mathbb{R}^{Md}$ with only d non-zero elements. We may then solve the resulting LB problem using a sparse LB algorithm (e.g., [133]). This approach would result in a regret bound of $\widetilde{\mathcal{O}}(d\sqrt{MT})$, which may not be desirable when the number of models M is large.

Another approach is to use the EXP4 (or SquareCB) algorithm [138] to obtain a regret that scales only logarithmically with M. If we partition the linear space of each model into $\mathcal{O}(2^d)$ predictors, we will have the total number of $\mathcal{O}(M2^d)$ predictors. Predictor $(i, j) \in ([M], [2^d])$ is associated with a linear map $\theta^{ij} \in \mathbb{R}^d$ and recommends the action $\text{argmax}_{a\in A} \langle \phi^i(a), \theta^{ij} \rangle$. The regret of EXP4 with this set of experts is of $\widetilde{\mathcal{O}}(\sqrt{dKT\log M})$. Although this solution has logarithmic dependence on M, it is still not desirable, since

it is not computationally efficient (requires handling $M2^d$ predictors).

To have computational efficiency, we can use the approach of [139], but this results in a $O(T^{2/3})$ regret. They designed a model selection strategy using an EXP4 algorithm with a set of experts that are instances of the S-EXP3 algorithm of [140]. The interesting fact is that each S-EXP3 expert is a learning algorithm and competes against a set of mappings. The overall regret of this algorithm is of $\widetilde{\mathcal{O}}\big(T^{2/3}(|S|K\log K)^{1/3}\sqrt{\log M}\big)$ (see [141, Chapter 4.2]). If we apply this algorithm to our setting, the resulting regret bound is of $\widetilde{\mathcal{O}}(T^{2/3}d^{1/3}K^{1/3}\sqrt{\log M})$. Although the algorithm is computationally more efficient than EXP4 and its regret has logarithmic dependence on M, it is still not desirable as its dependence on T is of $\widetilde{\mathcal{O}}(T^{2/3})$, which is not optimal.

The novelty of our results is that we propose a computationally efficient algorithm, whose regret has better dependence on M and T, i.e., $\widetilde{\mathcal{O}}(\sqrt{KT\log M})$, than all the existing methods. Our FS-SCB algorithm achieves this by **1)** using a novel instantiation of SquareCB, or more precisely by constructing a proper full information algorithm (expert), and **2)** using SquareCB with a set of *adaptive (learning)*, and not static, least-squares experts. Note that SquareCB is a reduction that turns any online regression oracle into an algorithm for contextual bandits [131].

More recently, [142] studied a feature selection problem where the reward function is linear in *all* M feature maps (all models are *realizable*). Under this *stronger* assumption (than ours), they prove a regret bound that is competitive (up to a $\log M$ factor) with that of a linear bandit algorithm that uses the best feature map. More specifically, if one of the feature maps is such that a constant regret is achievable, the overall model selection strategy also achieves a constant regret. Although our focus is not on constant regret, we are able to achieve our results without requiring all models to be *realizable*.

6.4 Parameter Selection Algorithm

We propose a UCB-style algorithm for the parameter selection setting described in Section 6.2.3, which we refer to as *parameter selection OFUL* (PS-OFUL). We then provide an upper-bound on its transfer regret and conclude with a discussion on the existing results related to this setting.

Algorithm 7 contains the pseudo-code of PS-OFUL. The novel idea in PS-OFUL is the construction of its confidence set $\mathcal{C}_t$ (Eq. 6.4.2), which is based on the predictions $\{\widehat{y}_s\}_{s=1}^{t-1}$ by a regression oracle `SqAlg`. As described in Section 6.2.4, `SqAlg` is a meta algorithm that consists of M learning algorithms (or experts), and its predictions $\widehat{y}_t$ are aggregates of its experts' predictions $f^i(\phi_t(a_t); H_t)$, $\forall i \in [M]$. In PS-OFUL, each expert $i \in [M]$ is a *biased regularized least-squares* algorithm with bias $\widehat{\mu}_i$, i.e., our estimate of the center of the i^{th} ball (model). Expert i predicts the reward of the context-action $\phi_t(a_t)$ as $f^i(\phi_t(a_t); H_t) = \langle \phi_t(a_t), \widehat{\theta}_t^i \rangle$, where $\widehat{\theta}_t^i = \operatorname{argmin}_\theta \|\Phi_t^\top \theta - Y_t\|^2 + \lambda_i \|\theta - \widehat{\mu}_i\|^2$. We may write $\widehat{\theta}_t^i$ in closed-form as $\widehat{\theta}_t^i = (V_t^{\lambda_i})^{-1} \Phi_t^\top (Y_t - \Phi_t \widehat{\mu}_i) + \widehat{\mu}_i$. In these equations, the reward vector Y_t and $V_t^{\lambda_i}$ are defined as in Section 6.3; Φ_t is the feature matrix, whose rows are $\phi_1(a_1), \dots, \phi_{t-1}(a_{t-1})$; and λ_i is the regularization parameter of expert i. Our analysis in Section 6.4.1 and Appendix D.3 suggests to set them to $\lambda_i = \frac{1}{(b_i + c_i)^2}$.

The PS-OFUL algorithm takes the feature map ϕ and models $\{B(\widehat{\mu}_i, b_i + c_i)\}_{i=1}^M$ as input. At each round $t \in [T]$, it first constructs a confidence set $\mathcal{C}_{t-1}$ using the predictions of the regression oracle $\{\widehat{y}_s\}_{s=1}^{t-1}$. The radius $\gamma_t(\delta)$ of the confidence set $\mathcal{C}_t$ is defined by two terms: **1)** the regret $\mathcal{R}_{\texttt{Sq}}(t)$ of the regression oracle `SqAlg`, defined by (6.2.4), and **2)** an upper-bound U_t on the prediction error of the true models (i.e., models that contain θ_*), i.e.,

$$\max_{i \in \mathcal{I}_*} \sum_{s=1}^{t-1} \left(\langle \phi_s(a_s), \widehat{\theta}_t^i \rangle - \langle \phi_s(a_s), \theta_* \rangle \right)^2 \leq U_t. \tag{6.4.1}$$

The exact values of U_t, $\mathcal{R}_{\texttt{Sq}}(t)$, and $\gamma_t(\delta)$ come from our analysis and have been stated in Eq. D.3.22 in Appendix D-Section D.3.3. PS-OFUL then computes action a_t as the one that attains the maximum optimistic reward w.r.t. the confidence set $\mathcal{C}_{t-1}$. Using a_t, it calculates $\widehat{y}_t = \texttt{SqAlg}_t(\phi_t(a_t); H_t))$. As described in Section 6.2.4, `SqAlg` makes use of Algorithm 12 in Appendix D-Section D.1 to return its prediction $\widehat{y}_t$ as an aggregate of its experts' predictions (see Remark 6.4.1). Finally, PS-OFUL takes action a_t, observes reward y_t, and pass the sample $(\phi_t(a_t), y_t)$ to `SqAlg`. This sample is then used within `SqAlg` to evaluate its experts and to update their weights.

Algorithm 7: Parameter Selection OFUL (PS-OFUL)

Input: Feature Map ϕ, Confidence Parameter δ, Models $\{B(\widehat{\mu}_i, b_i + c_i)\}_{i=1}^M$
for $t = 1, \ldots, T$ **do**
  Construct the confidence set:

$$\mathcal{C}_{t-1} = \left\{ \theta : \sum_{s=1}^{t-1} \left(\widehat{y}_s - \langle \phi_s(a_s), \theta \rangle\right)^2 \leq \gamma_{t-1}(\delta) \right\} \tag{6.4.2}$$

  Take action: $a_t = \arg\max_{a \in \mathcal{A}_t} \max_{\theta \in \mathcal{C}_{t-1}} \langle \phi_t(a), \theta \rangle$
  Oracle predicts: $\widehat{y}_t = \texttt{SqAlg}_t(\phi_t(a_t); H_t)$
  Observe reward: $y_t = \langle \phi_t(a_t), \theta_* \rangle + \eta_t$
  Update `SqAlg` with $(\phi_t(a_t), y_t)$;
end for

Remark 6.4.1 (Admissible Experts) *Similar to FS-SCB, in each round $t \in [T]$, PS-OFUL only uses predictions by admissible experts, i.e., experts i that belong to the set*

$$\mathcal{S}_t := \Big\{ i \in \mathcal{S}_{t-1} : \langle \phi_t(a_t), \widehat{\theta}_t^i \rangle \leq G + RL\sqrt{d \log\left(\frac{1 + \frac{tL^2}{\lambda_i d}}{\delta}\right)} + L\sqrt{\lambda_i}(b_i + c_i) \Big\}, \tag{6.4.3}$$

with $\mathcal{S}_0 = [M]$. This is the set of experts i whose prediction $f^i(\phi_t(a_t); H_t) = \langle \phi_t(a_t), \widehat{\theta}_t^i \rangle$

is within a bound defined by (6.4.3). When an expert was removed from the admissible set in a round t, it will remain out for the rest of the game. We discuss the technical reasons for defining this set in the proof of Lemma 6.4.5 in Appendix D-Section D.3.2.

6.4.1 Regret Analysis of PS-OFUL

We state a regret bound for PS-OFUL followed by a proof sketch. The detailed proofs are all reported in Appendix D-Section D.3.

Theorem 6.4.2 *Let Assumption 14 hold and $\lambda_i = \frac{1}{(b_i+c_i)^2} \geq 1,\ \forall i \in [M]$. Then, for any $\delta \in (0, 1/4]$, with probability at least $1-\delta$, the transfer-regret defined by (6.2.3) of PS-OFUL is bounded as*

$$\mathcal{R}(T) = \mathcal{O}\Bigg(dRL \max\{1, G\} \sqrt{1 + \log(M)} \times \sqrt{T \log\left(1 + \frac{T}{d}\right) \log\left(\frac{1 + \frac{TL^2 \max_{i\in[M]}(b_i+c_i)^2}{d}}{\delta}\right)} \Bigg). \tag{6.4.4}$$

Proof Sketch. The proof consists of two main steps.

Step 1. We first fully specify the confidence set $\mathcal{C}_t$ and prove its validity i.e., $\mathbb{P}(\theta_* \in \mathcal{C}_t) \geq 1-\delta,\ \forall t \in [T]$.

Theorem 6.4.3 *Under the same assumptions as Theorem 6.4.2, the radius $\gamma_t(\delta)$ of the confidence set $\mathcal{C}_t$ is fully specified by Eq. D.3.22 in Appendix D.3.3. Moreover, for any $\delta \in (0, 1/4]$, with probability at least $1-\delta$, the true reward parameter θ_* lies in $\mathcal{C}_t$, i.e., $\mathbb{P}(\theta_* \in \mathcal{C}_t) \geq 1-\delta$.*

The definition of $\gamma_t(\delta)$ in Eq. D.3.22 shows its dependence on U_t and $\mathcal{R}_{\mathtt{Sq}}(t)$, defined by (6.4.1) and (6.2.4), respectively. Thus, the proof of Thm. 6.4.3 requires finding expressions for these quantities, which we derive them in the following lemmas.

Lemma 6.4.4 *Setting* $\lambda_i = \frac{1}{(b_i+c_i)^2}$, $\forall i \in [M]$, *with probability* $1-\delta$, *we may write* U_t, *defined by* (6.4.1), *as (see Eq. D.3.9 in Appendix D-Section D.3.1 for the exact expression)*

$$U_t = \mathcal{O}\Big(dR^2 \log\big(\frac{1 + \frac{tL^2 \max_{i\in[M]}(b_i+c_i)^2}{d}}{\delta}\big)\Big).$$

Lemma 6.4.5 *Setting* $\lambda_i = \frac{1}{(b_i+c_i)^2}$, $\forall i \in [M]$, *with probability* $1-\delta$, *we may write* $\mathcal{R}_{\texttt{Sq}}(t)$, *defined by* (6.2.4), *as (see Eq. D.3.15 in Appendix D.3.2 for the exact expression)*

$$\mathcal{R}_{\texttt{Sq}}(t) = \mathcal{O}\Big(dR^2L^2 \log(M) \log\big(\frac{1 + \frac{tL^2 \max_{i\in[M]}(b_i+c_i)^2}{d}}{\delta}\big)\Big).$$

Step 2. We then show how the regret is related to the confidence sets using the following lemma:

Lemma 6.4.6 *Under the same assumptions as Theorem 6.4.2, for any* $\delta \in (0, 1/4]$, *with probability at least* $1-\delta$, *the regret of PS-OFUL is bounded as*

$$\mathcal{R}_{\text{PS-OFUL}}(T) \leq 2Gd + 2\max\{1, G\}\sqrt{2dT \log\big(1+\frac{T}{d}\big) \max_{d<t\leq T} \gamma_t(\delta)}. \tag{6.4.5}$$

We conclude the proof of Theorem 6.4.2 by plugging the confidence radius $\gamma_t(\delta)$ computed in Theorem 6.4.3 (Eq. D.3.22 in Appendix D-Section D.3.3) into the regret bound (6.4.5).

6.4.2 Related Work (Parameter Selection)

[143] and [144] studied *meta learning* in stochastic linear bandit (LB), where the agent solves a sequence of LB problems, whose reward parameters θ_* are drawn from an unknown distribution ρ of bounded support in $\mathbb{R}^d$. For each LB task, the agent is given an

estimate of the mean of the distribution ρ and an upper-bound of its error, and its goal is to minimize the transfer regret $\mathcal{R}(T,\rho) = \mathbb{E}_{\theta_*\sim\rho}\big[\mathbb{E}[\mathcal{R}(T,\theta_*)]\big]$. Their proposed algorithms assume knowing the variance term $\mathrm{Var}_h = \mathbb{E}_{\theta_*\sim\rho}\big[\|\theta_* - h\|^2\big]$, for any $h \in \mathbb{R}^d$, in order to properly set their regularization parameter λ. Thus, the parameter selection setting studied in our paper can be seen as an extension of their transfer learning setting to multiple (M) models. Moreover, we allow the reward parameter of the new LB problem θ_* to be selected arbitrarily from the M models, and consider a worst-case transfer regret (see Eq. 6.2.3) for our algorithm (instead of a regret in expectation w.r.t. ρ). Despite these differences, our setting is similar to theirs as we are also given an estimate of the center of each model $\widehat{\mu}_i$, together with an upper-bound on its error c_i, plus the radius b_i of each model. Also similar to their results, our analysis clearly shows the importance of the choice of the regularization parameters, $\lambda_i = 1/(b_i+c_i)^2$, for obtaining a regret bound that only logarithmically depends on the maximum model uncertainty, i.e., $\max_{i\in[M]}(b_i+c_i)^2$.

Our parameter selection setting is also related to *latent* bandits [145, 146] in which identifying the true latent variable is analogous to finding the correct model. The latest work in this area is by [146] in which the agent faces a K-armed LB problem selected from a set of M known K-dimensional reward vectors. They proposed UCB and TS algorithms for this setting and showed that their regret (Bayes regret in case of TS) are bounded as $3M + 2T\varepsilon + 2R\sqrt{6MT\log T}$, where the reward vectors are known up to an error of ε. Comparing to their results, the regret of PS-OFUL in (6.4.4) has a better dependence on the number of models, $\sqrt{\log M}$ vs. M, and the model uncertainty, $\sqrt{\log(\max_{i\in[M]}(b_i+c_i)^2)}$ vs. ε. However, the number of actions K does not appear in their bound, while the bound of PS-OFUL will have a $\sqrt{K}$ factor when applied to K-armed bandit problems. If the objective is to have a better scaling in K, we can use a different bandit model selection strategy, called *regret balancing* [147, 148], to obtain an improved regret that scales as $\min\{\varepsilon T + \sqrt{MT}, \sqrt{KMT}\}$ (see Appendix D-Section D.5

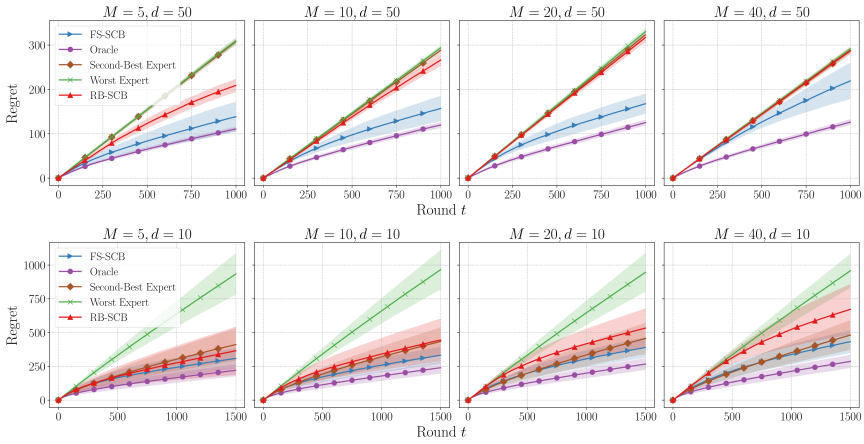

Figure 6.1: Feature selection in the synthetic LB problem *(top)* and MNIST *(bottom)*. The regrets are averaged over 100 LB problems.

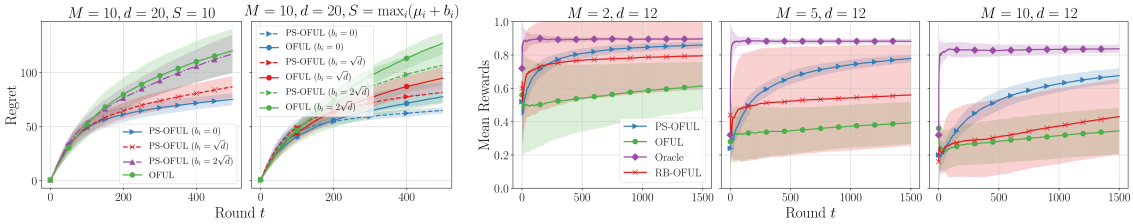

Figure 6.2: Parameter selection in the synthetic LB problem *(left)* and CIFAR-10 *(right)*. Results are averaged over 50 runs.

for details).

In another closely related work, [149] approach a similar problem by initializing TS with a prior that is a mixture of M distributions. They prove a Bayes regret bound for their algorithm in case of Gaussian mixtures that has $\sqrt{M}$ dependence on the number of models and $\sqrt{\max_{i\in[M]} \sigma_{0,i}^2}$ dependence on the maximum variance of the Gaussian priors. Both these dependences are logarithmic $\sqrt{\log M}$ and $\sqrt{\log(\max_{i\in[M]}(b_i + c_i)^2)}$ in the regret of PS-OFUL.

6.5 Experiments

We evaluate the performances of our FS-SCB and PS-OFUL algorithms using a synthetic LB problem and image classification problems: MNIST [150] and CIFAR-10 [151]. We report the details of our experimental setup and additional results in Appendix D-Section D.6.

Feature Selection (Synthetic): We first sample the parameter of the linear bandit problem from a $d = 50$ dimensional Gaussian with variance 0.01: $\theta_* \sim \mathcal{N}(0, 0.01 I_d)$. We generate all feature maps, $\{\phi^i(a)\}_{i=1}^M$, by sampling $10,000$ vectors from the Gaussian with mean θ_* and covariance $0.1 I_d$, i.e., $\phi^i(a) \sim \mathcal{N}(\theta_*, 0.1 I_d)$, for $a = 1, \ldots, 10,000$. This implies that all M feature maps have the same bias. We set $\phi^1(\cdot)$ to be the *true* feature map. At each round $t \in [T]$, the learner is given an action set consist of 10 numbers from $\mathcal{A} = \{1, 2, \ldots, 10,000\}$. The reward of each action a is $\langle \phi^1(a), \theta_\star \rangle + \eta_t$, where $\eta_t \sim \mathcal{U}[-0.5, 0.5]$.

Feature Selection (MNIST): We train a convolutional neural network (CNN) with M different number of epochs on MNIST data, and use their second layer to the last as our $d = 10$-dimensional feature maps $\{\phi^i\}_{i=1}^M$. These feature maps have test accuracy between 20% (worst model) and 97% (best model). We set the best one as true model ϕ^{i_*}. For each class $s \in \mathcal{S} = \{0, \ldots, 9\}$, we fit a linear model, given the feature map ϕ^{i_*}, and obtain parameters $\{\theta_s^{i_*}\}_{s=0}^9$. At the beginning of each LB task, we select a class $s_* \in \mathcal{S}$ uniformly at random and set its parameter to $\theta_{s_*}^{i_*}$. At each round $t \in [T]$, the learner is given an action set consists of 10 images, one from class s_* and the rest randomly selected from the other classes. The reward of each action a is defined as $\langle \phi^{i_*}(a), \theta_{s_*}^{i_*} \rangle + \eta_t \in [0, 1]$, where $\phi^{i_*}(a)$ is the application of the feature map ϕ^{i_*} to the image corresponding to action a and $\eta_t \sim \mathcal{U}[-0.5, 0.5]$ is the noise.

In Figure 6.1, we compare the regret of our FS-SCB algorithm for different number

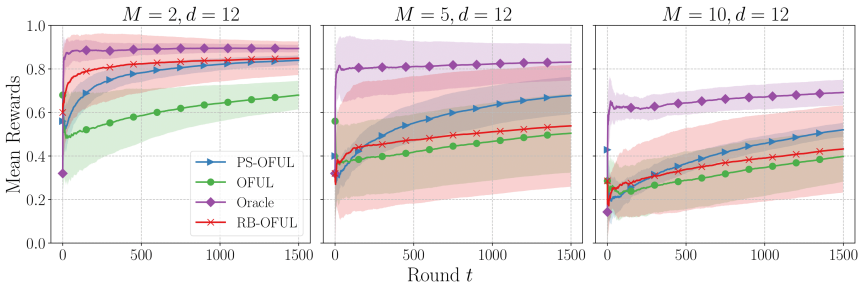

Figure 6.3: Parameter selection in CIFAR-10 with models less accurate than those in Figure 6.2 *(right)*. The results are averaged over 50 runs.

of models M with a regret balancing algorithm that uses SquareCB baselines (RB-SCB), and three SquareCB algorithms that use the best (Oracle), second-best (with test accuracy 84% for MNIST), and worst feature maps. The results in Figure 6.1 show that **1)** FS-SCB always performs between the best and second-best experts, **2)** the regret of FS-SCB that scales as $\sqrt{\log M}$ is close to RB-SCB (scales as $\sqrt{M}$) for small M, but gets much better as M grows, and **3)** RB-SCB has much higher variance than the other algorithms in MNIST.

Parameter Selection (Synthetic): We first sample the center of $M = 10$ balls from a $d = 20$-dimensional Gaussian, i.e., $\{\mu_i\}_{i=1}^{M} \sim \mathcal{N}(0, I_d)$, and set their radii to $b_i = b,\ \forall i \in [M]$. At the beginning of each LB task, we select a model $i_* \in [M]$ uniformly at random, and then sample the problem's parameter from its ball, i.e., $\theta_* \sim B(\mu_{i_*}, b_{i_*})$. The action set in each round $t \in [T]$ consists of 10 vectors $\{\phi_t(a_j)\}_{j=1}^{10} \sim \mathcal{N}(0, 0.01 I_d)$, and the reward of the selected action a_t is defined as $\langle \phi_t(a_t), \theta_* \rangle + \eta_t,\ \eta_t \sim \mathcal{U}[-0.5, 0.5]$. Figure 6.2 *(left)* compares the regret of our PS-OFUL algorithm with OFUL [42] for different sizes of the balls $b \in \{0, \sqrt{d}, 2\sqrt{d}\}$. We run OFUL with the upper-bounds $\|\theta_*\|_2 \leq S = 10$ and $S = \max_i(\mu_i + b_i)$ on the reward parameter. Note that the second bound is tighter and shows the best performance of OFUL. Our results indicate that the regret of PS-OFUL is better than OFUL, and gets closer to it as we increase the size of the balls from $b = 0$ to $b = 2\sqrt{d} \approx 9$. This clearly shows the potential advantage of

transfer (PS-OFUL) over individual (OFUL) learning.

Parameter Selection (CIFAR-10): We modify the EfficientNetV2-S network [152] by adding a layer of $d = 12$ neurons before the last layer and fine-tuning it on CIFAR-10 dataset. We then select this d-dimensional layer as our feature map ϕ. To define our M models (balls), we sample $100M$ datasets of size 500. For each dataset, we randomly select a class $s_* \in [M]$, assign reward 1 to images from s_* and 0 to other images, and fit a linear model to it to obtain a parameter vector. Finally, we fit a Gaussian mixture model with M components to these $100M$ parameter vectors and use the means and covariances of the resulting clusters as the center and radii of our M models (balls). At the beginning of each LB task, we select a class $s_* \in [M]$ uniformly at random. In each round $t \in [T]$, the learner is given an action set consists of 10 images, one from class s_* and the rest randomly selected from the other classes. The learner receives a reward from Ber(0.9), if it selects the image from class s_*, and from Ber(0.1), otherwise.

In Figure 6.2 *(right)*, we compare the mean reward of PS-OFUL for different values of M with a regret balancing algorithm that uses OFUL baselines (RB-OFUL) [147], OFUL (individual learning), and BIAS-OFUL [143] with bias being the center of the true model (Oracle). The results show **1)** the good performance of PS-OFUL, **2)** the performance of PS-OFUL gets better than RB-OFUL as M grows ($\sqrt{\log M}$ vs. $\sqrt{M}$ scaling), **3)** the large variance of RB-OFUL, especially in comparison to PS-OFUL, and finally **4)** the advantage of transfer (PS-OFUL) over individual (OFUL) learning.

In order to show the impact of the model accuracy (the accuracy of the center of the balls and their radii) on the performance of the algorithms, we defined a less accurate set of M models (balls) using $10M$ datasets of size 50 (as opposed to $100M$ datasets of size 500 used in the results reported in Figure 6.2 *(right)*). In Figure 6.3, we compare the mean reward of PS-OFUL for different number of models M with RB-OFUL, OFUL, and BIAS-OFUL. The results indicate that with decreasing in the accuracy of the models,

the performance of PS-OFUL and RB-OFUL get closer to that for OFUL.

6.6 Conclusions

We studied two model selection settings in LB, where the mean reward is linear in at least one of M models *(feature selection)*, and where the reward parameter is arbitrarily selected from M misspecified models *(parameter selection)*. We derived computationally efficient algorithms in these settings that are based on reductions from bandits to full-information problems, and proved regret bounds with desirable dependence on the horizon and number of models. An interesting future direction is to extend our results to the meta learning and learning-to-learn setting, where the agent starts with M models, and instead of solving a single LB problem, has to solve N of them one after another. [153, 144] are the results of this chapter.

Appendix A

Supplements to Chapter 2

A.1 Proof of Lemma 2.3.3

We prove this by contradiction. Suppose there is another optimal solution for problem (2.3.8) $(\mathbf{R}^\star, \boldsymbol{\lambda}^\star)$ in which for customers with traveling preference $\mathcal{G}_\ell$, station s_n has empty capacity, and assume that customers with type (i, j, ℓ) are assigned to stations s_m with $m > n$. However, we can have another set of routing probabilities such that for a small $0 \leq \epsilon$, $r_{i,j,\ell}^{(n)'} = r_{i,j,\ell}^{(n)\star} + \epsilon$, and $r_{i,j,\ell}^{(m)'} = r_{i,j,\ell}^{(m)\star} - \epsilon$ which is a feasible solution, and it will increase the objective function (2.3.8) due to the structure we found in lemma 2.3.2. Hence, it is contradictory to the optimality of this solution.

A.2 Proof of Theorem 2.3.4

We first assume that all the charging stations are used at full capacity, i.e., potential customers are more than the available capacity of charging stations. We need to show that the Algorithm 1 will find the optimal solution of problem (2.3.8). For convenience, denote as $f(.)$ the objective function of (2.3.8), and $g(.)$ as the resulting linear program of

problem (2.3.8) when we consider virtual station s_{Q+1} and we fix $\lambda_{i,j,\ell} = \Lambda_{i,j,\ell}$, $\forall(i,j,\ell)$. Assume the optimal solution of problem f to be $A^\star = \left([r^{(k)^\star}_{i,j,\ell}]_{k=1,\dots,\rho}, \lambda^\star_{i,j,\ell}\right)$, $\forall(i,j,\ell)$, and the optimal solution of linear program g to be $\hat{B}^\star = \left([h^{(k)^\star}_{i,j,\ell}]_{k=1,\dots,\rho,Q+1}, \lambda_{i,j,\ell} = \Lambda_{i,j,\ell}\right)$, $\forall(i,j,\ell)$. We define $\hat{A} = \left([r^{(k)'}_{i,j,\ell}]_{k=1,\dots,\rho,Q+1}, \lambda'_{i,j,\ell} = \Lambda_{i,j,\ell}\right)$, $\forall(i,j,\ell)$, such that:

$$[r^{(k)'}_{i,j,\ell}]_{k=1,\dots,\rho} = [r^{(k)^\star}_{i,j,\ell}]_{k=1,\dots,\rho}, \tag{A.2.1}$$

$$r^{(Q+1)'}_{i,j,\ell} = \frac{\Lambda_{i,j,\ell} - \lambda^\star_{i,j,\ell}}{\Lambda_{i,j,\ell}}, \tag{A.2.2}$$

and we define $B = \left([h^{(k)''}_{i,j,\ell}]_{k=1,\dots,\rho}, \lambda''_{i,j,\ell}\right)$, $\forall(i,j,\ell)$, such that:

$$[h^{(k)''}_{i,j,\ell}]_{k=1,\dots,\rho} = [h^{(k)^\star}_{i,j,\ell}]_{k=1,\dots,\rho}, \tag{A.2.3}$$

$$\lambda''_{i,j,\ell} = \Lambda_{i,j,\ell}(1 - h^{(Q+1)^\star}_{i,j,\ell}). \tag{A.2.4}$$

Therefore, $\hat{A}$ and B are in the feasible set of solutions of problems g and f, respectively. By the definition of optimality, we can write:

f	g
$A^\star$	$\hat{A}$
B	$\hat{B}^\star$

$$f(A^\star) + g(\hat{B}^\star) \geq g(\hat{A}) + f(B), \text{or} \tag{A.2.5}$$

$$\alpha = f(A^\star) - g(\hat{A}) \geq f(B) - g(\hat{B}^\star) = \beta, \tag{A.2.6}$$

where α is the negative effect of admitting all customers to the system and adding the virtual station s_{Q+1} on the problem 2.3.8 for optimal solution $A^\star$, and β is that of solution B. Hence, $\alpha > \beta$ is contradictory to the optimality of $A^\star$. Therefore, $\alpha = \beta$, which means

the solution structure B is the optimal solution for problem (2.3.8). Therefore, Algorithm 1 will propose the optimal solution of (2.3.8). Now consider the case where all charging stations are not used at full capacity, i.e., the potential customers are less than the available capacity of charging stations. As is it shown in lemma 2.3.3, in the optimal solution of problem (2.3.8), customers will be assigned to the charging stations starting from charging station s_1. The same structure holds in the case where Algorithm 1 adds a virtual station since the coefficients of decision variables will have the same structure as they have in (2.3.8). Therefore, if the available capacity of charging stations is more than the potential set of customers can use, Algorithm 1 will not send any customers to the station s_{Q+1}, and in the optimal solution of problem (2.3.8) all customers will be admitted to the system in full.

A.3 Proof of Proposition 2.4.1

We know that

$$\begin{aligned} &P_{i+1,j,\ell} - P_{i,j,\ell} = v_{i+1}(W_{i,j,\ell} - W_{i+1,j,\ell}) \\ &v_{i+1}(W_{i,j,\ell} - W_{i+1,j,\ell}) \geq v_i(W_{i,j,\ell} - W_{i+1,j,\ell}), \end{aligned} \tag{A.3.1}$$

and hence, we can conclude that

$$P_{i+1,j,\ell} - P_{i,j,\ell} \geq v_i(W_{i,j,\ell} - W_{i+1,j,\ell}), \tag{A.3.2}$$

which satisfies the vertical IC constraints using Lemma (2.2.2). For proving Horizontal IC, we know from (2.4.7) that $P_{i,j+1,\ell} - P_{i,j,\ell} = v_i(W_{i,j,\ell} - W_{i,j+1,\ell})$, which satisfies the condition stated in Lemma (2.2.2) for Horizontal IC. For proving (2.2.6), we need to

show that $P_{i,j,\ell} + v_i W_{i,j,\ell} \leq P_{i,j,m} + v_i W_{i,j,m}$ if $m \in \mathcal{B}_\ell$ that we can get with considering constraint (2.4.5) and equations (2.4.6)-(2.4.8). We prove IR by induction for customers with traveling preference $\mathcal{G}_\ell$. We know that IR requires that $P_{i,j,\ell} \leq R_i - v_i W_{i,j,\ell}$. Starting with $i = 1$ we have $P_{1,j,\ell} = R_1 - v_1 W_{1,j,\ell}$. Now, assume that IR holds for type (i, j, ℓ). For type $(i+1, j, \ell)$, we can write $P_{i+1,j,\ell} = \big(P_{i,j,\ell} + v_{i+1} W_{i,j,\ell} - v_{i+1} W_{i+1,j,\ell}\big) \leq \big(R_i - v_i W_{i,j,\ell} + v_{i+1} W_{i,j,\ell} - v_{i+1} W_{i+1,j,\ell}\big)$. Also, we know that $W_{i+1,j,\ell} \leq W_{i,j,\ell} \leq \frac{R_i}{v_i} \leq \big(\frac{R_{i+1}-R_i}{v_{i+1}-v_i}\big)$, which leads to $P_{i+1,j,\ell} \leq \left(R_i + (v_{i+1} - v_i)\frac{R_{i+1}-R_i}{v_{i+1}-v_i} - v_{i+1} W_{i+1,j,\ell}\right)$. Accordingly, $P_{i+1,j,\ell} \leq R_i + R_{i+1} - R_i - v_{i+1} W_{i+1,j,\ell} = R_{i+1} - v_{i+1} W_{i+1,j,\ell}$, which concludes that: $P_{i+1,j,\ell} \leq R_{i+1} - v_{i+1} W_{i+1,j,\ell}$. This proves IR.

Appendix B

Supplements to Chapter 4

B.1 Proof of Lemma 4.4.2

We first state the standard results that plays an important role in most proofs for linear bandits problems.

Proposition B.1.1 *(from[42]) Let $\lambda \geq 1$. For any arbitrary sequence of actions $(x_1, \ldots, x_t) \in \mathcal{D}^t$, let V_t be the corresponding Gram matrix, then*

$$\sum_{s=1}^{t} \|x_s\|^2_{V_s^{-1}} \leq 2 \log \frac{det(V_{t+1})}{det(\lambda I)} \leq 2d \log (1 + \frac{tL^2}{\lambda}). \tag{B.1.1}$$

In particular, we have

$$\begin{aligned} \sum_{s=1}^{T} \|x_s\|_{V_s^{-1}} &\leq \sqrt{T} \left(\sum_{s=1}^{T} \|x_s\|^2_{V_s^{-1}} \right)^{\frac{1}{2}} \\ &\leq \sqrt{2Td \log \left(1 + \frac{TL^2}{\lambda} \right)}. \end{aligned} \tag{B.1.2}$$

Also, we recall the Azuma's concentration inequality for super-martingales.

Proposition B.1.2 *(Azuma's inequality [154]) If a super-martingale $(Y_t)_{t\geq 0}$ corresponding to a filtration $\mathcal{F}_t$ satisfies $|Y_t - Y_{t-1}| < c_t$ for some positive constant c_t, for all $t = 1, \ldots, T$, then, for any $u > 0$,*

$$\mathbb{P}\left(Y_T - Y_0 \geq u\right) \leq 2e^{-\frac{u^2}{2\sum_{t=1}^{T} c_t^2}}. \tag{B.1.3}$$

Next, we define the high probability confidence regions for the RLS-estimates that er use in the rest of the proof.

Definition B.1.3 *Let $\delta \in (0,1)$, $\delta' = \frac{\delta}{6T}$, and $t \in [T]$. We define the following events:*

- *$\hat{E}_t$ is the event that the RLS-estimate $\hat{\theta}$ concentrates around $\theta_\star$ for all steps $s \leq t$, i.e., $\hat{E}_t = \{\forall s \leq t, \left\|\hat{\theta}_s - \theta_\star\right\|_{V_s} \leq \beta_s(\delta')\}$;*
- *$\hat{Z}_t$ is the event that the RLS-estimate $\hat{\mu}$ concentrates around $\mu_\star$, i.e., $\hat{Z}_t = \{\forall s \leq t, \|\hat{\mu}_s - \mu_\star\|_{V_s} \leq \beta_s(\delta')\}$. Moreover, define Z_t such that*

$$Z_t = \hat{E}_t \cap \hat{Z}_t.$$

- *$\tilde{E}_t$ is the event that the sampled parameter $\tilde{\theta}_t$ concentrates around $\hat{\theta}_t$ for all steps $s \leq t$, i.e., $\tilde{E}_t = \{\forall s \leq t, \left\|\tilde{\theta}_s - \hat{\theta}_s\right\|_{V_s} \leq \gamma_s(\delta')\}$. Let E_t be such that $E_t = \tilde{E}_t \cap Z_t$.*

Lemma B.1.4 *Under Assumptions 6, 7, we have $\mathbb{P}(Z) = \mathbb{P}(\hat{E} \cap \hat{Z}) \geq 1 - \frac{\delta}{3}$ where $\hat{E} = \hat{E}_T \subset \cdots \subset \hat{E}_1$, and $\hat{Z} = \hat{Z}_T \subset \cdots \subset \hat{Z}_1$.*

Proof: The proof is similar to the one in Lemma 1 of [43] and is ommited for brevity. ■

Lemma B.1.5 *Under Assumptions 6, 7, we have $\mathbb{P}(E) = \mathbb{P}(\tilde{E} \cap Z) \geq 1 - \frac{\delta}{2}$, where $\tilde{E} = \tilde{E}_T \subset \cdots \subset \tilde{E}_1$.*

Proof: We show that $\mathbb{P}(\tilde{E}) \geq 1 - \frac{\delta}{6}$. Then, from Lemma B.1.4 we know that $\mathbb{P}(Z) \geq 1 - \frac{\delta}{3}$, thus we can conclude that $\mathbb{P}(E) \geq 1 - \frac{\delta}{2}$. Bounding $\tilde{E}$ comes directly from concentration inequality (4.3.9). Specifically, for $1 \leq t \leq T$

$$\mathbb{P}\left(\left\|\tilde{\theta}_t - \hat{\theta}_t\right\|_{V_t} \leq \gamma_t(\delta')\right) = \mathbb{P}\left(\|\eta_t\|_2 \leq \frac{\gamma_t(\delta')}{\beta_t(\delta')}\right)$$
$$= \mathbb{P}\left(\|\eta_t\|_2 \leq \left(1 + \frac{2}{C}LS\right)\sqrt{cd\log\left(\frac{c'd}{\delta'}\right)}\right) \geq 1 - \delta'.$$

Applying union bound on this ensures that $\mathbb{P}(\tilde{E}) \geq 1 - T\delta' = 1 - \frac{\delta}{6}$. ■

Now we are ready to provide the formal proof of Lemma 4.4.2. First, we provide a formal statement and a detailed proof of Lemma 4.4.2. Here, we need several modifications compared to [43] that are required because in our setting, actions x_t belong to inner approximations of the true safe set $\mathcal{D}_0^s$. Moreover, we follow an algebraic treatment that is perhaps simpler compared to the geometric viewpoint in [43].

Lemma B.1.6 *Let* $\Theta_t^{\text{opt}} = \{\theta \in \mathbb{R}^d : \max_{x \in \mathcal{D}_t^s} x^\top \theta \geq x_\star^\top \theta_\star\} \cap \mathcal{E}_t^{TS}$ *be the set of optimistic parameters,* $\tilde{\theta}_t = \hat{\theta}_t + \beta_t(\delta')V_t^{-\frac{1}{2}}\eta_t$ *with* $\eta_t \sim \mathcal{D}^{TS}$*, then* $\forall t \geq 1$, $\mathbb{P}\left(\tilde{\theta}_t \in \Theta_t^{\text{opt}} | \mathcal{F}_t, Z_t\right) \geq \frac{p}{2}$.

Proof: First, we provide the shrunk version $\tilde{\mathcal{D}}_t^s$ of $\mathcal{D}_t^s$ as follows:

A shrunk safe decision set $\tilde{\mathcal{D}}_t^s$. Consider the enlarged confidence region $\tilde{\mathcal{C}}_t$ centered at $\mu_\star$ as

$$\tilde{\mathcal{C}}_t := \{v \in \mathbb{R}^d : \|v - \mu_\star\|_{V_t} \leq 2\beta_t(\delta')\}. \tag{B.1.4}$$

We know that $\mathcal{C}_t \subseteq \tilde{\mathcal{C}}_t$, since $\forall v \in \mathcal{C}_t$, we know that $\|v - \mu_\star\|_{V_t} \leq \|v - \hat{\mu}_t\|_{V_t} + \|\hat{\mu}_t - \mu_\star\|_{V_t} \leq 2\beta(t)$. From the definition of enlarged confidence region, we can get the following defini-

tion for shrunk safe decision set:

$$\begin{aligned}\tilde{\mathcal{D}}_t^s :=& \{x \in \mathcal{D}_0 : x^\top v \leq C, \forall v \in \tilde{\mathcal{C}}_t\} \\ &= \{x \in \mathcal{D}_0 : \max_{v \in \tilde{\mathcal{C}}_t} x^\top v \leq C\} \\ &= \{x \in \mathcal{D}_0 : x^\top \mu_\star + 2\beta_t(\delta') \|x\|_{V_t^{-1}} \leq C\}, \end{aligned} \tag{B.1.5}$$

and note that $\tilde{\mathcal{D}}_t^s \subseteq \mathcal{D}_t^s$, and they are not empty, since they include zero due to Assumption 8.

Then, we define the parameter α_t such that the vector $z_t = \alpha_t x_\star$ in direction $x_\star$ belongs to $\tilde{\mathcal{D}}_t^s$ and is closest to $x_\star$. Hence, we have:

$$\alpha_t := \max \left\{ \alpha \in [0,1] : z_t = \alpha x_\star \in \tilde{\mathcal{D}}_t^s \right\}. \tag{B.1.6}$$

Since $\mathcal{D}_0$ is convex by Assumption 8 and both $0, x_\star \in \mathcal{D}_0$, we have

$$\alpha_t = \max \left\{ \alpha \in [0,1] : \alpha \left(x_\star^\top \mu_\star + 2\beta_t(\delta') \|x_\star\|_{V_t^{-1}} \right) \leq C \right\}. \tag{B.1.7}$$

From constraint (4.2.1), we know that $x_\star^\top \mu_\star \leq C$. We choose α_t such that

$$1 + \frac{2}{C} \beta_t(\delta') \|x_\star\|_{V_t^{-1}} = \frac{1}{\alpha_t}. \tag{B.1.8}$$

We need to study the probability that a sampled $\tilde{\theta}_t$ drawn from $\mathcal{H}^{\mathrm{TS}}$ distribution at round t is optimistic, i.e.,

$$p_t = \mathbb{P}\left((x_t(\tilde{\theta}_t))^\top \tilde{\theta}_t \geq x_\star^\top \theta_\star \;\middle|\; \mathcal{F}_t, Z_t \right).$$

Using the definition of α_t in (B.1.7), we have

$$(x_t(\tilde{\theta}_t))^\top \tilde{\theta}_t = \max_{x \in \mathcal{D}_t^s} x^\top \tilde{\theta}_t \geq \alpha_t x_\star^\top \tilde{\theta}_t. \tag{B.1.9}$$

Hence, we can write

$$\begin{aligned} p_t \geq & \mathbb{P}\left(\alpha_t x_\star^\top \tilde{\theta}_t \geq x_\star^\top \theta_\star \;\middle|\; \mathcal{F}_t, Z_t\right) \\ = & \mathbb{P}\left(x_\star^\top \left(\hat{\theta}_t + \beta_t(\delta') V_t^{-\frac{1}{2}} \eta_t\right) \geq \frac{x_\star^\top \theta_\star}{\alpha_t} \;\middle|\; \mathcal{F}_t, Z_t\right) \end{aligned}$$

Then, we use the value that we chose for α_t in (B.1.8), and we have

$$\begin{aligned} = \mathbb{P}\Bigg(& x_\star^\top \hat{\theta}_t + \beta_t(\delta') x_\star^\top V_t^{-\frac{1}{2}} \eta_t \geq \\ & x_\star^\top \theta_\star + \frac{2}{C} \beta_t(\delta') \left\|x_\star\right\|_{V_t^{-1}} x_\star^\top \theta_\star | \mathcal{F}_t, Z_t \Bigg) \end{aligned}$$

we know that $|x_\star^\top \theta_\star| \leq \|x_\star\|_2 \|\theta_\star\|_2 \leq LS$. Hence,

$$\begin{aligned} p_t \geq \mathbb{P}\Bigg(& \beta_t(\delta') x_\star^\top V_t^{-\frac{1}{2}} \eta_t \geq \\ & x_\star^\top (\theta_\star - \hat{\theta}_t) + \frac{2}{C} LS \beta_t(\delta') \left\|x_\star\right\|_{V_t^{-1}} \mid \mathcal{F}_t, Z_t \Bigg) \end{aligned}$$

From Cauchy–Schwarz inequality and (4.3.2), we have

$$\left| x_\star^\top \left(\theta_\star - \hat{\theta}_t\right) \right| \leq \left\|x_\star\right\|_{V_t^{-1}} \left\|\theta_\star - \hat{\theta}_t\right\|_{V_t} \leq \beta_t(\delta') \left\|x_\star\right\|_{V_t^{-1}}.$$

Therefore, we can write

$$p_t \geq \mathbb{P}\Bigg(x_\star^\top V_t^{-\frac{1}{2}} \eta_t \geq \\ \|x_\star\|_{V_t^{-1}} + \frac{2}{C} LS \|x_\star\|_{V_t^{-1}} \mid \mathcal{F}_t, Z_t \Bigg) \tag{B.1.10}$$

We define $u^\top = \frac{x_\star^\top V_t^{-\frac{1}{2}}}{\|x_\star\|_{V_t^{-1}}}$, and hence $\|u\|_2 = 1$. It follows from (B.1.10) that

$$p_t \geq \mathbb{P}\left(u^\top \eta_t \geq 1 + \frac{2}{C} LS \right) \geq p, \tag{B.1.11}$$

where the last inequality follows the concentration inequality (4.3.9) of the TS distribution. We also need to show that the high probability concentration inequality event does not effect the TS of being optimistic. This is because the chosen confidence bound $\delta' = \frac{\delta}{6T}$ is small enough compared to the anti-concentration property (4.3.8). Moreover, we assume that $T \geq \frac{1}{3p}$ which implies that $\delta' \leq \frac{p}{2}$. We know that for any events A ans B, we have

$$\mathbb{P}(A \cap B) = 1 - \mathbb{P}(A^c \cup B^c) \geq \mathbb{P}(A) - \mathbb{P}(B^c). \tag{B.1.12}$$

We apply (B.1.12) with $A = \{J_t(\tilde{\theta}_t) \geq J(\theta_\star)\}$ and $B = \{\tilde{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}\}$ which leads to

$$\mathbb{P}\left(\tilde{\theta}_t \in \Theta_t^{\mathrm{opt}} \;\middle|\; \mathcal{F}_t, Z_t \right) \geq p - \delta' \geq \frac{p}{2}.$$

■

B.2 Proof of Theorem 4.4.1

The proof presented below follows closely the proof of [43] and is primarily presented here for completeness. Specifically, we have identified that the only critical change that needs to be made to account for safety is the proof of actions being frequently optimistic in the face of constraints thanks to the modified anti-concentration property 4.3.8. This was handled in the previous section B.1. For completeness, we also prove in Lemma B.2.1 that the first action of Safe-LTS is always safe under our assumptions.

We use the following decomposition for bounding the regret:

$$R(T) \leq \sum_{t=1}^{T} \left(x_\star^\top \theta_\star - x_t \theta_\star\right) \mathbb{1}\{E_t\} = \sum_{t=1}^{T} \left(\underbrace{x_\star^\top \theta_\star - x_t^\top \tilde{\theta}_t}_{\text{Term I}} \right) \mathbb{1}\{E_t\} + \sum_{t=1}^{T} \left(\underbrace{x_t^\top \tilde{\theta}_t - x_t^\top \theta_\star}_{\text{Term II}} \right) \mathbb{1}\{E_t\}. \tag{B.2.1}$$

B.2.1 Bounding Term I.

For any θ, we denote $x_t(\theta) = \arg\max_{x \in \mathcal{D}_t^s} x^\top \theta$. On the event E_t, $\tilde{\theta}_t$ belongs to $\mathcal{E}_t^{\mathrm{TS}}$ which leads to

$$\begin{aligned} (\text{Term I}) \mathbb{1}\{E_t\} &:= R_t^{\mathrm{TS}} \mathbb{1}\{E_t\} \\ &\leq \left(x_\star^\top \theta_\star - \inf_{\theta \in \mathcal{E}_t^{\mathrm{TS}}} (x_t(\theta))^\top \theta \right) \mathbb{1}\{Z_t\}. \end{aligned} \tag{B.2.2}$$

Here and onwards, we use $\mathbb{1}\{\mathcal{E}\}$ as the indicator function applied to an event $\mathcal{E}$. We have also used the fact that E_t is a subset of Z_t. Next, we can also bound (B.2.2) by the expectation over any random choice of $\tilde{\theta} \in \Theta_t^{\mathrm{opt}}$ (recall (4.4.6)) that leads to

$$R_t^{\mathrm{TS}} \leq \mathbb{E}\left[\left((x_t(\tilde{\theta}))^\top \tilde{\theta} - \inf_{\theta \in \mathcal{E}_t^{\mathrm{TS}}} (x_t(\theta))^\top \theta \right) \mathbb{1}\{Z_t\} \,\middle|\, \mathcal{F}_t, \tilde{\theta} \in \Theta_t^{\mathrm{opt}} \right].$$

Equivalently, we can write

$$R_t^{\mathrm{TS}} \leq \mathbb{E}\left[\sup_{\theta\in\mathcal{E}_t^{\mathrm{TS}}}\left(\left(x_t(\tilde{\theta})\right)^\top\tilde{\theta} - \left(x_t(\theta)\right)^\top\theta\right)\mathbb{1}\{Z_t\}\,\middle|\,\mathcal{F}_t, \tilde{\theta}\in\Theta_t^{opt}\right], \tag{B.2.3}$$

Then, using Cauchy–Schwarz and the definition of $\gamma_t(\delta')$ in (4.4.10)

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\theta\in\mathcal{E}_t^{\mathrm{TS}}}\left(x_t(\tilde{\theta})\right)^\top\left(\tilde{\theta}-\theta\right)\mathbb{1}\{Z_t\}\,\middle|\,\mathcal{F}_t,\tilde{\theta}\in\Theta_t^{opt}\right]\\
&\leq \mathbb{E}\left[\left\|x_t(\tilde{\theta})\right\|_{V_t^{-1}}\sup_{\theta\in\mathcal{E}_t^{\mathrm{TS}}}\left\|\tilde{\theta}-\theta\right\|_{V_t}\,\middle|\,\mathcal{F}_t,\tilde{\theta}\in\Theta_t^{opt},Z_t\right]\mathbb{P}(Z_t)\\
&\leq 2\gamma_t(\delta')\mathbb{E}\left[\left\|x_t(\tilde{\theta})\right\|_{V_t^{-1}}\,\middle|\,\mathcal{F}_t,\tilde{\theta}\in\Theta_t^{opt},Z_t\right]\mathbb{P}(Z_t).
\end{aligned}$$

This property shows that the regret R_t^{TS} is upper bounded by V_t^{-1}-norm of the optimal safe action corresponding to the any optimistic parameter $\tilde{\theta}$. Hence, we need to show that TS samples from the optimistic set with high frequency. We prove in Lemma B.1.6 that TS is optimistic with a fixed probability $(\frac{p}{2})$ which leads to bounding R_t^{TS} as follows:

$$R_t^{\mathrm{TS}}\frac{p}{2} \leq 2\gamma_t(\delta')\mathbb{E}\left[\left\|x_t(\tilde{\theta}_t)\right\|_{V_t^{-1}}\,\middle|\,\mathcal{F}_t,\tilde{\theta}_t\in\Theta_t^{opt},Z_t\right]\mathbb{P}(Z_t)\frac{p}{2} \leq \tag{B.2.4}$$

$$\begin{aligned}
&2\gamma_t(\delta')\mathbb{E}\left[\left\|x_t(\tilde{\theta}_t)\right\|_{V_t^{-1}}\,\middle|\,\mathcal{F}_t,\tilde{\theta}_t\in\Theta_t^{opt},Z_t\right]\mathbb{P}(Z_t)\mathbb{P}\left(\tilde{\theta}_t\in\Theta_t^{\mathrm{opt}}\,\middle|\,\mathcal{F}_t,Z_t\right)\\
&\leq 2\gamma_t(\delta')\mathbb{E}\left[\left\|x_t(\tilde{\theta}_t)\right\|_{V_t^{-1}}\,\middle|\,\mathcal{F}_t,Z_t\right]\mathbb{P}(Z_t).
\end{aligned} \tag{B.2.5}$$

By reintegrating over the event Z_t we get

$$R_t^{\mathrm{TS}} \leq \frac{4\gamma_t(\delta')}{p}\mathbb{E}\left[\left\|x_t(\tilde{\theta}_t)\right\|_{V_t^{-1}}\mathbb{1}\{Z_t\}\,\middle|\,\mathcal{F}_t\right]. \tag{B.2.6}$$

Recall that $E_t \subset Z_t$, hence

$$R^{\mathrm{TS}}(T) \leq \sum_{t=1}^{T} R_t^{\mathrm{TS}} \mathbb{1}\{E_t\} \leq \frac{4\gamma_T(\delta')}{p} \sum_{t=1}^{T} \mathbb{E}\left[\left\| x_t(\tilde{\theta}_t) \right\|_{V_t^{-1}} \,\middle|\, \mathcal{F}_t \right]. \tag{B.2.7}$$

For bounding this term, we rewrite the RHS above as:

$$R^{\mathrm{TS}}(T) \leq \sum_{t=1}^{T} \|x_t\|_{V_t^{-1}} + \sum_{t=1}^{T} \left(\mathbb{E}\left[\left\| x_t(\tilde{\theta}_t) \right\|_{V_t^{-1}} \,\middle|\, \mathcal{F}_t \right] - \|x_t\|_{V_t^{-1}} \right). \tag{B.2.8}$$

We can now bound the first expression using Proposition B.1.1. For the second expression we proceed as follows:

- First, the sequence

$$Y_t = \sum_{s=1}^{t} \left(\mathbb{E}\left[\left\| x_s(\tilde{\theta}_s) \right\|_{V_s^{-1}} \,\middle|\, \mathcal{F}_s \right] - \|x_s\|_{V_s^{-1}} \right)$$

  is a martingale by construction.

- Second, under Assumption 8, $\|x_t\|_2 \leq L$, and since $V_t^{-1} \leq \frac{1}{\lambda} I$, we can write

$$\mathbb{E}\left[\left\| x_s(\tilde{\theta}_s) \right\|_{V_s^{-1}} \,\middle|\, \mathcal{F}_s \right] - \|x_s\|_{V_s^{-1}} \leq \frac{2L}{\sqrt{\lambda}}, \forall t \geq 1. \tag{B.2.9}$$

- Third, for bounding Y_T, we use Azuma's inequality, and we have that with probability $1 - \frac{\delta}{2}$,

$$Y_T \leq \sqrt{\frac{8TL^2}{\lambda} \log \frac{4}{\delta}}. \tag{B.2.10}$$

Putting these together, we conclude that with probability $1 - \frac{\delta}{2}$,

$$R^{\mathrm{TS}}(T) \leq \frac{4\gamma_T(\delta')}{p}\left(\sqrt{2Td\log\left(1+\frac{TL^2}{\lambda}\right)} + \sqrt{\frac{8TL^2}{\lambda}\log\frac{4}{\delta}}\right).$$

B.2.2 Bounding Term II

We can bound on Term II using the general result of [42]. In fact, we can use the following general decomposition:

$$\begin{aligned}
&\sum_{t=1}^{T}(\text{Term II})\mathbb{1}\{E_t\} := R^{\mathrm{RLS}}(T) \\
&= \sum_{t=1}^{T}\left(x_t^\top\tilde{\theta}_t - x_t^\top\theta_\star\right)\mathbb{1}\{E_t\} \\
&\leq \sum_{t=1}^{T} \mid x_t^\top(\tilde{\theta}_t - \hat{\theta}_t) \mid \mathbb{1}\{E_t\} + \sum_{t=1}^{T} \mid x_t^\top(\hat{\theta}_t - \theta_\star) \mid \mathbb{1}\{E_t\}. \qquad \text{(B.2.11)}
\end{aligned}$$

By Definition B.1.3, we have $E_t \subseteq Z_t$ and $E_t \subseteq \tilde{E}_t$, and hence

$$\mid x_t^\top(\tilde{\theta}_t - \hat{\theta}_t) \mid \mathbb{1}\{E_t\} \leq \|x\|_{V_t^{-1}}\gamma_t(\delta')$$

$$\mid x_t^\top(\hat{\theta}_t - \theta_\star) \mid \mathbb{1}\{E_t\} \leq \|x\|_{V_t^{-1}}\beta_t(\delta').$$

Therefore, from Proposition B.1.1, we have with probability $1 - \frac{\delta}{2}$

$$R^{\mathrm{RLS}}(T) \leq (\beta_T(\delta') + \gamma_T(\delta'))\sqrt{2Td\log\left(1+\frac{TL^2}{\lambda}\right)}. \qquad \text{(B.2.12)}$$

B.2.3 Overall Regret Bound

Recall that from (4.4.2), $R(T) \leq R^{TS}(T) + R^{\mathrm{RLS}}(T)$. As shown previously, each term is bounded separately with probability $1 - \frac{\delta}{2}$. Using union bound over two terms, we get

the following expression:

$$R(T) \leq (\beta_T(\delta') + \gamma_T(\delta')(1 + \frac{4}{p}))\sqrt{2Td\log(1 + \frac{TL^2}{\lambda})} + \frac{4\gamma_T(\delta')}{p}\sqrt{\frac{8TL^2}{\lambda}\log\frac{4}{\delta}}, \tag{B.2.13}$$

holds with probability $1-\delta$ where $\delta' = \frac{\delta}{6T}$. For completeness we show below that action x_1 is safe. Having established that, it follows that the rest of the actions $x_t, t > 1$ are also safe with probability at least $1 - \delta'$. This is by construction of the feasible sets $\mathcal{D}_t^s$ and by the fact that $\mu_\star \in \mathcal{C}_t(\delta')$ with the same probability for each t.

Lemma B.2.1 *The first action that Safe-LTS chooses is safe, that is* $x_1^\top \mu_\star \leq C$.

Proof: At round $t = 1$, the RLS-estimate $\hat{\mu}_1 = 0$ and $V_1 = \lambda I$. Thus, Safe-LTS chooses the action which maximizes the expected reward while satisfying $x_1^\top \hat{\mu}_1 + \beta_1(\delta') \|x_1\|_{V_1^{-1}} \leq C$. Hence, x_1 satisfies:

$$\beta_1(\delta') \|x_1\|_{V_1^{-1}} \leq C.$$

From Theorem 4.3.1 and $V_1^{-1} = (1/\lambda)I$ leads to $S\|x_1\|_2 \leq C$ which completes the proof. ■

Appendix C

Supplements to Chapter 5

C.1 Proof of Proposition 5.4.1

From the definition of regret, we can write

$$\begin{aligned} R(T) &= \sum_{t \in N_T} \left(\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle \right) + \sum_{t \in N_T^c} \left(\langle x_\star, \theta_\star \rangle - \langle (1-\rho_1) x_{b_t} - \rho_1 \zeta_t, \theta_\star \rangle \right) \\ &= \sum_{t \in N_T} \left(\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle \right) + \sum_{t \in N_T^c} \Big(\langle x_\star, \theta_\star \rangle - \langle x_{b_t}, \theta_\star \rangle + \rho_1 \langle x_{b_t}, \theta_\star \rangle + \rho_1 \langle \zeta_t, \theta_\star \rangle \Big) \\ &\leq \sum_{t \in N_T} \left(\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle \right) + |N_T^c| \left(\kappa_h + \rho_1 (r_h + S) \right). \end{aligned} \tag{C.1.1}$$

C.2 Proof of Lemma 5.3.2

In order to ensure that the conservative action $x_t = (1-\rho) x_{b_t} + \rho \zeta_t$ is safe, we need to show that it satisfies (5.2.2). Hence, it suffices to show that

$$\langle (1-\rho) x_{b_t} + \rho \zeta_t, \theta_\star \rangle \geq (1-\alpha) r_{b_t}. \tag{C.2.1}$$

We can lower bound the LHS of (C.2.1) as follows:

$$\langle (1-\rho)x_{b_t} + \rho\zeta_t, \theta_\star \rangle = r_{b_t} - \rho r_{b_t} + \rho\langle \zeta_t, \theta_\star \rangle \geq r_{b_t} - \rho r_{b_t} - \rho S.$$

Recall that $\|\zeta_t\|_2 = 1$ almost surely, and due to Assumption 10, we know that $\|\theta_\star\|_2 \leq S$. Hence, it suffices to show that

$$r_{b_t} - \rho r_{b_t} - \rho S \geq (1-\alpha) r_{b_t},$$

or equivalently,

$$\rho r_{b_t} + \rho S \leq \alpha r_{b_t} \tag{C.2.2}$$

From (C.2.2) we can write

$$\rho \leq \frac{\alpha r_{b_t}}{S + r_{b_t}}. \tag{C.2.3}$$

Therefore, for any ρ satisfying (C.2.3), the conservative action $x_t = (1-\rho)x_{b_t} + \rho\zeta_t$ is guaranteed to be safe almost surely. Then, we lower bound the right hand side of (C.2.3) using Assumption 12, and we establish the following upper bound on ρ,

$$\rho \leq \frac{\alpha r_l}{S + r_h}. \tag{C.2.4}$$

Therefore, for any $\rho \in (0, \bar{\rho})$, where $\bar{\rho} = \frac{\alpha r_l}{S + r_h}$, the conservative actions are safe.

C.3 Proof of Theorem 5.4.2

In this section, we provide an upper bound on the regret of Term I in (5.4.1). We first rewrite Term I as follows:

$$\sum_{t \in N_T} \left(\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle \right) \tag{C.3.1}$$

Clearly, it would be beneficial to show that (C.3.1) is non-positive. However, as stated in [43] (in the case of linear TS applied to the standard stochastic linear bandit problem with no safety constraints), this cannot be the case in general. Instead, to bound regret in the unconstrained case, [43] argues that it suffices to show that (C.3.1) is non-positive with a constant probability. But what happens in the safety-constrained scenario? It turns out that once the above stated event happens with constant probability (in our case, in the presence of safety constraints), the rest of the argument by [43] remains unaltered. Therefore, our main contribution in the proof of Theorem 5.4.2 is to show that (C.3.1) is non-positive with a constant probability in spite of the limitations on actions imposed because of the safety constraints. To do so, let

$$\Theta_t^{\text{opt}} = \{\theta \in \mathbb{R}^d : \langle x(\theta), \theta \rangle \geq \langle x_\star, \theta_\star \rangle\}, \tag{C.3.2}$$

be the so-called *set of optimistic parameters*, where $x(\tilde{\theta}_t) = \arg\max_{x \in \mathcal{X}_t^s} \langle x, \tilde{\theta}_t \rangle$ is the optimal safe action for the sampled parameter $\tilde{\theta}_t$ chosen from the estimated safe action set $\mathcal{X}_t^s$. LTS is considered optimistic at round t, if it samples the parameter $\tilde{\theta}_t$ from the set of optimistic parameters Θ_t^{opt} and plays the action $x(\tilde{\theta}_t)$. In Lemma C.3.2, we show that SCLTS is optimistic with constant probability despite the safety constraints. Before that, let us restate the distributional properties put forth in [43] for the noise $\eta \sim \mathcal{H}^{\text{TS}}$ that are required to ensure the right balance of exploration and exploitation.

Definition C.3.1 *(Definition 1. in [43]) $\mathcal{H}^{TS}$ is a multivariate distribution on $\mathbb{R}^d$ absolutely continuous with respect to the Lebesgue measure which satisfies the following properties:*

- *(anti-concentration) there exists a strictly positive probability p such that for any $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$,*

$$\mathbb{P}_{\eta \sim \mathcal{H}^{TS}} \left(\langle u, \eta \rangle \geq 1 \right) \geq p. \tag{C.3.3}$$

- *(concentration) there exists positive constants c, c' such that $\forall \delta \in (0,1)$*

$$\mathbb{P}_{\eta \sim \mathcal{H}^{TS}} \left(\|\eta\| \leq \sqrt{cd \log(\frac{c'd}{\delta})} \right) \geq 1 - \delta. \tag{C.3.4}$$

Lemma C.3.2 *Let $\Theta_t^{opt} = \{\theta \in \mathbb{R}^d : \langle x(\theta), \theta \rangle \geq \langle x_\star, \theta_\star \rangle\}$ be the set of the optimistic parameters. For round $t \in N_T$, SCLTS samples the optimistic parameter $\tilde{\theta}_t \in \Theta_t^{opt}$ and plays the corresponding safe action $x(\tilde{\theta}_t)$ frequently enough, i.e.,*

$$\mathbb{P}(\tilde{\theta}_t \in \Theta_t^{opt}) \geq p. \tag{C.3.5}$$

Proof: We need to show that for rounds $t \in N_T$

$$\mathbb{P}\left(\langle x(\tilde{\theta}_t), \tilde{\theta}_t \rangle \geq \langle x_\star, \theta_\star \rangle \right) \geq p. \tag{C.3.6}$$

First, we show that for rounds $t \in N_T$, $x_\star$ falls in the estimated safe set, i.e., $x_\star \in \mathcal{X}_t^s$. To do so, we need to show that

$$\langle x_\star, \hat{\theta}_t \rangle - \beta_t \|x_\star\|_{V_t^{-1}} \geq (1 - \alpha) r_{b_t}, \tag{C.3.7}$$

using $\|\theta_\star - \hat{\theta}_t\|_{V_t} \leq \beta_t$, it suffices that

$$\langle x_\star, \theta_\star \rangle - 2\beta_t \|x_\star\|_{V_t^{-1}} \geq (1-\alpha) r_{b_t}. \tag{C.3.8}$$

But we know that $\|x_\star\|_{V_t^{-1}} \leq \frac{\|x_\star\|_2}{\sqrt{\lambda_{\min}(V_t)}} \leq \frac{L}{\sqrt{\lambda_{\min}(V_t)}}$, where we also used Assumption 11 to bound $\|x_\star\|_2$. Hence, we can get

$$\langle x_\star, \theta_\star \rangle - 2\beta_t \|x_\star\|_{V_t^{-1}} \geq \langle x_\star, \theta_\star \rangle - \frac{2\beta_t L}{\sqrt{\lambda_{\min}(V_t)}}. \tag{C.3.9}$$

By substituting (C.3.9) in (C.3.8), it suffices to show that

$$\kappa_{b_t} + \alpha r_{b_t} \geq \frac{2\beta_t L}{\sqrt{\lambda_{\min}(V_t)}}, \tag{C.3.10}$$

or equivalently,

$$\lambda_{\min}(V_t) \geq \left(\frac{2L\beta_t}{\kappa_t + \alpha r_{b_t}} \right)^2. \tag{C.3.11}$$

To show (C.3.11), simply recall that $\lambda_{\min}(V_t) \geq k_t^1$, where $k_t^1 = \left(\frac{2L\beta_t}{\kappa_l + \alpha r_l}\right)^2$. Therefore, $x_\star \in \mathcal{X}_t^s$ for $t \in N_T$. Note that we are not interested in expanding the safe set in all possible directions. Instead, what aligns with the objective of minimizing regret, is expanding the safe set in the "correct" direction, that of $x_\star$. Therefore, $\lambda_{\min}(V_t) \geq \mathcal{O}(\log t)$ provides enough expansion of the safe set to bound the Term I in (5.4.1).

The rest of the proof is similar as in [43, Lemma 3]; we include in here for completeness.

For rounds $t \in N_T$, we know that

$$\langle x(\tilde{\theta}_t), \tilde{\theta}_t \rangle \geq \langle x_\star, \tilde{\theta}_t \rangle,$$

since $x(\tilde{\theta}_t) = \arg\max_{x \in \mathcal{X}_t^s} \langle x, \tilde{\theta}_t \rangle$ and we have already shown that $x_\star \in \mathcal{X}_t^s$. Therefore, it suffices to show that

$$\mathbb{P}\left(\langle x_\star, \tilde{\theta}_t \rangle \geq \langle x_\star, \theta_\star \rangle\right) \geq p. \tag{C.3.12}$$

From the definition of $\tilde{\theta}_t$, we can rewrite (C.3.12) as

$$\mathbb{P}\left(\langle x_\star, \hat{\theta}_t \rangle + \beta_t \langle x_\star, V_t^{-1/2} \eta_t \rangle \geq \langle x_\star, \theta_\star \rangle\right) \geq p,$$

or equivalently,

$$\mathbb{P}\left(\beta_t \langle x_\star, V_t^{-1/2} \eta_t \rangle \geq \langle x_\star, \theta_\star - \hat{\theta}_t \rangle\right) \geq p. \tag{C.3.13}$$

Then, we use Cauchy-Schwarz for the LHS of (C.3.13), and given the fact that $\|\theta_\star - \hat{\theta}_t\|_{V_t} \leq \beta_t$, we get

$$\mathbb{P}\left(\langle x_\star, V_t^{-1/2} \eta_t \rangle \geq \|x_\star\|_{V_t^{-1/2}}\right) \geq p,$$

or equivalently,

$$\mathbb{P}\left(\langle u_t, \eta_t \rangle \geq 1\right) \geq p, \tag{C.3.14}$$

where $u_t = \frac{x_\star V_t^{-1/2}}{\|x_\star\|_{V_t^{-1/2}}}$. Therefore, $\|u_t\|_2 = 1$ by construction. At last, we know that (C.3.14) is true thanks to the anti-concentration distributional property of the parameter η_t in Definition C.3.1. ■

As mentioned, after showing that SCLTS for rounds $t \in N_T$ samples from the set of optimistic parameters with a constant probability, the rest of the proof for bounding the

regret of Term I is similar to that of [43]. In particular, we conclude with the following bound

$$\text{Term I} := \sum_{t \in N_T} \left(\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle \right) \\ (\beta_T(\delta') + \gamma_T(\delta')(1 + \frac{4}{p})) \sqrt{2|N_T| d \log\left(1 + \frac{|N_T| L^2}{\lambda}\right)} + \frac{4\gamma_T(\delta')}{p} \sqrt{\frac{8|N_T| L^2}{\lambda} \log \frac{4}{\delta}}, \tag{C.3.15}$$

where $\delta' = \frac{\delta}{6|N_T|}$, and,

$$\gamma_t(\delta) = \beta_t(\delta') \left(1 + \frac{2}{C}\right) \sqrt{cd \log\left(\frac{c'd}{\delta}\right)}, \tag{C.3.16}$$

and since $N_T \leq T$, the proof is completed.

C.4 Proof of Theorem 5.4.3

In this section, we prove an upper bound of order $\mathcal{O}(\log T)$ on the number of times that SCLTS plays the conservative actions.

Let τ be any round that the algorithm plays the conservative action, i.e., at round τ, either $F = 0$ or $\lambda_{\min}(V_\tau) < k_\tau^1 = \left(\frac{2L\beta_\tau}{\kappa_\tau + \alpha r_{b_\tau}}\right)^2$. By definition, if $F = 0$, we have

$$\nexists x \in \mathcal{X} : \langle x, \hat{\theta}_\tau \rangle - \beta_\tau \|x\|_{V_\tau^{-1}} \geq (1 - \alpha) r_{b_\tau}, \tag{C.4.1}$$

and since we know that $x_\star \in \mathcal{X}$, and $\theta_\star \in \mathcal{E}_t$ with high probability, we can write

$$\langle x_\star, \theta_\star \rangle - 2\beta_\tau \|x_\star\|_{V_\tau^{-1}} \leq \langle x_\star, \hat{\theta}_\tau \rangle - \beta_\tau \|x_\star\|_{V_\tau^{-1}} < (1 - \alpha) r_{b_\tau}. \tag{C.4.2}$$

From (C.4.2), we can get

$$\kappa_{b_\tau} + \alpha r_{b_\tau} < 2\beta_\tau \|x_\star\|_{V_\tau^{-1}} \leq \frac{2\beta_\tau L}{\sqrt{\lambda_{\min}(V_\tau)}}, \tag{C.4.3}$$

and hence the following upper bound on minimum eigenvalue of the Gram matrix:

$$\lambda_{\min}(V_\tau) < \left(\frac{2\beta_\tau L}{\kappa_{b_\tau} + \alpha r_{b_\tau}}\right)^2 \leq k_\tau^1. \tag{C.4.4}$$

Therefore, at any round τ that a conservative action is played, whether it is because $F = 0$, or because we have $\{\lambda_{\min}(V_\tau) < k_\tau\}$, we can always conclude that

$$\lambda_{\min}(V_\tau) < k_\tau^1. \tag{C.4.5}$$

The remainder of the proof builds on two auxiliary lemmas. First, in Lemma C.4.1, we show that the minimum eigenvalue of the Gram matrix V_t is lower bounded with the number of times SCLTS plays the conservative actions.

Lemma C.4.1 *Under Assumptions 9, 10, and 11, it holds that*

$$\mathbb{P}(\lambda_{min}(V_t) \leq t) \leq d \exp\left(-\frac{(\rho_1^2 |N_t^c| \sigma_\zeta^2 - t)^2}{8|N_t^c| h_1^2}\right), \tag{C.4.6}$$

where $h_1 = 2\rho_1(1-\rho_1)L + 2\rho_1^2$ *and* $\rho_1 = (\frac{r_l}{S+r_h})\alpha$.

Using (C.4.5) and applying Lemma C.4.1, it can be checked that with probability $1-\delta$,

$$\left(\frac{2L\beta_\tau}{\kappa_l + \alpha r_l}\right)^2 > \rho_1^2 |N_\tau^c| \sigma_\zeta^2 - \sqrt{8|N_\tau^c| h_1^2 \log(\frac{d}{\delta})}. \tag{C.4.7}$$

This gives an explicit inequality that must be satisfied by τ. Solving with respect to

τ leads to the desired. In particular, we apply simple Lemma C.4.2 below.

Lemma C.4.2 *For any $a, b, c > 0$, if $ax - \sqrt{bx} < c$, then the following holds for $x \geq 0$*

$$0 \leq x < \frac{2ac + b + \sqrt{b^2 + 4abc}}{2a^2}. \tag{C.4.8}$$

Using Lemma C.4.2 results in the following upper bound on the $|N_\tau^c|$

$$|N_\tau^c| \leq \left(\frac{2L\beta_\tau}{\rho_1\sigma_\zeta(\kappa_l + \alpha r_l)}\right)^2 + \frac{2h_1^2}{\rho_1^4\sigma_\zeta^4}\log(\frac{d}{\delta}) + \frac{h_1 2L\beta_\tau}{(\kappa_l + \alpha r_l)\rho_1^3\sigma_\zeta^3}\sqrt{8\log(\frac{d}{\delta})}. \tag{C.4.9}$$

Therefore, we can upper bound N_T^c with the following:

$$|N_T^c| \leq \left(\frac{2L\beta_T}{\rho_1\sigma_\zeta(\kappa_l + \alpha r_l)}\right)^2 + \frac{2h_1^2}{\rho_1^4\sigma_\zeta^4}\log(\frac{d}{\delta}) + \frac{2Lh_1\beta_T\sqrt{8\log(\frac{d}{\delta})}}{\rho_1^3\sigma^3(\kappa_l + \alpha r_l)}, \tag{C.4.10}$$

which has order $\mathcal{O}\left(\frac{L^2 d\log(\frac{T}{\delta})}{\alpha^2 r_l^2(\kappa_l+\alpha r_l)^2\sigma_\zeta^2} + \left(\frac{L^2}{\alpha^2 r_l^2\sigma_\zeta^4} + d^2\right)\log(\frac{d}{\delta})\right)$, as promised.

C.4.1 Proof of Lemma C.4.1

Our objective is to establish a lower bound on $\lambda_{\min}(V_t)$ for all t. It holds that

$$\begin{aligned}
V_t &= \lambda I + \sum_{s=1}^{t} x_s x_s^\top \\
&\succeq \sum_{s \in N_t^c} \left((1-\rho_1)x_{b_s} - \rho_1 \zeta_s\right)\left((1-\rho_1)x_{b_s} - \rho_1 \zeta_s\right)^\top \\
&= \sum_{s \in N_t^c} \Big((1-\rho_1)^2 x_{b_s} x_{b_s}^\top - \rho_1(1-\rho_1) x_{b_s}\zeta_s^\top - \rho_1(1-\rho_1)\zeta_s x_{b_s}^\top + \rho_1^2 \zeta_s \zeta_s^\top \Big) \\
&\succeq \sum_{s \in N_t^c} \Big(- \rho_1(1-\rho_1) x_{b_s}\zeta_s^\top - \rho_1(1-\rho_1)\zeta_s x_{b_s}^\top + \rho_1^2 \zeta_s \zeta_s^\top \Big) \\
&= \sum_{s \in N_t^c} \Big(\rho_1^2 \mathbb{E}[\zeta_s \zeta_s^\top] - \rho_1(1-\rho_1) x_{b_s}\zeta_s^\top - \rho_1(1-\rho_1)\zeta_s x_{b_s}^\top + \rho_1^2 \zeta_s \zeta_s^\top - \rho_1^2 \mathbb{E}[\zeta_s \zeta_s^\top] \Big) \\
&\succeq \rho_1^2 \sigma_\zeta^2 |N_t^c| I + \sum_{s \in N_t^c} U_s, \qquad \text{(C.4.11)}
\end{aligned}$$

where U_s is defined as

$$U_s = \Big(- \rho_1(1-\rho_1) x_{b_s}\zeta_s^\top - \rho_1(1-\rho_1)\zeta_s x_{b_s}^\top + \rho_1^2 \zeta_s \zeta_s^\top - \rho_1^2 \mathbb{E}[\zeta_s \zeta_s^\top] \Big). \qquad \text{(C.4.12)}$$

Then, using Weyl's inequality, it follows that

$$\lambda_{\min}(V_t) \geq \rho_1^2 \sigma_\zeta^2 |N_t^c| - \lambda_{\max}\Big(\sum_{s \in N_t^c} U_s\Big).$$

Next, we apply the matrix Azuma inequality (see Theorem C.4.3) to find an upper bound on $\lambda_{\max}(\sum_{s \in N_t^c} U_s)$. For this, we first need to show that the sequence of matrices U_s satisfies the conditions of Theorem C.4.3. By definition of U_s in (C.4.12), it follows that $\mathbb{E}[U_s | \mathcal{F}_{s-1}] = 0$, and $U_s^\top = U_s$. Also, we construct the sequence of deterministic matrices A_s such that $U_s^2 \preceq A_s^2$ as follows. We know that for any matrix B, $B^2 \leq \|B\|_2^2 I$,

where $\|B\|_2$ is the maximum singular value of B, i.e.,

$$\sigma_{\max}(B) = \max_{\|u\|_1=\|v\|_2=1} u^\top B v.$$

Thus, we first show the following bound on the maximum singular value of the matrix U_s defined in (C.4.12):

$$\begin{aligned}\max_{\|u\|_1=\|v\|_2=1} u^\top U_s v &= -\rho_1(1-\rho_1)(u^\top x_{b_s})(v^\top \zeta_s)^\top - \rho_1(1-\rho_1)(u^\top \zeta_s)(v^\top x_{b_s})^\top + \\ &\quad \rho_1^2 (u^\top \zeta_s)(v^\top \zeta_s)^\top - \rho_1^2 \mathbb{E}\left[(u^\top \zeta_s)(v^\top \zeta_s)^\top\right] \\ &\leq \rho_1(1-\rho_1)\|x_{b_s}\|_2\|\zeta_s\|_2 + \rho_1(1-\rho_1)\|\zeta_s\|_2\|x_{b_s}\|_2 + \rho_1^2\|\zeta_s\|_2^2 + \rho_1^2\mathbb{E}\left[\|\zeta_s\|_2^2\right] \\ &\leq 2\rho_1(1-\rho_1)L + 2\rho_1^2, \end{aligned} \tag{C.4.13}$$

where we have used Cauchy-Schwarz inequality and the last inequality comes from the fact that $\|\zeta_s\|_2 = 1$ almost surely, and $\|x_{b_s}\|_2 \leq L$ by Assumption 11. From the derivations above, and choosing $A_s = h_1 I$, with $h_1 = 2\rho_1(1-\rho_1)L + 2\rho_1^2$, it almost surely holds that $U_s^2 \preceq \sigma_{\max}(U_s)^2 I \preceq h_1^2 I = A_s^2$. Moreover, using triangular inequality, it holds that

$$\|\sum_{s\in N_t^c} A_s^2\| \leq \sum_{s\in N_t^c} \|A_s^2\| \leq |N_t^c| h_1^2.$$

Now we apply the the matrix Azuma inequality, to conclude that for any $c \geq 0$,

$$\mathbb{P}\left(\lambda_{\max}(\sum_{s\in N_t^c} U_s) \geq c\right) \leq d \exp\left(-\frac{c^2}{8|N_t^c|h_1^2}\right).$$

Therefore, it holds that with probability $1-\delta$, $\lambda_{\max}(\sum_{s\in N_t^c} U_s) \leq \sqrt{8|N_t^c|h_1^2 \log(\frac{d}{\delta})}$, and

hence with probability $1-\delta$,

$$\lambda_{\min}(V_t) \geq \rho^2 |N_t^c| \sigma_\zeta^2 - \sqrt{8|N_t^c| h_1^2 \log(\frac{d}{\delta})},$$

or equivalently,

$$\mathbb{P}(\lambda_{\min}(V_t) \leq t) \leq d \exp\left(-\frac{(\rho_1^2 |N_t^c| \sigma_\zeta^2 - t)^2}{8|N_t^c| h_1^2}\right),$$

where $h_1 = 2\rho_1(1-\rho_1)L + 2\rho_1^2$ and $\rho_1 = (\frac{r_l}{S+r_h})\alpha$. This completed the proof of lemma.

C.4.2 Matrix Azuma Inequality

Theorem C.4.3 (Matrix Azuma Inequality, [155]) *Consider a sequence $\{Y_k\}$ of independent, random matrices adapted to the filtration $\{\mathcal{F}_k\}$. Each $\{Y_k\}$ is a self-adjoint matrix such that $\mathbb{E}[Y_k \,|\, \mathcal{F}_{k-1}] = 0$. Consider a fixed matrix A_k such that $Y_k^2 \preceq A_k^2$ holds almost surely. Then, for $t \geq 0$, it holds that*

$$\mathbb{P}\left(\lambda_{max}\left(\sum_{k=1}^{s} Y_k\right) \geq t\right) \leq d \exp\left(-\frac{t^2}{8\|\sum_{k=1}^{s} A_k^2\|}\right). \tag{C.4.14}$$

C.5 Upper Bounding the Regret of SCLTS-BF

In this section we provide the variation of our algorithm for the case of constraints with bandit feedback, which we refer to as SCLTS-BF in Algorithm 8. We then provide a regret bound for SCLTS-BF. The summary of SCLTS-BF is presented in Algorithm 8.

In this setting, we assume that at each round t, with playing an action x_t, the learner

Algorithm 8: SCLTS-BF

Input: δ, T, λ, ρ
Set $\delta' = \frac{\delta}{4T}$
for $t = 1, \dots, T$ **do**
  Sample $\eta_t \sim \mathcal{H}^{\mathrm{TS}}$
  Compute RLS-estimate $\hat{\theta}_t$ and V_t according to (5.3.1) and $\hat{\mu}_t$
  Set $\tilde{\theta}_t = \hat{\theta}_t + \beta_t V_t^{-1/2} \eta_t$
  Build the confidence region $\mathcal{E}_t(\delta')$ in (C.5.2) and $\mathcal{C}_t(\delta')$ in (C.5.3)
  Compute the estimated safe set $\mathcal{P}_t^s = \{x \in \mathcal{X} : \langle x, v \rangle \geq (1-\alpha) q_{b_t}, \forall v \in \mathcal{C}_t\}$
  if the following optimization has a feasible solution:
   $x(\tilde{\theta}_t) = \mathrm{argmax}_{x \in \mathcal{P}_t^s} \langle x, \tilde{\theta}_t \rangle$, **then**
  Set $F = 1$, **else** $F = 0$
  if $F = 1$ **and** $\lambda_{\min}(V_t) \geq \left(\frac{2L\beta_t}{\nu_l + \alpha q_l} \right)^2$, **then**
  Play $x_t = x(\tilde{\theta}_t)$
  else
  play $x_t = x_t^{\mathrm{cb}}$ defined in (C.5.6)
  Observe reward r_t
end for

observes the reward $y_t = \langle x_t, \theta_\star \rangle + \xi_t$ and the following bandit feedback:

$$w_t = \langle x_t, \mu_\star \rangle + \chi_t, \tag{C.5.1}$$

where χ_t is assumed to be a zero-mean R-sub-Gaussian noise.

The main difference of SCLTS-BF with SCLTS is in the definition of the estimated safe action set. In particular, at each round t, SCLTS-BF constructs the following confidence regions:

$$\mathcal{E}_t(\delta') = \{\theta \in \mathbb{R} : \left\| \theta - \hat{\theta}_t \right\|_{V_t} \leq \beta_t(\delta')\}, \tag{C.5.2}$$

$$\mathcal{C}_t(\delta') = \{v \in \mathbb{R} : \| v - \hat{\mu}_t \|_{V_t} \leq \beta_t(\delta')\}, \tag{C.5.3}$$

where $\hat{\mu}_t = V_t^{-1} \sum_{s=1}^{t-1} w_s x_s$ is the RLS-estimate of $\mu_\star$. The radius in (C.5.2) and (C.5.3) is

chosen according to Proposition 5.3.1 such that $\theta_\star \in \mathcal{E}_t$ and $\mu_\star \in \mathcal{C}_t$ with high probability. In order to ensure safety at each round t, SCLTS-BF constructs the following estimated safe action set

$$\mathcal{P}_t^s = \{x \in \mathcal{X} : \langle x, v\rangle \geq (1-\alpha)q_{b_t}, \forall v \in \mathcal{C}_t\}. \tag{C.5.4}$$

The challenge with $\mathcal{P}_t^s$ is that it contains all the actions that are safe with respect to all the parameters in $\mathcal{C}_t$. Thus, there may exist some rounds that $\mathcal{P}_t^s$ is empty. To handle this case, SCLTS-BF proceed as follows. At each round t, given the sampled parameter $\tilde{\theta}_t$, if the estimated safe action set $\mathcal{P}_t^s$ defined in (C.5.4) is not empty, SCLTS-BF plays the safe action

$$x(\tilde{\theta}_t) = \arg\max_{x \in \mathcal{P}_t^s} \langle x, \tilde{\theta}_t\rangle \tag{C.5.5}$$

only if $\lambda_{\min}(V_t) \geq k_t^2$, where $k_t^2 = \left(\frac{2L\beta_t}{\nu_l + \alpha q_l}\right)^2$. Otherwise, it plays the following conservative action

$$x_t^{\text{cb}} = (1-\rho_2)x_{b_t} + \rho_2 \zeta_t, \tag{C.5.6}$$

where $\rho_2 = \alpha(\frac{q_l}{S+q_h})$ in order to ensure that the conservative actions are safe.

Next, we provide a regret guarantee for SCLTS-BF. First, we use the following decomposition of regret:

$$\begin{aligned} R(T) &= \sum_{t=1}^{T} \langle x_\star, \theta_\star\rangle - \langle x_t, \theta_\star\rangle \\ &= \underbrace{\sum_{t \in N_T} \Big(\langle x_\star, \theta_\star\rangle - \langle x_t, \theta_\star\rangle\Big)}_{\text{Term I}} + \underbrace{\sum_{t \in N_T^c} \Big(\langle x_\star, \theta_\star\rangle - \langle (1-\rho)x_{b_t} - \rho\zeta_t, \theta_\star\rangle\Big)}_{\text{Term II}}, \end{aligned} \tag{C.5.7}$$

where N_t^c is the set of rounds $i < t$ that SCLTS-BF plays the conservative actions, and $N_t = \{1, \dots, t\} - N_t^c$. In the following, we upper bound both Term I and Term II, separately.

Bounding Term I. Bounding Term I follows the same steps as that of Theorem 5.4.2. Here, we show that for SCLTS-BF, at rounds $t \in N_T$, the optimal action $x_\star$ belongs to the estimated safe safe, i.e., $x_\star \in \mathcal{P}_t^s$. Then, we conclude that regret of Term I similar to Theorem 5.4.2 has the order of $\mathcal{O}(d^{3/2} \log^{1/2} d \; T^{1/2} \log^{3/2} T)$.

At rounds $t \in N_T$, we know

$$\lambda_{\min}(V_t) \geq k_t^2 \geq \left(\frac{2L\beta_t}{\nu_{b_t} + \alpha q_{b_t}} \right)^2. \tag{C.5.8}$$

Then, in order to show that $x_\star \in \mathcal{X}_t^s$, we need to show

$$\langle x_\star, \hat{\mu}_t \rangle - \beta_t \|x_\star\|_{V_t^{-1}} \geq \langle x_\star, \mu_\star \rangle - 2\beta_t \|x_\star\|_{V_t^{-1}} \geq (1-\alpha) q_{b_t}. \tag{C.5.9}$$

First inequality comes from the fact that $\|\mu_\star - \hat{\mu}_t\|_{V_t} \leq \beta_t$. Therefore, it suffices to show the second inequality holds. We use the fact that $\|x_\star\|_{V_t^{-1}} \leq \frac{\|x_\star\|_2}{\sqrt{\lambda_{\min}(V_t)}} \leq \frac{L}{\sqrt{\lambda_{\min}(V_t)}}$, where we use Assumption 11 to bound $\|x_\star\|_2$. Hence, we have

$$\langle x_\star, \mu_\star \rangle - 2\beta_t \|x_\star\|_{V_t^{-1}} \geq \langle x_\star, \mu_\star \rangle - \frac{2\beta_t L}{\sqrt{\lambda_{\min}(V_t)}}. \tag{C.5.10}$$

Then, it suffices to show that

$$\nu_{b_t} + \alpha q_{b_t} \geq \frac{2\beta_t L}{\sqrt{\lambda_{\min}(V_t)}}, \tag{C.5.11}$$

From (C.5.8), we know that (C.5.11) holds, and hence, $x_\star \in \mathcal{P}_t^s$. Therefore, we can use the result of Theorem 5.4.2, and obtain the desired regret bound.

Bounding Term II. First, we provide the formal statement of the theorem.

Theorem C.5.1 *Let $\lambda, L \geq 1$. On event $\left\{\{\theta_\star \in \mathcal{E}_t, \forall t \in [T]\} \cap \{\mu_\star \in \mathcal{C}_t, \forall t \in [T]\}\right\}$, and Assumptions 12, we can upper bound the number of times SCLTS-BF plays the conservative actions, i.e., $|N_T^c|$ as:*

$$|N_T^c| \leq \left(\frac{2L\beta_T}{\rho_2\sigma_\zeta(\alpha q_l + \nu_l)}\right)^2 + \frac{2h_2^2}{\rho_2^4\sigma_\zeta^4}\log(\frac{d}{\delta}) + \frac{2Lh_2\beta_T\sqrt{8\log(\frac{d}{\delta})}}{\rho_2^3\sigma_\zeta^3(\alpha q_l + \nu_l)} \tag{C.5.12}$$

where $h_2 = 2\rho_2(1-\rho_2)L + 2\rho_2^2$ and $\rho_2 = (\frac{q_l}{S+q_h})\alpha$.

In order to prove Theorem C.5.1, we proceed as follows:

$$\begin{aligned}\sum_{t\in N_T^c}\Big(\langle x_\star, \theta_\star\rangle - \langle (1-\rho_2)x_{b_t} - \rho_2\zeta_t, \theta_\star\rangle\Big) &= \sum_{t\in N_T^c}\langle x_\star, \theta_\star\rangle - \langle x_{b_t}, \theta_\star\rangle + \rho_2(\langle x_{b_t} + \zeta_t, \theta_\star\rangle) \\ &\leq \sum_{t\in N_T^c}\nu_h + \rho_2(q_{b_t} + S) \leq |N_T^c|(\nu_h + \alpha q_l),\end{aligned} \tag{C.5.13}$$

where $q_h \geq q_{b_t} \geq q_l > 0$ and $\nu_h \geq \nu_{b_t} \geq \nu_l$ for all t. Therefore, in order to bound Term II, it suffices to upper bound $|N_T^c|$ which is the number of rounds that SCLTS-BF plays the conservative actions up to round T. In order to do so, we proceed as follows:

Let τ be any round that the algorithm plays the conservative action.

If $F = 0$, i.e.,

$$\nexists x \in \mathcal{X} : \langle x, \hat{\mu}_\tau\rangle - \beta_\tau\|x\|_{V_\tau^{-1}} \geq (1-\alpha)q_{b_\tau}, \tag{C.5.14}$$

and since we know that $x_\star \in \mathcal{X}$, and $\mu_\star \in \mathcal{C}_t$ with high probability, we can write

$$\langle x_\star, \mu_\star\rangle - 2\beta_\tau\|x_\star\|_{V_\tau^{-1}} \leq \langle x_\star, \hat{\mu}_\tau\rangle - \beta_\tau\|x_\star\|_{V_\tau^{-1}} < (1-\alpha)q_{b_\tau}. \tag{C.5.15}$$

Using (C.5.15), we can get

$$\nu_{b_\tau} + \alpha q_{b_\tau} < 2\beta_\tau \|x_\star\|_{V_\tau^{-1}} \leq \frac{2\beta_\tau L}{\sqrt{\lambda_{\min}(V_\tau)}}, \tag{C.5.16}$$

and hence the following upper bound on minimum eigenvalue of the Gram matrix:

$$\lambda_{\min}(V_\tau) < \left(\frac{2\beta_\tau L}{\nu_{b_\tau} + \alpha q_{b_\tau}}\right)^2 \leq \left(\frac{2\beta_\tau L}{\nu_l + \alpha q_l}\right)^2 = k_\tau \tag{C.5.17}$$

Therefore, we show that in the cases where either the event $\{\nexists x \in \mathcal{X} : \langle x, \hat{\mu}_\tau \rangle - \beta_\tau \|x\|_{V_\tau^{-1}} \geq (1-\alpha) q_{b_\tau}\}$ or the event $\{\lambda_{\min}(V_\tau) < k_\tau^2\}$ happen, we can conclude that at round τ

$$\lambda_{\min}(V_\tau) < k_\tau^2. \tag{C.5.18}$$

From Lemma C.4.1, we know that the minimum eigenvalue of the Gram matrix, i.e., $\lambda_{\min}(V_t)$ is lower bounded with the number of times that SCLTS-BF plays the conservative actions, i.e., $|N_T^c|$. Therefore, using (C.5.18), we can get

$$|N_T^c| \leq \left(\frac{2L\beta_T}{\rho_2 \sigma_\zeta (\alpha q_l + \nu_l)}\right)^2 + \frac{2h_2^2}{\rho_2^4 \sigma_\zeta^4} \log(\frac{d}{\delta}) + \frac{2Lh_2\beta_T \sqrt{2\log(\frac{d}{\delta})}}{\rho_2^3 \sigma_\zeta^3 (\alpha q_l + \nu_l)} \tag{C.5.19}$$

where $h_2 = 2\rho_2(1-\rho_2)L + 2\rho_2^2$ and $\rho_2 = \alpha(\frac{q_l}{S+q_h})$.

C.6 Proof of Theorem 5.5.1

In this section, we first present the SCLTS2 algorithm, for the case where the learner does not know the reward of the actions suggested by baseline policy in advance, i.e., r_{b_t}. The summary of SCLTS2 is presented in Algorithm 9.

The algorithm relies on the fact that we can find an upper bound over the value of r_{b_t}, using the fact that $\theta_\star \in \mathcal{E}_t$, i.e.,:

$$\max_{v \in \mathcal{E}_t} \langle x_{b_t}, v \rangle \geq \langle x_{b_t}, \theta_\star \rangle = r_{b_t}. \tag{C.6.1}$$

Then, we can write the safety constraint as follows:

$$\min_{v \in \mathcal{E}_t} \langle x(\tilde{\theta}_t), v \rangle \geq (1-\alpha) \max_{v \in \mathcal{E}_t} \langle x_{b_t}, v \rangle. \tag{C.6.2}$$

It is easy to show that safety constraint (5.2.2) holds when (C.6.2) is true. Therefore, if we choose actions that satisfy (C.6.2), we can ensure that they are safe with respect to the safety constrain in (5.2.2).

Then we propose the estimated safe action set $\mathcal{Z}_t^s$ as:

$$\mathcal{Z}_t^s = \{x \in \mathcal{X} : \min_{v \in \mathcal{E}_t} \langle x, v \rangle \geq (1-\alpha) \max_{v \in \mathcal{E}_t} \langle x_{b_t}, v \rangle\}, \tag{C.6.3}$$

which contains actions that are safe with respect to all the parameter in $\mathcal{E}_t$. At each round t, SCLTS2 plays the safe action $x(\tilde{\theta}_t)$ from $\mathcal{Z}_t^s$ that maximizes the expected reward given the sampled parameter $\tilde{\theta}_t$, i.e.,

$$x(\tilde{\theta}_t) = \arg\max_{x \in \mathcal{Z}_t^s} \langle x, \tilde{\theta}_t \rangle \tag{C.6.4}$$

only if $\lambda_{\min}(V_t) \geq k_t^3$, where $k_t^3 = \left(\frac{2L\beta_t(2-\alpha)}{\kappa_l + \alpha r_l}\right)^2$. Otherwise it plays the conservative action $x_{b_t}^{\text{cb}}$ as:

$$x_t^{\text{cb}} = (1-\rho_3) x_{b_t} + \rho_3 \zeta_t, \tag{C.6.5}$$

where $\rho_3 = \alpha(\frac{r_l}{S+1})$ such that the conservative action x_t^{cb} is safe, where we use Assumption 11 for upper bounding the reward, i.e., $r_{b_t} \leq 1$.

Algorithm 9: SCLTS2

Input: δ, T, λ, ρ
Set $\delta' = \frac{\delta}{4T}$
for $t = 1, \ldots, T$ **do**
  Sample $\eta_t \sim \mathcal{H}^{\text{TS}}$
  Compute RLS-estimate $\hat{\theta}_t$ and V_t according to (5.3.1)
  Set $\tilde{\theta}_t = \hat{\theta}_t + \beta_t V_t^{-1/2} \eta_t$
  Build the confidence region $\mathcal{E}_t(\delta')$ in (5.3.2)
  Compute the estimated safe set
   $\mathcal{Z}_t^s = \{x \in \mathcal{X} : \min_{v \in \mathcal{E}_t} \langle x, v \rangle \geq (1-\alpha) \max_{v \in \mathcal{E}_t} \langle x_{b_t}, v \rangle\}$
  if the following optimization is feasible: $x(\tilde{\theta}_t) = \arg\max_{x \in \mathcal{Z}_t^s} \langle x, \tilde{\theta}_t \rangle$, **then**
  Set $F = 1$, **else** $F = 0$
  if $F = 1$ **and** $\lambda_{\min}(V_t) \geq \left(\frac{2L\beta_t(2-\alpha)}{\kappa_l + \alpha r_l}\right)^2$, **then**
  Play $x_t = x(\tilde{\theta}_t)$
  else
  play $x_t = x_t^{\text{cb}}$ defined in (C.6.5)
  Observe reward y_t
end for

In order to bound the regret of SCLTS2, we first use the decomposition defined in Proposition 5.4.1. The regret of Term I is similar to that of SCLTS (i.e., Theorem 5.4.2). Hence, it suffices to upper bound the number of time SCLTS2 plays the conservative actions, i.e., $|N_T^c|$.

In order to bound $|N_T^c|$, we proceed as follows. Let τ be the round that SCLTS2 plays a conservative action. If $F = 0$, i.e.,

$$\nexists x \in \mathcal{X} : \min_{v \in \mathcal{C}_\tau} \langle x, v \rangle \geq (1-\alpha) \max_{v \in \mathcal{C}_\tau} \langle x_{b_\tau}, v \rangle. \tag{C.6.6}$$

Using the fact that $x_\star \in \mathcal{X}$, we can write

$$\langle x_\star, \hat{\theta}_\tau \rangle - \beta_\tau \|x_\star\|_{V_\tau^{-1}} < (1-\alpha)\left(\langle x_{b_\tau}, \hat{\theta}_\tau \rangle + \beta_\tau \|x_{b_\tau}\|_{V_\tau^{-1}}\right). \tag{C.6.7}$$

Then, since $\|\theta_\star - \hat{\theta}_t\|_{V_t} \leq \beta_t$, we can upper bound the RHS and lower bound the LHS of (C.6.7), and get

$$\langle x_\star, \theta_\star \rangle - 2\beta_\tau \|x_\star\|_{V_\tau^{-1}} < (1-\alpha)\left(\langle x_{b_\tau}, \theta_\star \rangle + 2\beta_\tau \|x_{b_\tau}\|_{V_\tau^{-1}}\right), \tag{C.6.8}$$

or equivalently,

$$\kappa_{b_\tau} + \alpha r_{b_\tau} < 2\beta_\tau \|x_\star\|_{V_\tau^{-1}} + 2(1-\alpha)\beta_\tau \|x_{b_\tau}\|_{V_\tau^{-1}}. \tag{C.6.9}$$

Then we can use the fact that $\|x_\star\|_{V_\tau^{-1}} \leq \frac{L}{\sqrt{\lambda_{\min}(V_\tau)}}$ and $\|x_{b_\tau}\|_{V_\tau^{-1}} \leq \frac{L}{\sqrt{\lambda_{\min}(V_\tau)}}$, where we use Assumption 11 for upper bounding $\|x_\star\|_2$. Thus, we upper bound the RHS of (C.6.9) as follows:

$$\kappa_{b_\tau} + \alpha r_{b_\tau} < 2\beta_\tau \frac{L}{\sqrt{\lambda_{\min}(V_\tau)}} + 2(1-\alpha)\beta_\tau \frac{L}{\sqrt{\lambda_{\min}(V_\tau)}}, \tag{C.6.10}$$

and hence, we can get the following upper bound $\lambda_{\min}(V_\tau)$ as follows:

$$\lambda_{\min}(V_\tau) < \left(\frac{2L\beta_T(2-\alpha)}{\kappa_{b_\tau} + \alpha r_{b_\tau}}\right)^2 \leq \left(\frac{2L\beta_T(2-\alpha)}{\kappa_l + \alpha r_l}\right)^2 = k_\tau^3. \tag{C.6.11}$$

Therefore, we show that whether the event $F = 0$ happens or $\lambda_{\min}(V_t) < k_t^3$, we can achieve the upper bound provided in (C.6.11). Then, using the result of Lemma C.4.1, where we show that $\lambda_{\min}(V_t)$ is lower bounded with the number of times the algorithm

plays the conservative actions, we obtain the following upper bound on the $|N_\tau^c|$

$$|N_\tau^c| \leq \left(\frac{2L\beta_\tau(2-\alpha)}{\rho_3\sigma_\zeta(\kappa_l+\alpha r_l)}\right)^2 + \frac{2h_3^2}{\rho_3^4\sigma_\zeta^4}\log(\frac{d}{\delta}) + \frac{2Lh_3\beta_\tau(2-\alpha)}{\rho_3^3\sigma_\zeta^3(\kappa_l+\alpha r_l)}\sqrt{2\log(\frac{d}{\delta})}, \tag{C.6.12}$$

where $h_3 = 2\rho_3(1-\rho_3)L + 2\rho_3^2$ and $\rho_3 = \alpha(\frac{r_l}{S+1})$.

C.7 Stage-wise Conservative Linear UCB (SCLUCB) Algorithm

In this section we propose a UCB-based safe stochastic linear bandit algorithm called Stage-wise Conservative Linear-UCB (SCLUCB), which is a safe counterpart of LUCB for the stage-wise conservative bandit setting. In particular, at each round t, given the RLS-estimate $\hat{\theta}_t$ of $\theta_\star$, SCLUCB constructs the confidence region $\mathcal{E}_t$ as follows:

$$\mathcal{E}_t(\delta) = \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|_{V_t} \leq \beta_t(\delta)\}. \tag{C.7.1}$$

The radius $\beta_t(\delta)$ is chosen as in Proposition 5.3.1 such that $\theta_\star \in \mathcal{E}_t(\delta)$ with probability $1-\delta$. Then, similar to SCLTS, it builds the estimated safe set $\mathcal{X}_t^s$ such that it includes actions that are safe with respect to all the parameter in $\mathcal{E}_t$, i.e.,

$$\mathcal{X}_t^s = \{x \in \mathcal{X} : \langle x, v\rangle \geq (1-\alpha)r_{b_t}, \forall v \in \mathcal{E}_t\}. \tag{C.7.2}$$

Similar to SCLTS, the challenge with $\mathcal{X}_t^s$ is that there may exist some rounds that $\mathcal{X}_t^s$ is empty. In order to face this problem, SCLUCB proceed as follows. In order to guarantee

safety, at each round t, if $\mathcal{X}_t^s$ is not empty, SCLUCB plays the action $\bar{x}_t$ as

$$(\bar{x}_t, \bar{\theta}_t) = \max_{x \in \mathcal{X}_t^s} \max_{v \in \mathcal{E}_t} \langle x, v \rangle \tag{C.7.3}$$

only if $\lambda_{\min}(V_t) \geq \left(\frac{2L\beta_t}{\kappa_l + \alpha r_{b_l}}\right)^2$, otherwise it plays the conservative action x_t^{cb} defined in (5.3.6). The summary of SCLUCB is presented in Algorithm (10).

Algorithm 10: Stage-wise Conservative Linear UCB (SCLUCB)

Input: δ, T, λ, ρ
for $t = 1, \ldots, T$ **do**
  Compute RLS-estimate $\hat{\theta}_t$ and V_t according to (5.3.1)
  Build the confidence region $\mathcal{E}_t(\delta)$ in (C.7.1)
  Compute the estimated safe set $\mathcal{X}_t^s$ in (C.7.2)
  if the following optimization is feasible: $\bar{x}_t = \arg\max_{x \in \mathcal{X}_t^s} \max_{v \in \mathcal{E}_t} \langle x, v \rangle$, **then**
    Set $F = 1$, **else** $F = 0$
  if $F = 1$ **and** $\lambda_{\min}(V_t) \geq \left(\frac{2L\beta_t}{\kappa_l + \alpha r_{b_l}}\right)^2$, **then**
    Play $x_t = \bar{x}_t$
  else
    play $x_t = x_t^{\text{cb}}$ defined in (5.3.6)
  Observe reward y_t
end for

Next, we provide the regret guarantee for SCLUCB. Recall, N_{t-1} be the set of rounds $i < t$ at which SCLUCB plays the action in (5.3.5). Similarly, $N_{t-1}^c = \{1, \ldots, t-1\} - N_{t-1}$ is the set of rounds $j < t$ at which SCLUCB plays the conservative actions.

Proposition C.7.1 *The regret of SCLUCB can be decomposed into two terms as follows:*

$$R(T) \leq \underbrace{\sum_{t \in N_T} \left(\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle\right)}_{\textit{Term I}} + \underbrace{|N_T^c| \left(\kappa_h + \rho_1(r_h + S)\right)}_{\textit{Term II}} \tag{C.7.4}$$

In the following, we bound both terms, separately.

Bounding Term I. The first Term in (C.7.4) is the regret caused by playing the safe actions that maximize the reward given the true parameter is $\bar{\theta}_t$. The idea of bounding Term I is similar to [42]. We use the fact that for $t \in N_T$, $x_t = \bar{x}_t$, and start with the following decomposition of the instantaneous regret for $t \in N_T$:

$$\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle = \underbrace{\langle x_\star, \theta_\star \rangle - \langle \bar{x}_t, \bar{\theta}_t \rangle}_{\text{Term A}} + \underbrace{\langle \bar{x}_t, \bar{\theta}_t \rangle - \langle \bar{x}_t, \theta_\star \rangle}_{\text{Term B}} \tag{C.7.5}$$

Bounding Term A. Since for round $t \in N_t$, we require that $\lambda_{\min}(V_t) \geq k_t^1$, where $k_t^1 = \left(\frac{2L\beta_t}{\kappa_l + \alpha r_{b_l}}\right)^2$, we can conclude that $x_\star \in \mathcal{X}_t^s$. Therefore, due to (C.7.3), we have $\langle \bar{x}_t, \bar{\theta}_t \rangle \geq \langle x_\star, \theta_\star \rangle$, and hence Term A is not positive.

Bounding Term B. In order to bound Term B, we use the following chain of inequalities:

$$\begin{aligned}
\text{Term B} := \langle \bar{x}_t, \bar{\theta}_t \rangle - \langle \bar{x}_t, \theta_\star \rangle &= \langle \bar{x}_t, \bar{\theta}_t \rangle - \langle \bar{x}_t, \hat{\theta}_t \rangle + \langle \bar{x}_t, \hat{\theta}_t \rangle - \langle \bar{x}_t, \theta_\star \rangle \\
&\leq \|\bar{x}_t\|_{V_t^{-1}} \|\bar{\theta}_t - \hat{\theta}_t\|_{V_t} + \|\bar{x}_t\|_{V_t^{-1}} \|\hat{\theta}_t - \theta_\star\|_{V_t} \\
&\leq 2\beta_t \|\bar{x}_t\|_{V_t^{-1}},
\end{aligned} \tag{C.7.6}$$

where the last inequality follows from Proposition 5.3.1. Recall, from Assumption 11, we have the following trivial bound:

$$\langle x_\star, \theta_\star \rangle - \langle \bar{x}_t, \theta_\star \rangle \leq 2. \tag{C.7.7}$$

Thus, we conclude the following

$$\text{Term B} \leq 2 \min(\beta_t \|\bar{x}_t\|_{V_t^{-1}}, 1). \tag{C.7.8}$$

Next, we state a direct application of Lemma 11 in [42].

Lemma C.7.2 *For $\lambda > 0$, and under Assumptions 9, 10, and 11, we have*

$$\sum_{t=1}^{T} \min(\|\bar{x}_t\|^2_{V_t^{-1}}, 1) \leq 2d \log\left(1 + \frac{TL^2}{\lambda d}\right) \tag{C.7.9}$$

Therefore, from Lemma C.7.2, we can conclude the following bound on regret of Term B:

$$\sum_{t \in N_T} 2\min(\beta_t \|\bar{x}_t\|_{V_t^{-1}}, 1) \leq 2\beta_T \sqrt{2d|N_T| \log(1 + \frac{|N_T|L^2}{\lambda d})}. \tag{C.7.10}$$

Next, in Theorem C.7.3, we provide an upper bound on the regret of Term I which is of order $\mathcal{O}\left(d\sqrt{T}\log(\frac{TL^2}{\lambda\delta})\right)$.

Theorem C.7.3 *On event $\{\theta_\star \in \mathcal{E}_t\}$ for a fixed $\delta \in (0,1)$, with probability $1-\delta$, it holds that:*

$$\sum_{t \in N_T} (\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle) \leq 2\beta_T \sqrt{2dT \log(1 + \frac{TL^2}{\lambda d})} \tag{C.7.11}$$

Bounding Term II. In order to bound Term II in (C.7.4), we need to find an upper bound on the number of times that SCLUCB plays the conservative actions up to time T, i.e., $|N_T^c|$. We prove an upper bound on $|N_T^c|$ in Theorem C.7.4 which has the order of $\mathcal{O}\left(\frac{L^2 d \log(\frac{T}{\delta}) \log(\frac{d}{\delta})}{\alpha^4 (r_l^2 \wedge r_l^4) \kappa_l (\sigma_\zeta^2 \wedge \sigma_\zeta^4)}\right)$.

Theorem C.7.4 *Let $\lambda, L \geq 1$. On event $\{\theta_\star \in \mathcal{E}_t, \forall t \in [T]\}$, and under Assumption 12, we can upper bound the number of times SCLUCB plays the conservative actions, i.e.,*

$|N_T^c|$ *as:*

$$|N_T^c| \leq \left(\frac{2L\beta_T}{\rho_1\sigma_\zeta(\kappa_l+\alpha r_l)}\right)^2 + \frac{2h_1^2}{\rho_1^4\sigma_\zeta^4}\log(\frac{d}{\delta}) + \frac{2Lh_1\beta_T\sqrt{8\log(\frac{d}{\delta})}}{\rho_1^3\sigma^3(\kappa_l+\alpha r_l)}, \tag{C.7.12}$$

where $h_1 = 2\rho_1(1-\rho_1)L + 2\rho_1^2$ *and* $\rho_1 = (\frac{r_l}{S+r_h})\alpha$.

The proof is similar to that of Theorem 5.4.3, and we omit its proof here.

C.8 Comparison with Safe-LUCB

In this section, we extend our results to an alternative safe bandit formulation proposed in [70], where the algorithm Safe-LUCB was proposed. In order to do so, we first present the safety constraint in [70], and then we show the required modification of SCLUCB to handle this case, which we refer to as SCLUCB2. Then, we provide a problem-dependent regret bound for SCLUSB2, and we show that it matches the problem dependent regret bound of Safe-LUCB in [70]. We need to note that in [70], they also provide a general regret bound of order $\tilde{\mathcal{O}}(T^{2/3})$ for Safe-LUCB which we do not discuss in this paper.

In [70], it is assumed that the learner is given a convex and compact decision set $\mathcal{D}_0$ which contains the origin, and with playing the action x_t, she observes the reward of $y_t = x_t^\top\theta_\star + \eta_t$, where $\theta_\star$ is the fixed unknown parameter, and η_t is R-sub-Gaussian noise. Moreover, The learning environment is subject to the linear safety constraint

$$x^\top B\theta_\star \leq C, \tag{C.8.1}$$

which needs to be satisfied at all rounds t with high probability, and an action x_t is called safe, if it satisfies (C.8.1). In (C.8.1), the matrix $B \in \mathbb{R}^{d\times d}$ and the positive constant C

are known to the learner. However, the learner does not receive any bandit feedback on the value $x^\top B\theta_\star$ and her information is restricted to those she receives from the reward.

Given the above constraint, the learner is restricted to choose actions from the safe set $\mathcal{D}_0^s$ as:

$$\mathcal{D}_0^s(\theta_\star) = \{x \in \mathcal{D}_0 : x^\top B\theta_\star \leq C\}. \tag{C.8.2}$$

Since $\theta_\star$ in unknown, the safe set $\mathcal{D}_0^s$ is unknown to the learner. Then, in [70], they provide the problem-dependent regret bound (for the case where $\Delta := C - x^\top B\theta_\star > 0$) of order $\mathcal{O}(\sqrt{T}\log T)$. In the following, we present the required modification of SCLUSB to handle this safe bandit formulation, and propose the new algorithm called SCLUCB2 that we prove a problem dependnt regret bound of order $\mathbb{O}(\sqrt{T}\log T)$. We need to note that [70] also provide a general regret bound of order $\tilde{\mathcal{O}}(T^{2/3})$ for the case where $\Delta = 0$; however, we do not discuss this case in this paper.

At each round t, given the RLS-estimate $\hat{\theta}_t$ of $\theta_\star$, SLUCB2 builds the confidence region $\mathcal{E}_t$ as:

$$\mathcal{E}_t = \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|_{V_t} \leq \beta_t\}, \tag{C.8.3}$$

and the radius β_t is chosen according to Proposition 5.3.1 such that $\theta_\star \in \mathcal{E}_t$ with high probability. The learner does not know the safe set $\mathcal{D}_0^s$; however, she knows that $\theta_\star \in \mathcal{E}_t$ with high probability. Hence, SLUCB2 constructs the estimated safe set $\mathcal{D}_t^s$ such that it

contains actions that are safe with respect to all the parameter in $\mathcal{E}_t$, i.e.,

$$\begin{aligned}\mathcal{D}_t^s &= \{x \in \mathcal{D}_0 : x^\top Bv \leq C, \forall v \in \mathcal{E}_t\} \\ &= \{x \in \mathcal{D}_0 : \max_{v \in \mathcal{E}_t} x^\top Bv \leq C\} \\ &= \{x \in \mathcal{D}_0 : x^\top B\hat{\theta}_t + \beta_t \|Bx\|_{V_t^{-1}} \leq C\} \end{aligned} \tag{C.8.4}$$

Clearly, action $x = [0]^d$ (origin) is a safe action since $C > 0$, and also $[0]^d \in \mathcal{D}_0$. Thus, $[0]^d \in \mathcal{D}_t^s$. Since $x = [0]^d$ is a known safe action, we define the conservative action x_0^{c} as:

$$x_0^{\mathrm{c}} = (1-\rho)[0]^d + \rho\zeta_t = \rho\zeta_t, \tag{C.8.5}$$

where ζ_t is a sequence of IID random vectors such that $\|\zeta_t\|_2 = 1$ almost surly, and $\sigma_\zeta = \lambda_{\min}(\mathrm{Cov}(\zeta_t)) > 0$. We choose the constant ρ according to the Lemma C.8.1 in order to ensure that the conservative action x_0^{c} is safe.

Lemma C.8.1 *At each round t, for any $\rho \in (0, \bar{\rho})$, where*

$$\bar{\rho} = \frac{C}{\|B\|S}, \tag{C.8.6}$$

the conservative action $x_0^c = \rho\zeta_t$ is guaranteed to be safe almost surly.

We choose $\rho = \frac{C}{\|B\|S}$ for the rest of this section, and hence the conservative action is

$$x_0^{\mathrm{c}} = \frac{C}{\|B\|S}\zeta_t. \tag{C.8.7}$$

Let $\Delta = C - x_\star^\top B\theta_\star$. We consider the case where $\Delta > 0$. At each t, in order to guarantee safety, SCLUCB2 only chooses its action from the estimated safe set $\mathcal{D}_t^s$. The challenge with $\mathcal{D}_t^s$ is that it includes actions that are safe with respect to all parameter

in $\mathcal{E}_t$, and not only $\theta_\star$. Thus, there may exist some rounds that $\mathcal{D}_t^s$ is empty. At round t, if $\mathcal{D}_t^s$ is not empty, SCLUCB2 plays the safe action

$$\bar{x}_t = \arg\max_{x \in \mathcal{D}_t^s} \max_{v \in \mathcal{E}_t} \langle x, v \rangle \tag{C.8.8}$$

only if $\lambda_{\min}(V_t) \geq \left(\frac{2L\beta_t \|B\|}{\Delta}\right)^2$, otherwise it plays the conservative action x_0^c in (C.8.7). The summary of SCLUCB2 is presented in Algorithm 11.

Algorithm 11: SCLUCB2

Input: δ, T, λ, ρ
for $t = 1, \ldots, T$ **do**
  Compute RLS-estimate $\hat{\theta}_t$ and V_t according to (5.3.1)
  Build the confidence region $\mathcal{E}_t(\delta)$ in (C.8.3)
  Compute the estimated safe set $\mathcal{D}_t^s$ in (C.8.4)
  if the following optimization is feasible: $\bar{x}_t = \arg\max_{x \in \mathcal{D}_t^s} \max_{v \in \mathcal{E}_t} \langle x, v \rangle$, **then**
  Set $F = 1$, **else** $F = 0$
  if $F = 1$ **and** $\lambda_{\min}(V_t) \geq \left(\frac{2L\beta_t \|B\|}{\Delta}\right)^2$, **then**
  Play $x_t = \bar{x}_t$
  else
  play $x_t = x_0^c$ defined in (C.8.7)
  Observe reward y_t
end for

In the following we provide the regret guarantee for SCLUCB2. Let N_{t-1} be the set of rounds $i < t$ at which SCLUCB2 plays the action in (C.8.8). Similarly, $N_{t-1}^c = \{1, \ldots, t-1\} - N_{t-1}$ is the set of rounds $j < t$ at which SCLUCB2 plays the conservative action in (C.8.7).

First, we use the following decomposition of the regret, then we bound each term separately.

Proposition C.8.2 *The regret of SCLUCB2 can be decomposed to the following two*

terms:

$$
\begin{aligned}
R(T) &= \sum_{t=1}^{T} \langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle \\
&= \sum_{t \in N_T} \Big(\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle \Big) + \sum_{t \in N_T^c} \Big(\langle x_\star, \theta_\star \rangle - \langle x_0^c, \theta_\star \rangle \Big), \\
&\leq \underbrace{\sum_{t \in N_T} \Big(\langle x_\star, \theta_\star \rangle - \langle x_t, \theta_\star \rangle \Big)}_{\textit{Term I}} + \underbrace{2|N_t^c|}_{\textit{Term II}}.
\end{aligned} \tag{C.8.9}
$$

Bounding Term I. In order to bound Term I, we proceed as follows. First, we show that at rounds $t \in N_T$, the optimal action $x_\star$ belongs to the estimated safe set $\mathcal{D}_t^s$, i.e., $x_\star \in \mathcal{D}_t^s$. To do so, we need to show that

$$
x_\star^\top B \hat{\theta}_t + \beta_t \|Bx_\star\|_{V_t^{-1}} \leq C. \tag{C.8.10}
$$

Since $\|\theta_\star - \hat{\theta}_t\|_{V_t} \leq \beta_t$, it suffices to show that:

$$
x_\star^\top B \theta_\star + 2\beta_t \|Bx_\star\|_{V_t^{-1}} \leq C, \tag{C.8.11}
$$

or equivalently

$$
2\beta_t \|Bx_\star\|_{V_t^{-1}} \leq \Delta, \tag{C.8.12}
$$

where $\Delta = C - x_\star^\top B \theta_\star$. It is easy to see (C.8.10) is true whenever (C.8.11) holds. Using Assumption 11, we can get $\|Bx_\star\|_{V_t^{-1}} \leq \frac{\|B\| \|x_\star\|_2}{\sqrt{\lambda_{\min}(V_t)}} \leq \frac{\|B\| L}{\sqrt{\lambda_{\min}(V_t)}}$. Hence, from (C.8.12), it suffices to show that

$$
\frac{2\beta_t \|B\| L}{\sqrt{\lambda_{\min}(V_t)}} \leq \Delta, \tag{C.8.13}
$$

or equivalently

$$\lambda_{\min}(V_t) \geq \left(\frac{2\beta_t \|B\| L}{\Delta}\right)^2 \tag{C.8.14}$$

that we know it is true for $t \in N_T$. Therefore, on event $\{\theta_\star \in \mathcal{E}_t\}$, $x_\star \in \mathcal{D}_t^s$. We can bound the regret of Term I in (C.8.9) similar to Theorem C.7.3, and get the regret of order $\mathcal{O}\left(d\sqrt{T}\log(\frac{TL^2}{\lambda\delta})\right)$.

Bounding Term II. We need to upper bound the number of times that SCLUCB2 plays the conservative action x_0^c, i.e., $|N_T^c|$. We prove an upper bound on $|N_T^c|$ in Theorem C.8.3 which has the order of $\mathcal{O}\left(\frac{L^2S^2\|B\|^2 d\log(\frac{T}{\delta})\log(\frac{d}{\delta})}{\Delta^2(C\wedge C^2)(\sigma_\zeta^2\wedge\sigma_\zeta^4)}\right)$.

Theorem C.8.3 *Let $\lambda, L \geq 1$. On event $\{\theta_\star \in \mathcal{E}_t, \forall t \in [T]\}$, we can upper bound the number of times SCLUCB2 plays the conservative actions, i.e., $|N_T^c|$ as:*

$$|N_T^c| \leq \left(\frac{2LS\|B\|^2\beta_T}{C\Delta\sigma_\zeta}\right)^2 + \frac{32\log(\frac{d}{\delta})}{\sigma_\zeta^4} + \frac{8LS\|B\|^2\beta_T\sqrt{2\log(\frac{d}{\delta})}}{C\Delta\sigma_\zeta^3}. \tag{C.8.15}$$

Proof: Let τ be any round that the algorithm plays the conservative action, i.e., at round τ, either $F = 0$ or $\lambda_{\min}(V_\tau) < \left(\frac{2L\|B\|\beta_\tau}{\Delta}\right)^2$.

By definition, if $F = 0$, we have

$$\nexists x \in \mathcal{X} : x^\top B\hat{\theta}_\tau + \beta_\tau\|Bx\|_{V_\tau^{-1}} \leq C, \tag{C.8.16}$$

and since we know that $x_\star \in \mathcal{X}$, and $\theta_\star \in \mathcal{E}_t$ with high probability, we can write

$$x_\star^\top B\theta_\star + 2\beta_\tau\|Bx_\star\|_{V_\tau^{-1}} \geq\ x_\star^\top B\hat{\theta}_\tau + \beta_\tau\|Bx_\star\|_{V_\tau^{-1}} > C. \tag{C.8.17}$$

Then, using the LHS and RHS of (C.8.17), we can get

$$\frac{2L\|B\|\beta_\tau}{\sqrt{\lambda_{\min}(V_\tau)}} \geq 2\beta_\tau \|x_\star\|_{V_\tau^{-1}} \geq \Delta,$$

and hence the following upper bound on minimum eigenvalue of the Gram matrix:

$$\lambda_{\min}(V_\tau) < \left(\frac{2L\|B\|\beta_\tau}{\Delta}\right)^2.$$

Therefore, at any round τ that a conservative action is played, whether it is because $\{F = 0\}$ happens or beccause we have $\{\lambda_{\min}(V_\tau) < \left(\frac{2L\|B\|\beta_\tau}{\Delta}\right)^2\}$, we can always conclude that

$$\lambda_{\min}(V_\tau) < \left(\frac{2L\|B\|\beta_\tau}{\Delta}\right)^2 \tag{C.8.18}$$

The remaining of the proof builds on two auxiliary lemmas. First, in Lemma C.8.4, we show that the minimum eigenvalue of the Gram matrix V_t is lower bounded with the number of times SCLUCB2 plays the conservative actions.

Lemma C.8.4 *On event $\{\theta_\star \in \mathcal{E}_t\}$, it holds that*

$$\mathbb{P}(\lambda_{min}(V_t) \leq t) \leq d \exp\left(-\frac{(\rho^2\sigma_\zeta^2|N_t^c| - t)^2}{32\rho^4|N_t^c|}\right), \tag{C.8.19}$$

where $\rho = \frac{C}{\|B\|S}$.

Using (C.8.18) and applying Lemma C.8.4, it can checked that with probability $1-\delta$

$$\left(\frac{2L\|B\|\beta_\tau}{\Delta}\right)^2 > \rho^2\sigma_\zeta^2|N_\tau^c| - \sqrt{32\rho^4.|N_\tau^c|\log(\frac{d}{\delta})},$$

Then using Lemma C.4.2, we can conclude the following upper bound

$$|N_\tau^c| \leq \left(\frac{2LS\|B\|^2\beta_\tau}{C\Delta\sigma_\zeta}\right)^2 + \frac{32\log(\frac{d}{\delta})}{\sigma_\zeta^4} + \frac{8LS\|B\|^2\beta_\tau\sqrt{2\log(\frac{d}{\delta})}}{C\Delta\sigma_\zeta^3}.$$

■

C.8.1 Proof of Lemma C.8.4

Our objective is to establish a lower bound on $\lambda_{\min}(V_t)$ for all t. It holds that

$$\begin{aligned} V_t &= \lambda I + \sum_{s=1}^{t} x_s x_s^\top \\ &\succeq \sum_{s\in N_t^c} (\rho\zeta_s)(\rho\zeta_s)^\top \\ &= \sum_{s\in N_t^c} \left(\rho^2\mathbb{E}[\zeta_s\zeta_s^\top] + \rho^2\zeta_s\zeta_s^\top - \rho^2\mathbb{E}[\zeta_s\zeta_s^\top]\right) \\ &\succeq \rho^2\sigma_\zeta^2|N_t^c|I + \sum_{s\in N_t^c} G_s, \end{aligned} \tag{C.8.20}$$

where G_s is defined as

$$G_s = \left(\rho^2\zeta_s\zeta_s^\top - \rho^2\mathbb{E}[\zeta_s\zeta_s^\top]\right). \tag{C.8.21}$$

Thus, using Weyl's inequality, it follows that

$$\lambda_{\min}(V_t) \geq \rho^2\sigma_\zeta^2|N_t^c| - \lambda_{\max}(\sum_{s\in N_t^c} G_s).$$

Next, we apply the matrix Azuma inequality (see Theorem C.4.3) to find an upper bound on $\lambda_{\max}(\sum_{s\in N_t^c} G_s)$. For this, we first need to show that the sequence of matrices G_s satisfies the conditions of Theorem C.4.3. By definition of G_s in (C.8.21), it follows

that $\mathbb{E}[G_s|\mathcal{F}_{s-1}] = 0$, and $G_s^\top = G_s$. Also, we construct the sequence of deterministic matrices A_s such that $G_s^2 \preceq A_s^2$ as follows. We know that for any matrix K, $K^2 \leq \|K\|_2^2 I$, where $\|K\|_2$ is the maximum singular value of K, i.e.,

$$\sigma_{\max}(K) = \max_{\|u\|_1=\|v\|_2=1} u^\top K v.$$

Thus, we first show the following bound on the maximum singular value of the matrix G_s defined in (C.8.21):

$$\begin{aligned}\max_{\|u\|_1=\|v\|_2=1} u^\top G_s v &= \rho^2 (u^\top \zeta_s)(v^\top \zeta_s)^\top - \rho^2 \mathbb{E}\left[(u^\top \zeta_s)(v^\top \zeta_s)^\top\right] \\ &\leq \rho^2 \|\zeta_s\|_2^2 + \rho^2 \mathbb{E}\left[\|\zeta_s\|_2^2\right] \\ &\leq 2\rho^2,\end{aligned}$$

where we have used Cauchy-Schwarz inequality and the last inequality comes from the fact that $\|\zeta_s\|_2 = 1$ almost surely. From the derivations above, and choosing $A_s = 2\rho^2 I$, it almost surely holds that $G_s^2 \preceq \sigma_{\max}(G_s)^2 I \preceq 4\rho^4 I = A_s^2$. Moreover, using triangular inequality, it holds that

$$\|\sum_{s\in N_t^c} A_s^2\| \leq \sum_{s\in N_t^c} \|A_s^2\| \leq 4\rho^4 |N_t^c|.$$

Now we can apply the matrix Azuma inequality, to conclude that for any $c \geq 0$,

$$\mathbb{P}\left(\lambda_{\max}(\sum_{s\in N_t^c} G_s) \geq c\right) \leq d \exp\left(-\frac{c^2}{32\rho^4 |N_t^c|}\right).$$

Therefore, it holds that with probability $1-\delta$, $\lambda_{\max}(\sum_{s\in N_t^c} G_s) \leq \sqrt{32\rho^4 |N_t^c| \log(\frac{d}{\delta})}$, and

hence with probability $1-\delta$,

$$\lambda_{\min}(V_t) \geq \rho^2\sigma_\zeta^2|N_t^c| - \sqrt{32\rho^4|N_t^c|\log(\frac{d}{\delta})}, \tag{C.8.22}$$

or equivalently,

$$\mathbb{P}(\lambda_{\min}(V_t) \leq t) \leq d\exp\left(-\frac{(\rho^2\sigma_\zeta^2|N_t^c| - t)^2}{32\rho^4|N_t^c|}\right), \tag{C.8.23}$$

where $\rho = \frac{C}{\|B\|S}$. This completes the proof.

Appendix D

Supplements to Chapter 6

D.1 Sequential Prediction Algorithm

The sequential prediction algorithm `SqAlg` uses the following algorithm from [136] (also see [156, Chapter 3]) to aggregate its experts' predictions. Algorithm 12 takes the observations y_t and experts' predictions $f_t^i(H_t)$ that are bounded in the known range $[\beta, \beta + \ell]$ as input. It first scales these input to the range $[0, 1]$ and uses its current weights for the experts to generate its own prediction $\widehat{y}_t$.

The performance of `SqAlg` is evaluated as the accuracy (in terms of square loss) of its prediction w.r.t. the accuracy of the prediction by the best expert in the set, i.e.,

$$\sum_{t=1}^{T}(\widehat{y}_t - y_t)^2 - \min_{i\in[M]}\sum_{t=1}^{T}(f_t^i(H_t) - y_t)^2 \leq \mathcal{R}_{\texttt{Sq}}(T). \tag{D.1.1}$$

We call this the regret of `SqAlg` and denote it by $\mathcal{R}_{\texttt{Sq}}(T)$. [136] prove the following bound for $\mathcal{R}_{\texttt{Sq}}(T)$, which we use in the analysis of our algorithms.

Proposition D.1.1 (Theorem 4.2 in [136]) *For any arbitrary sequence* $\left\{\left(\{f_t^i(H_t)\}_{i=1}^M, \widehat{y}_t, y_t\right)\right\}_{t=1}^T$ *in which the experts' predictions* $\left\{\{f_t^i(H_t)\}_{i=1}^M\right\}_{t=1}^T$ *and observations* $\{y_t\}_{t=1}^T$ *are all bounded*

in $[\beta, \beta+\ell]$*, the regret defined by* (D.1.1) *of Algorithm 12 is bounded as*

$$\mathcal{R}_{\mathsf{Sq}}(T) \leq 2\ell^2 \log M.$$

Here we use the fact that $|\mathcal{S}_t| \leq M,\ \forall t \in [T]$.

Algorithm 12: Sequential Prediction with Expert Advice

Input: ℓ and β *(experts' predictions* $f_t^i(H_t)$ *are bounded in the known range* $[\beta, \beta+\ell]$*)*

Initialization: Set the weight $w_{1,i} = 1$ for all experts $i \in [M]$

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

Receive predictions $f_t^i(H_t)$ by experts $i \in \mathcal{S}_{t-1}$

Remove experts whose predictions are out of bound and construct the new set of admissible experts $\mathcal{S}_t$ *(see Remarks 6.3.1 and 6.4.1)*

Scale experts' predictions $h_{i,t} = \frac{f_t^i(H_t)-\beta}{\ell},\ \forall i \in \mathcal{S}_t$

Set $v_{t,i} = \frac{w_{t,i}}{W},\ \forall i \in \mathcal{S}_t$, where $W = \sum_{i\in\mathcal{S}_t} w_{t,i}$

Prediction: Compute:

$$\Delta(0) = \frac{-1}{2}\log\left(\sum_{i\in\mathcal{S}_t} v_{t,i} e^{-2h_{i,t}^2}\right) \quad , \quad \Delta(1) = \frac{-1}{2}\log\left(\sum_{i\in\mathcal{S}_t} v_{t,i} e^{-2(1-h_{i,t})^2}\right)$$

Predict a value $\widehat{y}'_t$ that satisfies the following conditions:

$$(\widehat{y}'_t)^2 \leq \Delta(0) \quad , \quad (1-\widehat{y}'_t)^2 \leq \Delta(1).$$

Update: Observing reward y_t, scale it as $y'_t = \frac{y_t - a}{\ell}$, and update the experts' weights

$$w_{t+1,i} = w_{t,i} e^{-2(y'_t - h_{i,t})^2} \tag{D.1.2}$$

Return prediction $\widehat{y}_t = \beta + \ell\widehat{y}'_t$

end for

D.2 Proofs of Section 6.3

In this section, we first provide a brief overview for the steps of our proof. Then, we provide the proofs of lemmas used in Section 6.3.

The performance analysis of the FS-SCB algorithm requires two steps. First, we control the sum of the prediction error of the agent. Second, we show how the regret is related to the prediction error of the agent, and then we bound the regret.

Step 1. To control the sum of the prediction error of the agent D_t, we need to find two upper bounds: 1) an upper bound on the prediction error of the true model i_* whose identity is unknown to the agent Q_t; 2) an upper bound on the regret caused by the online regression oracle $\mathcal{R}_{\mathtt{Sq}}$

First, in Lemma 6.3.4, we bound the sum of the prediction error of the true model as

$$\sum_{s=1}^{t-1}\left(\langle\phi^{i_*}(x_s,a_s),\widehat{\theta}_s^{i_*}\rangle-\langle\phi^{i_*}(x_s,a_s),\theta_*^{i_*}\rangle\right)^2\leq Q_t \tag{D.2.1}$$

where

$$Q_t=1+2\left(\max_{i\in[M]}\left\{\lambda_i S^2+4d\log\left(1+\frac{tL^2}{\lambda_i d}\right)\right\}\right)+ 32R^2\log\left(\frac{R\sqrt{8}+\sqrt{1+\max_{i\in[M]}\left\{\lambda_i S^2+4d\log\left(1+\frac{tL^2}{\lambda_i d}\right)\right\}}}{\delta}\right).$$

Next, in Lemma 6.3.5, we provide a high probability upper-bound on the regret caused by the online regression oracle as

$$\mathcal{R}_{\mathtt{Sq}}(t)\leq 8(\log M)R^2L^2\left(G^2+\max_{i\in[M]}\{\lambda_i S^2+d\log\left(1+\frac{tL^2}{\lambda_i d}\right)\}+\log(1/\delta)\right).$$

Then, in Lemma 6.3.3, we show the following upper bound on the sum of prediction error

of the agent:

$$D_t(\delta) \leq 1 + 2\mathcal{R}_{\texttt{Sq}}(t) + 2Q_t + 4R\sqrt{2(1+Q_t)\log\left(\frac{\sqrt{1+Q_t}}{\delta}\right)} + 32R^2 \log\left(\frac{R\sqrt{8} + \sqrt{1 + \mathcal{R}_{\texttt{Sq}}(t) + Q_t + 2R\sqrt{2(1+Q_t)\log\left(\frac{\sqrt{1+Q_t}}{\delta}\right)}}}{\delta}\right). \tag{D.2.2}$$

Step 2. First in Lemma 6.3.6, we show how the regret is related to the prediction error of the agent using the Azuma's inequality, i.e.,

$$\mathcal{R}_{\text{FS-SCB}}(T) \leq \sqrt{2T\log(2/\delta)} + \alpha D_T(\delta) \tag{D.2.3}$$
$$+ \sum_{t=1}^{T} \sum_{a\in[K]} p_t(x_t, a)\left(\langle \phi^{i_*}(x_t, a_t^*), \theta_*^{i_*}\rangle - \langle \phi^{i_*}(x_t, a), \theta_*^{i_*}\rangle - \frac{\alpha}{4}\left(\widehat{y}_t(x_t, a) - \langle \phi^{i_*}(x_t, a), \theta_*^{i_*}\rangle\right)^2\right).$$

Then in Appendix D.2.5, we put everything together and complete the proof.

D.2.1 Proof of Lemma 6.3.4

At each round t, each expert $i_* \in \mathcal{I}_*$ estimates its reward parameter as

$$\widehat{\theta}_t^{i_*} = \arg\min_{\theta} \left\|(\Phi_t^{i_*})^\top \theta - Y_t\right\|_2^2 + \lambda_{i_*} \|\theta\|_2^2. \tag{D.2.4}$$

Let $V_t^{\lambda_{i_*}} = \lambda_{i_*} I + \sum_{s=1}^{t-1} \phi^{i_*}(x_s, a_s)\phi^{i_*}(x_s, a_s)^\top$. From the standard least-squares analysis, we have

$$\sum_{s=1}^{t-1} \left(\langle \phi^{i_*}(x_s, a_s), \widehat{\theta}_s^{i_*} \rangle - y_s\right)^2 - \sum_{s=1}^{t-1} \left(\langle \phi^{i_*}(x_s, a_s), \theta_*^{i_*} \rangle - y_s\right)^2 \leq \lambda_{i_*} \left\|\theta_*^{i_*}\right\|_2^2 + 2 \sum_{s=1}^{t-1} \left\langle \phi^{i_*}(x_s, a_s)^\top, (V_s^{\lambda_{i_*}})^{-1} \phi^{i_*}(x_s, a_s) \right\rangle.$$

Therefore, we can write:

$$\sum_{s=1}^{t-1} \left(\langle \phi^{i_*}(x_s, a_s), \widehat{\theta}_s^{i_*} - \theta_*^{i_*} \rangle\right)^2 \leq \lambda_{i_*} \left\|\theta_*^{i_*}\right\|_2^2 + 2 \log\left(\frac{\det(V_t^{\lambda_{i_*}})}{\det(\lambda_{i_*} I)}\right) + 2 \sum_{s=1}^{t-1} \eta_s \left(\langle \phi^{i_*}(x_s, a_s), \widehat{\theta}_s^{i_*} - \theta_*^{i_*} \rangle\right). \tag{D.2.5}$$

The last term on the RHS of (D.2.5) can be bounded using Proposition D.4.1 in Appendix D.4 as

$$\left| \sum_{s=1}^{t-1} \eta_s \left(\langle \phi^{i_*}(x_s, a_s), \widehat{\theta}_s^{i_*} - \theta_*^{i_*} \rangle\right) \right| \leq R \sqrt{2 \left(1 + \sum_{s=1}^{t-1} \left(\langle \phi^{i_*}(x_s, a_s), \widehat{\theta}_s^{i_*} - \theta_*^{i_*} \rangle\right)^2\right) \log\left(\frac{1 + \sum_{s=1}^{t-1} \left(\langle \phi^{i_*}(x_s, a_s), \widehat{\theta}_s^{i_*} - \theta_*^{i_*} \rangle\right)^2}{\delta}\right)}. \tag{D.2.6}$$

Define $u = \sqrt{1 + \sum_{k=1}^{t-1} \left(\langle \phi^{i_*}(x_k, a_k), \widehat{\theta}_k^{i_*} - \theta_*^{i_*} \rangle\right)^2}$, $v = 1 + \lambda_{i_*} \left\|\theta_*^{\lambda_{i_*}}\right\|_2^2 + 2 \log\left(\frac{\det(V_t^{i_*})}{\det(\lambda_{i_*} I)}\right)$, and $w = 2R\sqrt{2 \log(s/\delta)}$. It is easy to see that (D.2.6) can be written in the form of $u^2 \leq v + uq$. Then, by applying Lemma D.4.5 in Appendix D.4, we may write $u \leq \sqrt{v} + w$. Substituting for w, we can get $u \leq \sqrt{v} + 2R\sqrt{2 \log(u/\delta)}$. Then, by

Lemma D.4.6 in Appendix D.4, for $\delta \in (0, 1/4]$, we have

$$u \leq \sqrt{v} + 4R\sqrt{\log\left(\frac{2\sqrt{2}R + \sqrt{v}}{\delta}\right)},$$

which using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, for any a and b, we can write it as

$$u^2 \leq 2v + 32R^2 \log\left(\frac{2\sqrt{2}R + \sqrt{v}}{\delta}\right).$$

Finally, we substitute u and v, and subtract 1 from both sides, and for $\delta \in (0, 1/4]$, we obtain

$$\sum_{s=1}^{t-1} \left(\langle \phi^{i_*}(x_s, a_s), \widehat{\theta}_s^{i_*} - \theta_*^{i_*} \rangle\right)^2 \leq 1 + 2\lambda_{i_*} \left\|\theta_*^{i_*}\right\|_2^2 + 4\log\left(\frac{\det(V_t^{\lambda_{i_*}})}{\det(\lambda_{i_*} I)}\right) + 32R^2 \log\left(\frac{R\sqrt{8} + \sqrt{1 + \lambda \left\|\theta_*^{i_*}\right\|_2^2 + 2\log\left(\frac{\det(V_t^{\lambda_{i_*}})}{\det(\lambda_{i_*} I)}\right)}}{\delta}\right). \tag{D.2.7}$$

We know that $\|\theta_*^{i_*}\|_2^2 \leq S^2$. Moreover, by Lemma D.4.3 in Appendix D.4, we can bound the term $\log\left(\frac{\det(V_t^{\lambda_{i_*}})}{\det(\lambda_{i_*} I)}\right)$. Replacing these in (D.2.7), we may write

$$\sum_{s=1}^{t-1} \left(\langle \phi^{i_*}(x_s, a_s), \widehat{\theta}_s^{i_*} - \theta_*^{i_*} \rangle\right)^2 \leq 1 + 2\lambda_{i_*} S^2 + 8d\log\left(1 + \frac{tL^2}{\lambda_{i_*} d}\right) + 32R^2 \log\left(\frac{R\sqrt{8} + \sqrt{1 + \lambda_{i_*} S^2 + 4d\log\left(1 + \frac{tL^2}{\lambda_{i_*} d}\right)}}{\delta}\right). \tag{D.2.8}$$

Since the algorithm does not know the identity of i_*, we derive an expression for Q_t, and conclude the proof by replacing i_* with the maximum over all $i \in [M]$ in (D.2.8) as

$$\sum_{s=1}^{t-1}\left(\langle \phi^{i_*}(x_s,a_s), \widehat{\theta}_s^{i_*} - \theta_*^{i_*}\rangle\right)^2 \leq 1 + 2\left(\max_{i\in[M]}\left\{\lambda_i S^2 + 4d\log\left(1+\frac{tL^2}{\lambda_i d}\right)\right\}\right) + 32R^2\log\left(\frac{R\sqrt{8}+\sqrt{1+\max_{i\in[M]}\left\{\lambda_i S^2 + 4d\log\left(1+\frac{tL^2}{\lambda_i d}\right)\right\}}}{\delta}\right) := Q_t. \quad \text{(D.2.9)}$$

D.2.2 Proof of Lemma 6.3.5

To bound the regret $\mathcal{R}_{\texttt{Sq}}(t)$ of the regression oracle `SqAlg`, similar to the proof of Lemma 6.4.5 in Appendix D.3.2, we show the reward signals and the experts' predictions are bounded with high probability. Then, we use Proposition D.1.1 in Appendix D.1 to complete the proof.

From (D.3.10), according to Assumption 13, we have $\langle \phi^{i_*}(x,a), \theta_*^{i_*}\rangle \leq LS$, for all $x \in \mathcal{X}$, $a \in [k]$, and $i \in [M]$. Hence, with probability at least $1-\delta$, we have

$$y_t \in \left[-\left(G + R\sqrt{2\log(2/\delta)}\right), \left(G + R\sqrt{2\log(2/\delta)}\right)\right]. \quad \text{(D.2.10)}$$

Next we bound the predictions of the experts that FS-SCB considers in its prediction. To do so, we first show an upper bound on the prediction of the any true model i_*. In particular, we can write for $t \in [T]$ and $\forall a \in [K]$:

$$\begin{aligned}\left|\langle \phi^{i_*}(x_t,a_t), \widehat{\theta}_t^{i_*}\rangle\right| &= \left|\langle \phi^{i_*}(x_t,a_t), \theta_*^{i_*}\rangle + \langle \phi^{i_*}(x_t,a_t), \widehat{\theta}_t^{i_*} - \theta_*^{i_*}\rangle\right| \\ &\overset{(a)}{\leq} \left|\langle \phi^{i_*}(x_t,a_t), \theta_*^{i_*}\rangle\right| + \left|\langle \phi^{i_*}(x_t,a_t), \widehat{\theta}_t^{i_*} - \theta_*^{i_*}\rangle\right| \\ &\overset{(b)}{\leq} G + \left\|\phi^{i_*}(x_t,a_t)\right\|_{(V_t^{\lambda_{i_*}})^{-1}}\left(\left\|\Phi_t^{i_*}\eta_t\right\|_{(V_t^{\lambda_{i_*}})^{-1}} + \sqrt{\lambda_{i_*}}S\right), \end{aligned} \quad \text{(D.2.11)}$$

(a) It results from a triangular inequality. **(b)** This is because of the Assumption 13, and

the fact that the true model is linearly realizable, we can apply Theorem 2 in [42]. Then, we use Theorem 1 in [42] and standard matrix analysis together with our assumption that $\|\phi^{i_*}(x_t, a_t)\| \leq L$, and bound the terms on the RHS of (D.2.11) with high probability as

$$\left\|\Phi_t^{i_*}\eta_t\right\|_{(V_t^{\lambda_{i_*}})^{-1}} \leq R\sqrt{2\log\left(\frac{\sqrt{\det(V_t^{\lambda_{i_*}})}}{\delta\sqrt{\det(\lambda_{i_*}I)}}\right)}, \tag{D.2.12}$$

and

$$\left\|\phi^{i_*}(x_t, a_t)\right\|_{(V_t^{\lambda_{i_*}})^{-1}} \leq \frac{\|\phi^{i_*}(x_t, a_t)\|}{\sqrt{\lambda_{\min}(V_t^{\lambda_{i_*}})}} \leq \frac{L}{\sqrt{\lambda_{i_*}}} \leq L, \tag{D.2.13}$$

where $\lambda_{\min}(V_t^{\lambda_{i_*}})$ is the smallest eigenvalue of the matrix $V_t^{\lambda_{i_*}}$. In the last step of (D.2.13), we use the fact that $\lambda_i \geq 1,\ \forall i \in [M]$. Putting Eqs. D.2.11, D.2.12, and D.2.13 together, with probability at least $1-\delta$, we have

$$\left|\langle\phi^{i_*}(x_t, a_t), \widehat{\theta}_t^{i_*}\rangle\right| \leq G + RL\sqrt{2\log\left(\frac{\sqrt{\det(V_t^{\lambda_{i_*}})}}{\delta\sqrt{\det(\lambda_{i_*}I)}}\right)} + L\sqrt{\lambda_i}S. \tag{D.2.14}$$

Using Lemma D.4.3 in Appendix D.4, we may write (D.2.14) as

$$\left|\langle\phi^{i_*}(x_t, a_t), \widehat{\theta}_t^{i_*}\rangle\right| \leq G + RL\sqrt{d\log\left(\frac{1+\frac{tL^2}{\lambda_{i_*}d}}{\delta}\right)} + L\sqrt{\lambda_{i_*}}S. \tag{D.2.15}$$

FS-SCB employees this idea that at any time step $t \in [T]$, any potentially true model (i.e., linearly realizable) should have a similar bound on its prediction. To do so, the set of admissible experts, $\mathcal{S}_t$, only considers experts that have the following bound on their

prediction at each time $t \in [T]$ and $\forall a \in [K]$ as:

$$\left|\langle \phi^i(x_t, a_t), \widehat{\theta}_t^i \rangle\right| \leq G + RL\sqrt{d \log\left(\frac{1 + \frac{tL^2}{\lambda_i d}}{\delta}\right)} + L\sqrt{\lambda_i} S. \tag{D.2.16}$$

If at some time step t, this bound does not hold for any expert i, then the algorithm simply eliminates that expert from the set of admissible experts, since that model is not a true model (i.e., the reward is not in the linear span of the prediction of that expert), and that expert will remain out for the rest of the game. Then, we may bound the range of the prediction of each expert $i \in \mathcal{S}_t$ at round $t \in [T]$ as

$$\langle \phi^i(x_t, a_t), \widehat{\theta}_t^i \rangle \in$$
$$\left[-\left(G + RL\sqrt{d \log\left(\frac{1 + \frac{tL^2}{\lambda_i d}}{\delta}\right)} + L\sqrt{\lambda_i} S\right), \left(G + RL\sqrt{d \log\left(\frac{1 + \frac{tL^2}{\lambda_i d}}{\delta}\right)} + L\sqrt{\lambda_i} S\right)\right]. \tag{D.2.17}$$

Putting together (D.2.10) and (D.2.17), we conclude that for all rounds $t \in [T]$ and experts $i \in \mathcal{S}_t$, with probability at least $1 - \delta$, the reward y_t and the expert's predictions $f_t^i(H_t)$ are in the range $[\beta, \beta + \ell]$ for

$$\beta = -\left(G + RL\sqrt{d \log\left(\frac{1 + \frac{tL^2}{\lambda_i d}}{\delta}\right)} + L\sqrt{\lambda_i} S\right), \ell = 2\left(G + RL\sqrt{d \log\left(\frac{1 + \frac{tL^2}{\lambda_i d}}{\delta}\right)} + L\sqrt{\lambda_i} S\right). \tag{D.2.18}$$

Using Proposition D.1.1 in Appendix D.1 with the bound on the observations and predictions in (D.2.18), with probability at least $1 - \delta$, we obtain the following regret bound for `SqAlg`:

$$\mathcal{R}_{\texttt{Sq}}(t) = 8R^2L^2 \log(M)\left(G^2 + \max_{i \in [M]}\left\{\lambda_i S^2 + d \log\left(1 + \frac{tL^2}{\lambda_i d}\right)\right\} + \log(1/\delta)\right), \tag{D.2.19}$$

in which we use the fact that for $a, b > 0$, $(a+b)^2 \leq 2a^2 + 2b^2$. This concludes our proof.

D.2.3 Proof of Lemma 6.3.3

Here, we bound the sum of the square loss of the oracle predictions, i.e.,

$$\sum_{s=1}^{t-1} \left(\widehat{y}_s(x_s, a_s) - \langle \phi^{i_*}(x_s, a_s), \theta_*^{i_*} \rangle\right)^2 \leq D_t(\delta). \tag{D.2.20}$$

We know that $y_t = \langle \phi^{i_*}(x_t, a_t), \theta_*^{i_*} \rangle + \eta_t$. Hence we can write

$$\begin{aligned}
&\left(\widehat{y}_t(x_t, a_t) - y_t\right)^2 - \left(\langle \phi^{i_*}(x_t, a_t), \widehat{\theta}_t^{i_*} \rangle - y_t\right)^2 = \\
&\qquad \left(\widehat{y}_t(x_t, a_t) - \langle \phi^{i_*}(x_t, a_t), \theta_*^{i_*} \rangle - \eta_t\right)^2 - \left(\langle \phi^{i_*}(x_t, a_t), \widehat{\theta}_t^{i_*} \rangle - \langle \phi^{i_*}(x_t, a_t), \theta_*^{i_*} \rangle - \eta_t\right)^2 \\
&= \left(\widehat{y}_t(x_t, a_t) - \langle \phi^{i_*}(x_t, a_t), \theta_*^{i_*} \rangle\right)^2 - \left(\langle \phi^{i_*}(x_t, a_t), \widehat{\theta}_t^{i_*} \rangle - \langle \phi^{i_*}(x_t, a_t), \theta_*^{i_*} \rangle\right)^2 \\
&\qquad + 2\eta_t\left(\langle \phi^{i_*}(x_t, a_t), \widehat{\theta}_t^{i_*} \rangle - \widehat{y}_t(x_t, a_t)\right) \\
&= \left(\widehat{y}_t(x_t, a_t) - \langle \phi^{i_*}(x_t, a_t), \theta_*^{i_*} \rangle\right)^2 - \left(\langle \phi^{i_*}(x_t, a_t), \widehat{\theta}_t^{i_*} \rangle - \langle \phi^{i_*}(x_t, a_t), \theta_*^{i_*} \rangle\right)^2 \\
&\qquad + 2\eta_t\left(\langle \phi^{i_*}(x_t, a_t), \widehat{\theta}_t^{i_*} \rangle - \langle \phi^{i_*}(x_t, a_t), \theta_*^{i_*} \rangle\right) + 2\eta_t\left(\langle \phi^{i_*}(x_t, a_t), \theta_*^{i_*} \rangle - \widehat{y}_t(x_t, a_t)\right).
\end{aligned} \tag{D.2.21}$$

Then, from Proposition D.4.1 in Appendix D.4, with probability at least $1 - \delta$, we have

$$\begin{aligned}
&\left|\sum_{s=1}^{t-1} \eta_s\left(\langle \phi^{i_*}(x_s, a_s), \widehat{\theta}_s^{i_*} - \theta_*^{i_*} \rangle\right)\right| \leq \\
&\qquad R\sqrt{2\left(1 + \sum_{s=1}^{t-1} \left(\langle \phi^{i_*}(x_s, a_s), \widehat{\theta}_s^{i_*} - \theta_*^{i_*} \rangle\right)^2\right) \log\left(\frac{\sqrt{1 + \sum_{s=1}^{t-1} \left(\langle \phi^{i_*}(x_s, a_s), \widehat{\theta}_s^{i_*} - \theta_*^{i_*} \rangle\right)^2}}{\delta}\right)},
\end{aligned} \tag{D.2.22}$$

and

$$\left|\sum_{s=1}^{t-1}\eta_s\big(\langle\phi^{i_*}(x_s,a_s),\theta_*^{i_*}\rangle-\widehat{y}_s(x_s,a_s)\big)\right|\leq$$
$$R\sqrt{2\left(1+\sum_{s=1}^{t-1}\big(\langle\phi^{i_*}(x_s,a_s),\theta_*^{i_*}\rangle-\widehat{y}_s\big)^2\right)\log\left(\frac{\sqrt{1+\sum_{s=1}^{t-1}\big(\langle\phi^{i_*}(x_s,a_s),\theta_*^{i_*}\rangle-\widehat{y}_s\big)^2}}{\delta}\right)}. \tag{D.2.23}$$

Using (D.2.22) and (D.2.23), the upper-bound $\mathcal{R}_{\mathtt{Sq}}(t)$ from (D.2.19) in Appendix D.2.2, and the upper-bound Q_t on the square error of the prediction of the true model in (D.2.9) in Appendix D.2.1, we may write (D.2.21) as

$$\sum_{s=1}^{t-1}\big(\widehat{y}_s(x_s,a_s)-\langle\phi^{i_*}(x_s,a_s),\theta_*^{i_*}\rangle\big)^2\leq\mathcal{R}_{\mathtt{Sq}}(t)+Q_t+2R\sqrt{2(1+Q_t)\log\left(\frac{\sqrt{1+Q_t}}{\delta}\right)}$$
$$+2R\sqrt{2\Big(1+\sum_{s=1}^{t-1}\big(\langle\phi^{i_*}(x_s,a_s),\theta_*^{i_*}\rangle-\widehat{y}_s(x_s,a_s)\big)^2\Big)\log\left(\frac{\sqrt{1+\sum_{s=1}^{t-1}\big(\langle\phi^{i_*}(a_s),\theta_*^{i_*}\rangle-\widehat{y}_s(a_s)\big)^2}}{\delta}\right)}. \tag{D.2.24}$$

Let $u=\sqrt{1+\sum_{k=1}^{t-1}\big(\widehat{y}_k(x_k,a_k)-\langle\phi^{i_*}(x_k,a_k),\theta_*^{i_*}\rangle\big)^2}$, $v=1+\mathcal{R}_{\mathtt{Sq}}(t)+Q_t+2R\sqrt{2(1+Q_t)\log(\frac{\sqrt{1+Q_t}}{\delta})}$, and $q=2R\sqrt{2\log(s/\delta)}$. Then, following the same machinery as the one in the proof of Lemma 6.3.4 in Section D.2.1, and with the use of Lemmas D.4.5 and D.4.6, for

$\delta \in (0, 1/4]$, with probability at least $1-\delta$, we have

$$\sum_{s=1}^{t-1} \left(\widehat{y}_s(x_s, a_s) - \langle \phi^{i_*}(x_s, a_s), \theta_*^{i_*} \rangle\right)^2 \leq 1 + 2\mathcal{R}_{\mathtt{Sq}}(t) + 2Q_t + 4R\sqrt{2(1+Q_t)\log\left(\frac{\sqrt{1+Q_t}}{\delta}\right)}$$
$$+ 32R^2 \log\left(\frac{R\sqrt{8} + \sqrt{1 + \mathcal{R}_{\mathtt{Sq}}(t) + Q_t + 2R\sqrt{2(1+Q_t)\log\left(\frac{\sqrt{1+Q_t}}{\delta}\right)}}}{\delta}\right) := D_t(\delta), \tag{D.2.25}$$

where

$$Q_t = 1 + 2\left(\max_{i\in[M]}\left\{\lambda_i S^2 + 4d\log\left(1 + \frac{tL^2}{\lambda_i d}\right)\right\}\right) +$$
$$32R^2 \log\left(\frac{R\sqrt{8} + \sqrt{1 + \max_{i\in[M]}\left\{\lambda_i S^2 + 4d\log\left(1+\frac{tL^2}{\lambda_i d}\right)\right\}}}{\delta}\right),$$

and

$$\mathcal{R}_{\mathtt{Sq}}(t) \leq 8R^2L^2\log(M)\left(G^2 + \max_{i\in[M]}\left\{\lambda_i S^2 + d\log\left(1 + \frac{tL^2}{\lambda_i d}\right)\right\} + \log(1/\delta)\right).$$

D.2.4 Proof of Lemma 6.3.6

The inequality can be obtained using Azuma's inequality and following similar steps as in Lemma 2 of [131]. We may write the regret as

$$\mathcal{R}_{\text{FS-SCB}}(T) = \sum_{t=1}^{T}\left(\langle \phi^{i_*}(x_t, a_t^*), \theta_*^{i_*}\rangle - \langle \phi^{i_*}(x_t, a_t), \theta_*^{i_*}\rangle - \frac{\alpha}{4}\left(\widehat{y}_t(x_t, a_t) - \langle \phi^{i_*}(x_t, a_t), \theta_*^{i_*}\rangle\right)^2\right)$$
$$+ \frac{\alpha}{4}\sum_{t=1}^{T}\left(\widehat{y}_t(x_t, a_t) - \langle \phi^{i_*}(x_t, a_t), \theta_*^{i_*}\rangle\right)^2. \tag{D.2.26}$$

The last term on the RHS of (D.2.26) is bounded with $D_t(\delta)$ in (D.2.25) from the result of Lemma 6.3.3 in Appendix D.2.3. Define filtration $F_{t-1} = \sigma\big((x_1, a_1, y_1), \ldots, (x_{t-1}, a_{t-1}, y_{t-1})\big)$. On the RHS of (D.2.26), the action a_t is random. We can use the Azuma's inequality and with probability at least $1-\delta$, upper-bound the first term on the RHS of (D.2.26) with its expectation counterparts using the probability distribution p_t as

$$\mathcal{R}_{\text{FS-SCB}}(T) \leq \sqrt{2T\log(2/\delta)} + \frac{\alpha}{4} D_T \tag{D.2.27}$$
$$+ \sum_{t=1}^{T} \sum_{a\in[K]} p_t(a) \left(\langle \phi^{i_*}(x_t, a_t^*), \theta_*^{i_*} \rangle - \langle \phi^{i_*}(x_t, a), \theta_*^{i_*} \rangle - \frac{\alpha}{4} \left(\widehat{y}_t(x_t, a) - \langle \phi^{i_*}(x_t, a_t), \theta_*^{i_*} \rangle \right)^2 \right).$$

D.2.5 Proof of Theorem 6.3.2

We first state the following lemma from [131] to bound the first term on the RHS of (D.2.28).

Lemma D.2.1 (Lemma 3 in [131]) *Under Assumption 13, for the probability distribution $p_t \in \Delta_K$ defined in the FS-SCB algorithm, we may write*

$$\sum_{a\in[K]} p_t(a) \left(\langle \phi^{i_*}(x_t, a_t^*), \theta_*^{i_*} \rangle - \langle \phi^{i_*}(x_t, a), \theta_*^{i_*} \rangle - \frac{\alpha}{4} \left(\widehat{y}_t(x_t, a) - \langle \phi^{i_*}(x_t, a_t), \theta_*^{i_*} \rangle \right)^2 \right) \leq \frac{2K}{\alpha}.$$

Putting everything together, with the choice of $\alpha = \sqrt{KT/D_T(\delta)}$, with probability at least $1-\delta$, we can show the following upper-bound on the regret of the FS-SCB algorithm:

$$\mathcal{R}_{\textit{FS-SCB}}(T) \leq 3\sqrt{KT D_T(\delta)} + \sqrt{2T\log(2/\delta)} \tag{D.2.28}$$

Here the upper-bound is of order

$$\mathcal{R}_{\textit{FS-SCB}}(T) \leq \mathcal{O}\left(\sqrt{2T\log(2/\delta)} + RLG\sqrt{KT(1+\log(M))\max_{i\in[M]}\left\{\lambda_i S^2 + 4d\log\left(\frac{1+\frac{TL^2}{\lambda_i d}}{\delta}\right)\right\}}\right).$$

D.3 Proofs of Section 6.4

The regret analysis of the PS-OFUL algorithms requires two steps. First, in Theorem 6.4.3, we show that the confidence set $\mathcal{C}_t$ is valid at each round t, i.e., for any $t, \delta > 0$, it includes the reward parameter θ_* with probability at least $1-\delta$. Second, we show how the regret is related to the valid confidence set, and then using Lemmas D.4.2 and D.4.3 complete the proof.

Step 1. The key idea for showing the validity of the confidence set $\mathcal{C}_t$ requires controlling the square prediction error of the online regression oracle $\widehat{y}_t$, i.e., upper-bounding γ_t. In Appendix D.3.3, we show that we can relate this distance to the sum of two terms: $\gamma_t \leq \mathcal{O}\left(U_t + \mathcal{R}_{\texttt{Sq}}(t)\right)$, and then show how we can bound each of them.

1) Bounding U_t: Lemma 6.4.4 shows the worst-case upper-bound on the square error of the prediction of true model i_*, given that the agent does not know the identity of the true model:

$$\sum_{s=1}^{t} \langle \phi_s(a_s), \widehat{\theta}_s^{i_*} - \theta_* \rangle^2 \leq U_t \tag{D.3.1}$$

where

$$U_t \leq 3 + 8d \log\left(1 + \frac{tL^2 \max_{i\in[M]}(b_i + c_i)^2}{d}\right) + 32R^2 \log(1/\delta) \tag{D.3.2}$$

Proof: The proof is provided in Appendix D.3.1. ■

2) Bounding $\mathcal{R}_{\texttt{Sq}}(t)$: In Lemma 6.4.5, we prove an upper-bound on the regret caused by the prediction oracle `SqAlg`, given our proposed expert predictions as (see Appendix D.3.2

for details).

$$\begin{aligned}\mathcal{R}_{\mathtt{Sq}}(t) \leq & 8(G+L)^2 \log(M) + 8R^2L^2 d \log(M) \log(1/\delta) \\ & + 8R^2L^2 d \log(M) \log\left(1 + \frac{tL^2 \max_{i\in[M]}(b_i + c_i)^2}{d}\right)\end{aligned}$$

Putting these together, in Appendix D.3.3, we prove Theorem 6.4.3 that shows the validity of the confidence set $\mathcal{C}_t$.

Step 2. In Appendix D.3.4, we first show how regret is related to the confidence set. In particular, we show that given the validity of of the confidence set $\mathcal{C}_t$, i.e., for any $\delta \in (0, 1/4]$, with probability at least $1-\delta$, $\theta_* \in \mathcal{C}_t$, we can bound the regret as

$$\mathcal{R}_{\text{PS-OFUL}}(T) \leq 2Gd + 2\max\{1, G\}\sqrt{2dT \log\left(1 + \frac{T}{d}\right) \max_{d<t\leq T} \gamma_t(\delta)}.$$

Then, in Appendix D.3.5, we set $\lambda_i = \frac{1}{(b_i+c_i)^2}$, for each $i \in [M]$, and use Lemmas D.4.2 and D.4.3 to complete the proof of Theorem 6.4.2. Here we prove a regret upper-bound of order

$$\mathcal{O}\left(dRL\max\{1, G\}\sqrt{1+\log(M)} \times \sqrt{T \log\left(1 + \frac{T}{d}\right) \log\left(\frac{1 + \frac{TL^2 \max_{i\in[M]}(b_i+c_i)^2}{d}}{\delta}\right)}\right).$$

D.3.1 Proof of Lemma 6.4.4

At each round $s \in [T]$, each expert $i_* \in \mathcal{I}_*$ estimates its reward parameter as

$$\widehat{\theta}_s^{i_*} = \arg\min_{\theta} \left\|\Phi_s^\top \theta - Y_s\right\|^2 + \lambda_{i_*} \left\|\theta - \widehat{\mu}_{i_*}\right\|^2,$$

which is the output of a Follow-The-Regularized-Leader (FTRL) algorithm with quadratic regularizer $\|\theta - \widehat{\mu}_{i_*}\|^2$. Following the standard FTRL analysis of online regression (see

e.g., [156, Chapter 11]), we have

$$\sum_{s=1}^{t}(\langle\phi_s(a_s),\widehat{\theta}_s^{i_*}\rangle-y_s)^2-\sum_{s=1}^{t}(\langle\phi_s(a_s),\theta_*\rangle-y_s)^2 \leq \lambda_{i_*}\left\|\theta_*-\widehat{\mu}_{i_*}\right\|^2+2\sum_{s=1}^{t}\langle\phi_s(a_s),(V_s^{\lambda_{i_*}})^{-1}\phi_s(a_s)\rangle, \tag{D.3.3}$$

where $V_t^{\lambda_{i_*}}=\lambda_{i_*}I+\sum_{s=1}^{t-1}\phi_s(a_s)\phi_s(a_s)^\top$. We may write (D.3.3) as

$$\sum_{s=1}^{t}\langle\phi_s(a_s),\widehat{\theta}_s^{i_*}-\theta_*\rangle^2 \leq \lambda_{i_*}\left\|\theta_*-\widehat{\mu}_{i_*}\right\|^2+2\log\left(\frac{\det(V_t^{\lambda_{i_*}})}{\det(\lambda_{i_*}I)}\right)+2\sum_{s=1}^{t}\eta_s\langle\phi_s(a_s),\widehat{\theta}_s^{i_*}-\theta_*\rangle. \tag{D.3.4}$$

Using Proposition D.4.1 in Appendix D.4, we may bound the last term on the RHS of (D.3.4) as

$$\left|\sum_{s=1}^{t}\eta_s\langle\phi_s(a_s),\widehat{\theta}_s^{i_*}-\theta_*\rangle\right| \leq R\sqrt{2\left(1+\sum_{s=1}^{t}\langle\phi_s(a_s),\widehat{\theta}_s^{i_*}-\theta_*\rangle^2\right)\log\left(\frac{1+\sum_{s=1}^{t}\langle\phi_s(a_s),\widehat{\theta}_s^{i_*}-\theta_*\rangle^2}{\delta}\right)}. \tag{D.3.5}$$

It is easy to see that (D.3.4) can be written in the form $u^2 \leq v+uw$, where $u=\sqrt{1+\sum_{s=1}^{t}\langle\phi_s(a_s),\widehat{\theta}_s^{i_*}-\theta_*\rangle^2}$, $v=1+\lambda_i\|\theta_*-\widehat{\mu}_{i_*}\|^2+2\log\left(\frac{\det(V_t^{\lambda_{i_*}})}{\det(\lambda_{i_*}I)}\right)$, and $w=2R\sqrt{2\log(u/\delta)}$. Then, by applying Lemma D.4.5 in Appendix D.4, we may write $u\leq\sqrt{v}+w$. Substituting for w, we can get $u\leq\sqrt{v}+2R\sqrt{2\log(u/\delta)}$. Then, by Lemma D.4.6 in Appendix D.4, for $\delta\in(0,1/4]$, we have

$$u\leq\sqrt{v}+4R\sqrt{\log\left(\frac{2\sqrt{2}R+\sqrt{v}}{\delta}\right)},$$

which using the inequality $(a+b)^2\leq 2a^2+2b^2$, for any a and b, we can write it as

$$u^2\leq 2v+32R^2\log\left(\frac{2\sqrt{2}R+\sqrt{v}}{\delta}\right).$$

Finally, we substitute u and v, and subtract 1 from both sides, and for $\delta \in (0, 1/4]$, we obtain

$$\sum_{s=1}^{t} \langle \phi_s(a_s), \widehat{\theta}_s^{i_*} - \theta_* \rangle^2 \leq 1 + 2\lambda_{i_*} \|\theta_* - \widehat{\mu}_{i_*}\|^2 + 4 \log \left(\frac{\det(V_t^{\lambda_{i_*}})}{\det(\lambda_{i_*} I)} \right) + 32R^2 \log \left(\frac{2\sqrt{2}R + \sqrt{1 + \lambda_{i_*} \|\theta_* - \widehat{\mu}_{i_*}\|^2 + 2 \log \left(\frac{\det(V_t^{\lambda_{i_*}})}{\det(\lambda_{i_*} I)} \right)}}{\delta} \right). \tag{D.3.6}$$

We know $\|\theta_* - \widehat{\mu}_{i_*}\|^2 \leq (b_{i_*} + c_{i_*})^2$. Moreover, by Lemma D.4.3 in Appendix D.4, we can bound the term $\log \left(\frac{\det(V_t^{\lambda_{i_*}})}{\det(\lambda_{i_*} I)} \right)$. Replacing these terms in (D.3.6), we have

$$\sum_{s=1}^{t} \langle \phi_s(a_s), \widehat{\theta}_s^{i_*} - \theta_* \rangle^2 \leq 1 + 2\lambda_{i_*} (b_{i_*} + c_{i_*})^2 + 8d \log \left(1 + \frac{tL^2}{d\lambda_{i_*}} \right) + 32R^2 \log \left(\frac{2\sqrt{2}R + \sqrt{1 + \lambda_{i_*} (b_{i_*} + c_{i_*})^2 + 4d \log(1 + \frac{tL^2}{d\lambda_{i_*}})}}{\delta} \right). \tag{D.3.7}$$

Setting $\lambda_{i_*} = \frac{1}{(b_{i_*} + c_{i_*})^2}$, as used by the PS-OFUL algorithm, we obtain

$$\sum_{s=1}^{t} \langle \phi_s(a_s), \widehat{\theta}_s^{i_*} - \theta_* \rangle^2 \leq 3 + 8d \log \left(1 + \frac{tL^2 (b_{i_*} + c_{i_*})^2}{d} \right) + 32R^2 \log \left(\frac{2\sqrt{2}R + \sqrt{2 + 4d \log \left(1 + \frac{tL^2 (b_{i_*} + c_{i_*})^2}{d}\right)}}{\delta} \right). \tag{D.3.8}$$

Since the algorithm does not know the identity of i_*, we derive an expression for U_t and conclude the proof by replacing i_* with the maximum over all $i \in [M]$ in (D.3.8), as

$$\sum_{s=1}^{t} \langle \phi_s(a_s), \widehat{\theta}_s^{i_*} - \theta_* \rangle^2 \leq 3 + 8d \log \left(1 + \frac{tL^2 \max_{i\in[M]}(b_i + c_i)^2}{d} \right) + 32R^2 \log \left(\frac{2\sqrt{2}R + \sqrt{2 + 4d \log \left(1 + \frac{tL^2 \max_{i\in[M]}(b_i+c_i)^2}{d}\right)}}{\delta} \right) := U_t. \tag{D.3.9}$$

D.3.2 Proof of Lemma 6.4.5

To obtain a high probability bound on the regret $\mathcal{R}_{\texttt{Sq}}(t)$ of the regression oracle `SqAlg`, we first show that the inputs to the regression oracle, i.e., reward signals $y_t = \phi_t(a_t) + \eta_t$ and the experts' predictions $f_t^i(H_t) = \langle \phi_t(a_t), \widehat{\theta}_t^i \rangle$ are all bounded with high probability. We then use Proposition D.1.1 in Appendix D.1 to complete the proof.

Since each noise η_t is R-sub-Gaussian, from Lemma D.4.4 in Appendix D.4, with probability at least $1 - \delta$, we have that $|\eta_t| \leq R\sqrt{2\log(2/\delta)}$. We also have from Assumption 14 that for each context and each action $a \in \bigcup_{t=1}^{T} \mathcal{A}_t$, their mean reward $|\langle \phi_t(a), \theta_* \rangle| \leq G$. Thus, by the triangular inequality, with probability at least $1 - \delta$, we obtain

$$y_t \in \left[-\left(G + R\sqrt{2\log(2/\delta)} \right), \left(G + R\sqrt{2\log(2/\delta)} \right) \right]. \tag{D.3.10}$$

Next we bound the prediction of the experts that PS-OFUL considers in its prediction. To do so, we employ the same idea as we mentioned in the proof of Lemma 6.3.5 in Appendix D.2.2, where we first show an upper bound on the prediction of the any true

model i_*. In particular, we can write for any time $t \in [T]$:

$$\begin{aligned}\left|\langle \phi_t(a_t), \widehat{\theta}_t^{i_*}\rangle\right| &= \left|\langle \phi_t(a_t), \theta_*\rangle + \langle \phi_t(a_t), \widehat{\theta}_t^{i_*} - \theta_*\rangle\right| \\ &\overset{(a)}{\leq} |\langle \phi_t(a_t), \theta_*\rangle| + \left|\langle \phi_t(a_t), \widehat{\theta}_t^{i_*} - \theta_*\rangle\right| \\ &\overset{(b)}{\leq} G + \|\phi_t(a_t)\|_{(V_t^{\lambda_i})^{-1}} \left(\|\Phi_t \eta_t\|_{(V_t^{\lambda_i})^{-1}} + \sqrt{\lambda_{i_*}}\, \|\widehat{\mu}_{i_*} - \theta_*\|\right) \\ &\overset{(c)}{\leq} G + RL\sqrt{d \log\left(\frac{1 + \frac{tL^2}{\lambda_i d}}{\delta}\right)} + L\sqrt{\lambda_i}(b_i + c_i) \end{aligned} \tag{D.3.11}$$

(a) It results from triangular inequality. **(b)** This comes from the Assumption 14 as well as Theorem 1 in [42]. **(c)** This is because of the Theorem 2 in [42] and the fact that i_* is the true model and hence $\theta_* \in B(\widehat{\mu}_{i_*}, b_{i_*})$. Thus, we can have $\|\widehat{\mu}_{i_*} - \theta_*\| \leq (b_i + c_i)$. PS-OFUL employees this idea that at any time step, any potentially true model should have a similar bound on its prediction. This is being enforced by the set of admissible expert, $\mathcal{S}_t$, where it only considers experts that have the following bound on their prediction at each time $t \in [T]$ as:

$$\left|\langle \phi_t(a_t), \widehat{\theta}_t^{i}\rangle\right| \leq G + RL\sqrt{d \log\left(\frac{1 + \frac{tL^2}{\lambda_i d}}{\delta}\right)} + L\sqrt{\lambda_i}(b_i + c_i). \tag{D.3.12}$$

If at some time step t, this bound does not hold for any expert i, then the algorithm simply eliminates that expert from the set of admissible experts, since that model is not a true model (i.e., the reward does not belong to the ball of that model), and that expert will remain out for the rest of the game.

Setting $\lambda_i = \frac{1}{(b_i + c_i)^2}$ in (D.3.12), we can bound the prediction of each expert $i \in \mathcal{S}_t$ at

round $t \in [T]$ as

$$\langle \phi_t(a_t), \widehat{\theta}_t^i \rangle \in \Bigg[-\Bigg(G + L + RL\sqrt{d \log\Big(\frac{1 + \frac{tL^2 \max_{i\in[M]}(b_i+c_i)^2}{d}}{\delta}\Big)}\Bigg) , \Bigg(G + L + RL\sqrt{d \log\Big(\frac{1 + \frac{tL^2 \max_{i\in[M]}(b_i+c_i)^2}{d}}{\delta}\Big)}\Bigg)\Bigg]. \quad \text{(D.3.13)}$$

Putting together (D.3.10) and (D.3.13), we conclude that for all rounds $t \in [T]$ and experts $i \in \mathcal{S}_T$, with probability at least $1-\delta$, the rewards y_t and the experts' predictions $f_t^i(H_t)$ are in the range $[\beta, \beta + \ell]$ for

$$\begin{aligned} \beta &= -\Bigg(G + L + RL\sqrt{d \log\Big(\frac{1 + \frac{tL^2 \max_{i\in[M]}(b_i+c_i)^2}{d}}{\delta}\Big)}\Bigg), \\ \ell &= 2\Bigg(G + L + RL\sqrt{d \log\Big(\frac{1 + \frac{tL^2 \max_{i\in[M]}(b_i+c_i)^2}{d}}{\delta}\Big)}\Bigg). \end{aligned} \quad \text{(D.3.14)}$$

Using Proposition D.1.1 in Appendix D.1 with the bound on the observations and predictions in (D.3.14), with probability at least $1-\delta$, we obtain the following regret bound for `SqAlg`:

$$\mathcal{R}_{\texttt{Sq}}(t) \leq 8(\log M)\Bigg((G + L)^2 + R^2L^2 d \log\Big(\frac{1 + \frac{tL^2 \max_{i\in[M]}(b_i+c_i)^2}{d}}{\delta}\Big)\Bigg), \quad \text{(D.3.15)}$$

in which we use the fact that for $a, b > 0$, $(a+b)^2 \leq 2a^2 + 2b^2$. This concludes our proof.

D.3.3 Proof of Theorem 6.4.3

In order to fully specify the confidence set $\mathcal{C}_t$ and prove its validity, i.e., $\theta_* \in \mathbb{P}(\theta_* \in \mathcal{C}_t) \geq 1-\delta$, we should find a high probability upper-bound $\gamma_t(\delta)$ for the sum of the square

loss of the oracle predictions, i.e.,

$$\sum_{s=1}^{t} (\widehat{y}_s - \langle \phi_s(a_s), \theta_* \rangle)^2 \leq \gamma_t(\delta).$$

Let $z_s = (\widehat{y}_s - y_s)^2 - (\langle \phi_s(a_s), \widehat{\theta}_s^{i_*} \rangle - y_s)^2$, where $i_* \in \mathcal{I}_*$ is the index of a ball that contains θ_*. Since $y_s = \langle \phi_s(a_s), \theta_* \rangle + \eta_s$, we may write

$$\begin{aligned} z_s &= (\widehat{y}_s - \langle \phi_s(a_s), \theta_* \rangle - \eta_s)^2 - (\langle \phi_s(a_s), \widehat{\theta}_s^{i_*} \rangle - \langle \phi_s(a_s), \theta_* \rangle - \eta_s)^2 \\ &= (\widehat{y}_s - \langle \phi_s(a_s), \theta_* \rangle)^2 - (\langle \phi_s(a_s), \widehat{\theta}_s^{i_*} \rangle - \langle \phi_s(a_s), \theta_* \rangle)^2 + 2\eta_s(\langle \phi_s(a_s), \widehat{\theta}_s^{i_*} \rangle - \widehat{y}_s). \end{aligned}$$

Since $\sum_{s=1}^{t} z_s \leq \mathcal{R}_{\mathtt{Sq}}(t)$, where $\mathcal{R}_{\mathtt{Sq}}(t)$ is the regret of the regression oracle at round t, we have

$$\sum_{s=1}^{t} (\widehat{y}_s - \langle \phi_s(a_s), \theta_* \rangle)^2 \leq \mathcal{R}_{\mathtt{Sq}}(t) + \sum_{s=1}^{t} (\langle \phi_s(a_s), \widehat{\theta}_s^{i_*} \rangle - \langle \phi_s(a_s), \theta_* \rangle)^2 + 2\sum_{s=1}^{t} \eta_s(\langle \phi_s(a_s), \widehat{\theta}_s^{i_*} \rangle - \widehat{y}_s). \tag{D.3.16}$$

From the definition of U_t in (6.4.1), we may upper-bound $\sum_{s=1}^{t} (\langle \phi_s(a_s), \widehat{\theta}_s^{i_*} \rangle - \langle \phi_s(a_s), \theta_* \rangle)^2$ with U_t and write (D.3.16) as

$$\begin{aligned} \sum_{s=1}^{t} (\widehat{y}_s - \langle \phi_s(a_s), \theta_* \rangle)^2 &\leq \mathcal{R}_{\mathtt{Sq}}(t) + U_t + 2\sum_{s=1}^{t} \eta_s(\langle \phi_s(a_s), \widehat{\theta}_s^{i_*} \rangle - \widehat{y}_s) \\ &\leq \mathcal{R}_{\mathtt{Sq}}(t) + U_t + 2\sum_{s=1}^{t} \eta_s \langle \phi_s(a_s), \widehat{\theta}_s^{i_*} - \theta_* \rangle + 2\sum_{s=1}^{t} \eta_s(\langle \phi_s(a_s), \theta_* \rangle - \widehat{y}_s). \end{aligned} \tag{D.3.17}$$

Then, from Proposition D.4.1 in Appendix D.4, with probability at least $1-\delta$, we have

$$\left|\sum_{s=1}^{t}\eta_s\langle\phi_s(a_s),\widehat{\theta}_s^{i*}-\theta_*\rangle\right|\leq \tag{D.3.18}$$
$$R\sqrt{2\left(1+\sum_{s=1}^{t}\langle\phi_s(a_s),\widehat{\theta}_s^{i*}-\theta_*\rangle^2\right)\log\left(\frac{\sqrt{1+\sum_{s=1}^{t}\langle\phi_s(a_k),\widehat{\theta}_s^{i*}-\theta_*\rangle^2}}{\delta}\right)},$$

and

$$\left|\sum_{s=1}^{t}\eta_s(\langle\phi_s(a_s),\theta_*\rangle-\widehat{y}_s)\right|\leq \tag{D.3.19}$$
$$R\sqrt{2\left(1+\sum_{s=1}^{t}(\langle\phi_s(a_s),\theta_*\rangle-\widehat{y}_s)^2\right)\log\left(\frac{\sqrt{1+\sum_{s=1}^{t}(\langle\phi_s(a_s),\theta_*\rangle-\widehat{y}_s)^2}}{\delta}\right)}.$$

Using (D.3.18) and (D.3.19), we may write (D.3.17) as

$$\sum_{s=1}^{t}\left(\widehat{y}_s-\langle\phi_s(a_s),\theta_*\rangle\right)^2\leq\mathcal{R}_{\mathtt{Sq}}(t)+U_t+2R\sqrt{2(1+U_t)\log\left(\sqrt{1+U_t}/\delta\right)}$$
$$+R\sqrt{8\left(1+\sum_{s=1}^{t}\left(\widehat{y}_s-\langle\phi^s(a_s),\theta_*\rangle\right)^2\right)\log\left(\frac{\sqrt{1+\sum_{s=1}^{t}\left(\widehat{y}_s-\langle\phi_s(a_s),\theta_*\rangle\right)^2}}{\delta}\right)}. \tag{D.3.20}$$

It is easy to see that (D.3.20) can be written in the form $u^2\leq v+uw$, where $u=\sqrt{1+\sum_{s=1}^{t}(\langle\phi_s(a_s),\theta_*\rangle-\widehat{y}_s)^2}$, $v=1+\mathcal{R}_{\mathtt{Sq}}(t)+U_t+2R\sqrt{2(1+U_t)\log(\frac{\sqrt{1+U_t}}{\delta})}$, and $w=R\sqrt{8\log(u/\delta)}$. Then, by applying Lemma D.4.5 in Appendix D.4, we may write $u\leq w+\sqrt{v}$. Substituting for w, we can get $u\leq\sqrt{v}+R\sqrt{8\log(u/\delta)}$. Then, by

Lemma D.4.6 in Appendix D.4, for $\delta \in (0, 1/4]$, we have

$$u \leq \sqrt{v} + 4R\sqrt{\log\left(\frac{R\sqrt{8}+\sqrt{v}}{\delta}\right)},$$

which using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, for any a and b, we can write it as

$$u^2 \leq 2v^2 + 32R^2 \log\left(\frac{R\sqrt{8}+\sqrt{v}}{\delta}\right).$$

Finally, we substitute u and v, and subtract 1 from both sides, and for $\delta \in (0, 1/4]$, we obtain

$$\sum_{s=1}^{t}\left(\widehat{y}_s - \langle \phi_s(a_s), \theta_* \rangle\right)^2 \leq 1 + 2\mathcal{R}_{\mathtt{Sq}}(t) + 2U_t + 4R\sqrt{2(1+U_t)\log\left(\sqrt{1+U_t}/\delta\right)}$$
$$+ 32R^2 \log\left(\frac{R\sqrt{8} + \sqrt{1 + \mathcal{R}_{\mathtt{Sq}}(t) + U_t + 2R\sqrt{2(1+U_t)\log\left(\sqrt{1+U_t}/\delta\right)}}}{\delta}\right) := \gamma_t(\delta). \tag{D.3.21}$$

Eq. D.3.21 shows that for $\delta \in (0, 1/4]$, with probability at least $1-\delta$, we have $\theta^* \in C_t$, which completes the proof of the validity of the confidence set $\mathcal{C}_t$.

We can now fully specify $\mathcal{C}_t$ by plugging U_t from (D.3.9) (see Appendix D.3.1) and

$\mathcal{R}_{\texttt{Sq}}(t)$ from (D.3.15) (see Appendix D.3.2) into (D.3.21), and write $\gamma_t(\delta)$ as

$$\gamma_t(\delta) := 1 + 2\mathcal{R}_{\texttt{Sq}}(t) + 2U_t + 4R\sqrt{2(1+U_t)\log\left(\sqrt{1+U_t}/\delta\right)} + 32R^2\log\left(\frac{R\sqrt{8} + \sqrt{1+\mathcal{R}_{\texttt{Sq}}(t) + U_t + 2R\sqrt{2(1+U_t)\log\left(\sqrt{1+U_t}/\delta\right)}}}{\delta}\right),$$

where (D.3.22)

$$U_t = 3 + 8d\log\left(1 + \frac{tL^2\max_{i\in[M]}(b_i+c_i)^2}{d}\right) + 32R^2\log\left(\frac{2\sqrt{2}R + \sqrt{2 + 4d\log\left(1 + \frac{tL^2\max_{i\in[M]}(b_i+c_i)^2}{d}\right)}}{\delta}\right),$$

$$\mathcal{R}_{\texttt{Sq}}(t) = 8\log(M)\left(G^2 + L^2 + 2GL + R^2L^2 d\log\left(\frac{1 + \frac{tL^2\max_{i\in[M]}(b_i+c_i)^2}{d}}{\delta}\right)\right),$$

which concludes the proof.

A closer look at U_t and $\mathcal{R}_{\texttt{Sq}}(t)$, the two main terms in the definition of $\gamma_t(\delta)$, we may write them in terms of the dominant terms as

$$\begin{aligned} U_t &\approx \overbrace{3 + 16R^2\log(2)}^{C_1} + 8d\log\left(1 + \frac{tL^2\max_{i\in[M]}(b_i+c_i)^2}{d}\right) + 32R^2\log(1/\delta) \\ &\quad + 32R^2\log\left(1 + 2R + d\log\left(1 + \frac{tL^2\max_{i\in[M]}(b_i+c_i)^2}{d}\right)\right) \\ &\approx C_1 + 32R^2\log(1/\delta) + 8d\log\left(1 + \frac{tL^2\max_{i\in[M]}(b_i+c_i)^2}{d}\right), \end{aligned} \tag{D.3.23}$$

and

$$\begin{aligned}\mathcal{R}_{\mathtt{Sq}}(t) &= \overbrace{8(G+L)^2\log(M)}^{C_2} + 8R^2L^2d\log(M)\log(1/\delta) \\ &\quad + 8R^2L^2d\log(M)\log\left(1+\frac{tL^2\max_{i\in[M]}(b_i+c_i)^2}{d}\right) \\ &= C_2 + 8R^2L^2d\log(M)\log(1/\delta) + 8R^2L^2d\log(M)\log\left(1+\frac{tL^2\max_{i\in[M]}(b_i+c_i)^2}{d}\right).\end{aligned} \tag{D.3.24}$$

Using (D.3.23) and (D.3.24), we may write $\gamma_t(\delta)$ in terms of the dominant terms as

$$\begin{aligned}\gamma_t(\delta) \approx 1 + 2C_1 + 2C_2 &+ 16R^2\left(4 + L^2d\log(M)\right)\log(1/\delta) \\ &+ 16d\left(1 + R^2L^2\log(M)\right)\log\left(1+\frac{tL^2\max_{i\in[M]}(b_i+c_i)^2}{d}\right).\end{aligned} \tag{D.3.25}$$

D.3.4 Proof of Lemma 6.4.6

In Theorem 6.4.3, we proved that at each round, with probability at least $1-\delta$, the true reward parameter θ_* belongs to the confidence set $\mathcal{C}_t$ of the PS-OFUL algorithm. Here, we show how the regret of PS-OFUL is related to the radius $\gamma_t(\delta)$ of this confidence set.

Here we assume that at the first d rounds, the algorithm plays actions whose features are of the form $\phi_i(a_i) = Le_i,\ \forall i \in [d]$, where $e_i = [0,\dots,1,\dots,0]$ is a d-dimensional vector whose elements are all 0, except a 1 at the i^{th} position. In this case, we can define a matrix V_t as

$$V_t = \sum_{s=1}^{t-1}\phi_s(a_s)^\top\phi_s(a_s) = L^2I + \sum_{s=d+1}^{t-1}\phi_t(a_t)^\top\phi_t(a_t), \tag{D.3.26}$$

and use it to rewrite the confidence set as

$$\mathcal{C}_{t-1} = \big\{\theta \in \mathbb{R}^d : (\theta - \widehat{\theta}_t)V_t(\theta - \widehat{\theta}_t) + \sum_{s=1}^{t-1} \big(\widehat{y}_s - \langle \phi_s(a_s), \widehat{\theta}_t\rangle\big)^2 \leq \gamma_t(\delta)\big\}, \tag{D.3.27}$$

where $\widehat{\theta}_t = \operatorname{argmin}_{\theta\in\mathbb{R}^d} \sum_{s=1}^{t-1} \big(\widehat{y}_s - \langle \phi_s(a_s), \theta\rangle\big)^2$. The confidence set $\mathcal{C}_t$ in (D.3.27) is contained in a larger ellipsoid

$$\mathcal{C}_{t-1} \subseteq \big\{\theta \in \mathbb{R}^d : (\theta - \widehat{\theta}_t)V_t(\theta - \widehat{\theta}_t) \leq \gamma_t(\delta)\big\} = \big\{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_t\|_{V_t}^2 \leq \gamma_t(\delta)\big\}. \tag{D.3.28}$$

Given $(a_t, \widetilde{\theta}_t) = \arg\max_{a\in\mathcal{A}_t} \max_{\theta\in\mathcal{C}_{t-1}} \langle \phi_t(a), \theta\rangle$ are the action and parameter resulted from solving the optimization problem at round t of the PS-OFUL algorithm, we may write

$$\begin{aligned}
\langle \phi_t(a_t^*), \theta_*\rangle - \langle \phi_t(a_t), \theta_*\rangle &\leq \langle \phi_t(a_t), \widetilde{\theta}_t\rangle - \langle \phi_t(a_t), \theta_*\rangle \\
&= \langle \phi_t(a_t), \widetilde{\theta}_t - \widehat{\theta}_t\rangle + \langle \phi_t(a_t), \widehat{\theta}_t - \theta_*\rangle \\
&\leq \|\phi_t(a_t)\|_{V_t^{-1}} \|\widetilde{\theta}_t - \widehat{\theta}_t\|_{V_t} + \|\phi_t(a_t)\|_{V_t^{-1}} \|\widehat{\theta}_t - \theta_*\|_{V_t} \\
&\leq 2\sqrt{\gamma_t(\delta)}\, \|\phi_t(a_t)\|_{V_t^{-1}} \qquad \text{(because } \theta_*, \widetilde{\theta}_t \in \mathcal{C}_{t-1}\text{)}.
\end{aligned} \tag{D.3.29}$$

Since $\forall a \in \bigcup_{t=1}^T \mathcal{A}_t$, we assume that $|\langle \phi(a), \theta_*\rangle| \leq G$, we can upper-bound the instantaneous regret in (D.3.29) as

$$\langle \phi_t(a_t^*), \theta_*\rangle - \langle \phi_t(a_t), \theta_*\rangle \leq 2\min\big\{G, \sqrt{\gamma_t(\delta)}\, \|\phi_t(a_t)\|_{V_t^{-1}}\big\}. \tag{D.3.30}$$

Using (D.3.30), we can bound the transfer-regret of PS-OFUL as

$$\begin{aligned}
\mathcal{R}_{\text{PS-OFUL}}(T) &= \sum_{t=1}^{T}\langle \phi_t(a_t^*) - \phi_t(a_t), \theta_* \rangle \leq 2Gd + \sum_{t=d+1}^{T} \langle \phi_t(a_t^*) - \phi_t(a_t), \theta_* \rangle \\
&\leq 2Gd + 2\sum_{t=d+1}^{T} \min\{G, \sqrt{\gamma_t(\delta)}\, \|\phi_t(a_t)\|_{V_t^{-1}}\} \\
&\leq 2Gd + 2\sum_{t=d+1}^{T} \sqrt{\gamma_t(\delta)}\ \min\{G, \|\phi_t(a_t)\|_{V_t^{-1}}\} \qquad (\text{since } \gamma_t(\delta) \geq 1) \\
&\leq 2Gd + 2\left(\max_{d<t\leq T} \sqrt{\gamma_t(\delta)}\right) \sum_{t=d+1}^{T} \min\{G, \|\phi_t(a_t)\|_{V_t^{-1}}\} \\
&\leq 2Gd + 2\left(\max_{d<t\leq T} \sqrt{\gamma_t(\delta)}\right) (\max\{1, G\}) \sum_{t=d+1}^{T} \min\{1, \|\phi_t(a_t)\|_{V_t^{-1}}\} \\
&\leq 2Gd + 2\left(\max_{d<t\leq T} \sqrt{\gamma_t(\delta)}\right) (\max\{1, G\}) \sqrt{T\sum_{t=d+1}^{T} \min\{1, \|\phi_t(a_t)\|^2_{V_t^{-1}}\}} \\
&\leq 2Gd + 2\left(\max_{d<t\leq T} \sqrt{\gamma_t(\delta)}\right) (\max\{1, G\}) \sqrt{2T\log\left(\frac{\det(V_T)}{\det(V_d)}\right)}, \qquad \text{(D.3.31)}
\end{aligned}$$

where the last inequality follows from Lemma D.4.2 in Appendix D.4. Then, using Lemma D.4.3 in Appendix D.4, we can bound $\det(V_T) \leq \left(L^2 + \frac{TL^2}{d}\right)^d$ and $\det(V_d) = L^{2d}$. Hence, we may write (D.3.31) as

$$\mathcal{R}_{\text{PS-OFUL}}(T) \leq 2Gd + 2\max\{1, G\}\sqrt{2dT\log\left(1 + \frac{T}{d}\right)\max_{d<t\leq T}\gamma_t(\delta)}. \qquad \text{(D.3.32)}$$

D.3.5 Proof of Theorem 6.4.2

If we substitute $\gamma_t(\delta)$ from (D.3.25) in the regret bound (D.3.32), we may write it (in terms of the dominant terms) as

$$\mathcal{R}_{\text{PS-OFUL}}(T) \leq 2Gd + 2\sqrt{2}\max\{1,G\}\sqrt{dT\log\left(1+\frac{T}{d}\right)}$$
$$\times\sqrt{C_3 + 16R^2\big(4+L^2 d\log(M)\big)\log(1/\delta) + 16d\big(1+R^2L^2\log(M)\big)\log\left(1+\frac{TL^2\max_{i\in[M]}(b_i+c_i)^2}{d}\right)}$$
$$= \mathcal{O}\left(dRL\max\{1,G\}\sqrt{1+\log(M)}\times\sqrt{T\log\left(1+\frac{T}{d}\right)\log\left(\frac{1+\frac{TL^2\max_{i\in[M]}(b_i+c_i)^2}{d}}{\delta}\right)}\right), \tag{D.3.33}$$

where $C_3 = 1+2C_1+2C_2$, and hence $C_3 = 7+32R^2\log(2)+16(G+L)^2\log(M)$.

D.4 Auxiliary Tools

Here we report auxiliary results that we use in our proofs in other appendices.

We start with stating Theorem 7 in [133], which is the self-normalized martingale tail inequality for the scalar random variables.

Proposition D.4.1 (Self-normalized bound for martingales) *Let $\{F_t\}_{t=1}^{\infty}$ be a filtration. Let τ be a stopping time w.r.t to the filtration $\{F_t\}_{t=1}^{\infty}$, i.e., the event $\{\tau \leq t\}$ belongs to F_{t+1}. Let $\{Z_t\}_{t=1}^{\infty}$ be a sequence of real-valued variables such that Z_t is F_t-measurable. Let $\{\eta_t\}_{t=1}^{\infty}$ be a sequence of real-valued random variables such that η_t is F_{t+1} measurable and is conditionally R-sub-Gaussian. Then, for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\left\| \sum_{t=1}^{\tau} \eta_t Z_t \right\| \leq R \sqrt{2 \left(1 + \sum_{t=1}^{\tau} Z_t^2\right) \log \left(\frac{\sqrt{1 + \sum_{t=1}^{\tau} Z_t^2}}{\delta} \right)}.$$

Next, we state a direct application of Lemma 11 in [42] that bounds the cumulative sum of $\sum_{s=1}^{t-1} \|\phi_s(a_s)\|^2_{V_s^{-1}}$ which plays an important role in most of the proofs for linear bandits problems.

Lemma D.4.2 *Let $\lambda > 0$ and $V_t = \lambda I + \sum_{s=1}^{t-1} \phi_s(a_s)\phi_s^{\top}(a_s)$. If for all $a \in \cup_{s=1}^{t-1} \mathcal{A}_s$, we have $\|\phi_s(a)\|_2 \leq L$, then we may write*

$$\sum_{s=1}^{t-1} \min\{1, \|\phi_s(a_s)\|^2_{V_s^{-1}}\} \leq 2 \log \left(\frac{\det(V_t)}{\det(\lambda I)} \right).$$

Next, we present a determinant-trace inequality matrix result.

Lemma D.4.3 (Determinant-Trace Inequality) *Suppose $X_1, \ldots, X_{t-1} \in \mathbb{R}^d$, and for any $1 \leq s \leq t-1$, we have $\|X_s\|_2 \leq L$. Let $V_t = \lambda I + \sum_{s=1}^{t-1} X_s X_s^{\top}$, for some*

$\lambda > 0$. *Then we have*

$$\det(V_t) \leq \left(\lambda + \frac{tL^2}{d}\right)^d.$$

Proof: Let $\alpha_1, \ldots, \alpha_d$ be the eigenvalues of V_t. Since V_t is positive definite, its eigenvalues are positive. Also not that $\det(V_t) = \Pi_{s=1}^d \alpha_s$ and $trace(V_t) = \sum_{s=1}^d \alpha_s$. By arithmetic-geometric means inequality we have

$$\sqrt[d]{\alpha_1 \ldots \alpha_d} \leq \frac{\alpha_1 + \cdots + \alpha_d}{d}.$$

Therefore, $\det(V_t) \leq \left(\frac{trace(V_t)}{d}\right)^d$. It suffices to upper-bound the trace of V_t as

$$trace(V_t) = trace(\lambda I) + \sum_{s=1}^{t-1} trace(X_s X_s^\top) = d\lambda + \sum_{s=1}^{t-1} \|X_s\|_2^2 \leq d\lambda + tL^2,$$

and the result follows. ■

Next, we state a bound on the absolute value of the R-sub-Gaussian random variable.

Lemma D.4.4 *Let $\{F_t\}_{t=1}^\infty$ be a filtration. Let $\{\eta\}_{t=1}^\infty$ be a real-valued stochastic process such that η_t is F_t-measurable and η_t is conditionally R-sub-Gaussian for some $R > 0$, i.e.,*

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}\left[\eta_t | F_t\right] = 0, \quad \mathbb{E}\left[e^{\lambda \eta_t} | F_t\right] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right).$$

Then, condition on filtration F_t, with probability at least $1-\delta$, we have $|\eta_t| \leq R\sqrt{2\log(2/\delta)}$.

Proof: Let $\lambda > 0$. Then,

$$\begin{aligned}\mathbb{P}(\eta_t \geq k|F_t) = \mathbb{P}(e^{\lambda\eta_t} \geq e^{\lambda k}|F_t) &\leq e^{-\lambda k}\ \mathbb{E}[e^{\lambda\eta_t}|F_t] \qquad \text{(by Markov's inequality)} \\ &\leq e^{-\lambda k}e^{\frac{\lambda^2R^2}{2}} = \exp\left(-\lambda k + \frac{\lambda^2R^2}{2}\right). \qquad \text{(D.4.1)}\end{aligned}$$

Optimizing for λ, and thus, selecting $\lambda = \frac{k}{R^2}$, we conclude that

$$\mathbb{P}(\eta_t \geq k|F_t) \leq e^{-\frac{k^2}{2R^2}}.$$

Repeating this argument for $-\eta_t$, we also obtain $\mathbb{P}(\eta_t \leq -k|F_t) \leq e^{-\frac{k^2}{2R^2}}$. Combining these two bounds, we can conclude that

$$\mathbb{P}(|\eta_t| \geq k|F_t) \leq 2e^{-\frac{k^2}{2R^2}}. \qquad \text{(D.4.2)}$$

From (D.4.2), with the choice of $\delta = 2e^{-\frac{k^2}{2R^2}}$, and thus $k = R\sqrt{2\log(2/\delta)}$, completes the proof. ■

Then, we state a square-root trick for positive numbers.

Lemma D.4.5 *Let $a, b > 0$. If $z^2 \leq a + bz$, then $z \leq \sqrt{a} + b$.*

Proof: Let $q(z) = z^2 - bz - a$. We can rewrite the condition $z^2 \leq a + bz$ as $q(z) \leq 0$. Then we know that the quadratic polynomial $q(z)$ has the following two roots

$$z_1^* = \frac{b + \sqrt{b^2 + 4a}}{2}\ z_2^* = \frac{b - \sqrt{b^2 + 4a}}{2}.$$

Then, we know that the condition $q(z) \leq 0$, implies that $\min\{z_1^*, z_2^*\} \leq z \leq \max\{z_1^*, z_2^*\}$.

Therefore, for positive numbers a, b, we get

$$z \leq \max\{z_1^*, z_2^*\} = \frac{b + \sqrt{b^2 + 4a}}{2} \leq b + \sqrt{a},$$

where for the last inequality, we use the fact that for $u, v > 0$, $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$. ■

Next, we restate a simple logarithmic trick from [133].

Lemma D.4.6 (Proposition 10 in [133]) *Let $c \geq 1$, $q > 0$, $\delta \in (0, 1/4]$. If $s \geq 1$ and $s \leq c + q\sqrt{\log(s/\delta)}$, then we have $s \leq c + q\sqrt{2\log(\frac{c+q}{\delta})}$.*

D.5 Relation to Latent Bandits

In this section, we informally show that if the goal in latent bandits is to have a better scaling with the number of actions K (e.g., the number of actions K is much larger than the number of latent states M), we can use a different bandit model selection strategy, called *regret balancing* [147, 148, 157, 158] to obtain an improved regret that scales as $\min\{\varepsilon T + \sqrt{MT}, \sqrt{KMT}\}$. This rate is the best of the regret of PS-OFUL, which scales as $\sqrt{KT}$, and the regret of the latent bandit algorithm of [146], which scales as $\varepsilon T + \sqrt{MT}$.

In regret balancing, in each round, the model selection strategy chooses one of M base algorithms. We denote by $N_{i,t}$, the number of times that the base algorithm i has been selected up to round t, and by $R_{i,t}$, the cumulative rewards of this base algorithm during these $N_{i,t}$ rounds. Given a reference regret bound $U : [T] \to \mathbb{R}$, in each round $t \in [T]$, the algorithm first finds the optimistic base algorithm I_t and its value b_t, i.e.,

$$I_t = \operatorname*{argmax}_{i \in [M]} \frac{R_{i,t}}{N_{i,t}} + \frac{U(N_{i,t})}{N_{i,t}} \quad , \quad b_t = \frac{R_{I_t,t}}{N_{I_t,t}} + \frac{U(N_{I_t,t})}{N_{I_t,t}} \, , \tag{D.5.1}$$

and then takes the action recommended by I_t and uses its observed reward to update the base algorithm I_t.

We can apply regret balancing to the problem of latent bandits in the following way. We consider $M + 1$ base algorithms: one that plays UCB, and M, each corresponds to a latent value and always plays the greedy action of that latent model (which is guaranteed to be ε-accurate by assumption). If the regret balancing strategy selects the UCB base algorithm in all rounds, it would suffer the regret $\sqrt{Kt} + \sqrt{t}$, and if it selects the optimal base algorithm, i.e., the base algorithm corresponding to the correct latent model, it would suffer the regret $\varepsilon t + \sqrt{t}$. Note that by regret, we mean the actual

regret and not pseudo-regret, and thus, $\sqrt{t}$ is the consequence of noise in the reward signal. Thus, we select the reference regret bound of our regret balancing strategy as $U(t) = \min\{\varepsilon t + \sqrt{t}, \sqrt{Kt} + \sqrt{t}\}$. We may write the regret of the resulting regret balancing strategy as follows:

$$\begin{aligned}
\mathcal{R}(T) &\overset{(a)}{=} \sum_{i=1}^{M+1} N_{i,T}\mu_* - R_{i,T} \overset{(b)}{\leq} \sum_{i=1}^{M+1} N_{i,T} b_T - R_{i,T} \overset{(c)}{=} \sum_{i=1}^{M+1} U(N_{i,T}) \\
&\leq \sum_{i=1}^{M+1} \min\left\{\varepsilon N_{i,T} + \sqrt{N_{i,T}}, \sqrt{KN_{i,T}} + \sqrt{N_{i,T}}\right\} \\
&\leq \min\left\{ \sum_{i=1}^{M+1} \left(\varepsilon N_{i,T} + \sqrt{N_{i,T}}\right), \sum_{i=1}^{M+1} \left(\sqrt{KN_{i,T}} + \sqrt{N_{i,T}}\right)\right\} \\
&\overset{(d)}{=} \min\left\{\varepsilon T + \sum_{i=1}^{M+1} \sqrt{N_{i,T}}, \sum_{i=1}^{M+1} \left(\sqrt{KN_{i,T}} + \sqrt{N_{i,T}}\right)\right\} \\
&\overset{(e)}{\leq} \min\left\{\varepsilon T + \sum_{i=1}^{M+1} \sqrt{\frac{T}{M+1}}, \sum_{i=1}^{M+1} \left(\sqrt{K\frac{T}{M+1}} + \sqrt{\frac{T}{M+1}}\right)\right\} \\
&= \min\left\{\varepsilon T + \sqrt{(M+1)T}, \sqrt{K(M+1)T} + \sqrt{(M+1)T}\right\} \\
&= \mathcal{O}\left(\min\left\{\varepsilon T + \sqrt{MT}, \sqrt{KMT}\right\}\right),
\end{aligned}$$

which concludes our claim. Note that we used the following steps in our above derivations: **(a)** μ_* is the mean of the optimal arm. **(b)** This is because with high probability we have $\mu_* \leq b_t$, $\forall t \in [T]$. **(c)** This is from the definition b_t in (D.5.1). **(d)** This is due to the fact that $\sum_{i=1}^{M+1} N_{i,T} = T$. **(e)** The maximizer of $\sum_{i=1}^{M+1} \sqrt{N_{i,T}}$, subject to $\sum_{i=1}^{M+1} N_{i,T} = T$, is when all $\{N_{i,T}\}_{i=1}^{M+1}$ are equal.

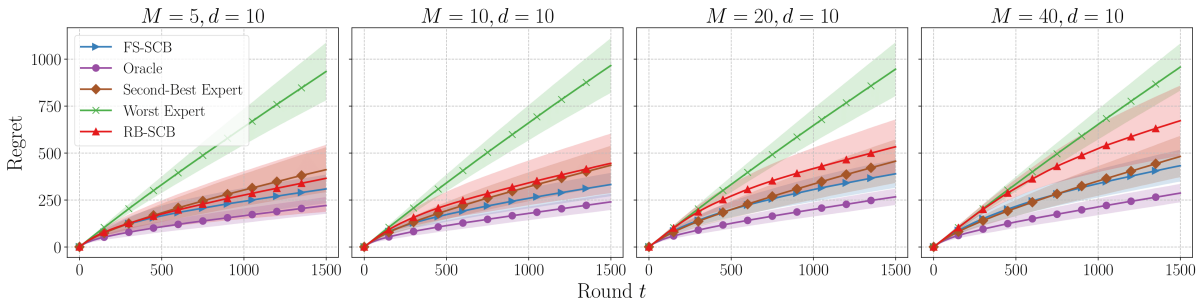

Figure D.1: Feature selection on MNIST dataset. The regrets are averaged over 100 LB problems.

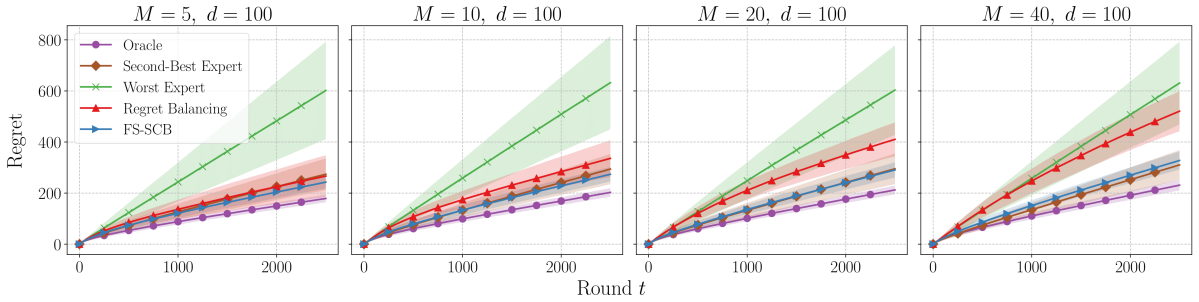

Figure D.2: Feature selection on CIFAR-100 dataset. The regrets are averaged over 100 LB problems.

D.6 More on Experimental Results

We evaluate the performances of FS-SCB and PS-OFUL algorithms in a synthetic linear bandit problem and real-world image classification problems on CIFAR-10, CIFAR-100 [151], and MNIST datasets [150].

D.6.1 Feature Selection

MNIST Dataset

MNIST dataset consists of 60000 training and 10000 test images of size 28×28, each belonging to one of 10 classes. We train a convolutional neural network (CNN) with M different number of epochs on MNIST data, and use their second layer to the last as our $d = 10$-dimensional feature maps $\{\phi^i\}_{i=1}^M$. These feature maps have test accuracy

between 20% (worst model) and 97% (best model). We set the best one as true model ϕ^{i_*}. For each class $s \in \mathcal{S} = \{0, \ldots, 9\}$, we fit a linear model, given the feature map ϕ^{i_*}, and obtain parameters $\{\theta_s^{i_*}\}_{s=0}^9$. At the beginning of each LB task, we select a class $s_* \in \mathcal{S}$ uniformly at random and set its parameter to $\theta_{s_*}^{i_*}$. At each round $t \in [T]$, the learner is given an action set consists of 10 images, one from class s_* and the rest randomly selected from the other classes. The reward of each action a is defined as $\langle \phi^{i_*}(a), \theta_{s_*}^{i_*} \rangle + \eta_t \in [0, 1]$, where $\phi^{i_*}(a)$ is the application of the feature map ϕ^{i_*} to the image corresponding to action a and $\eta_t \sim \mathcal{U}[-0.5, 0.5]$ is the noise.

In Figure D.1, we compare the regret of our FS-SCB algorithm for different number of models M with a regret balancing algorithm that uses SquareCB baselines (RB-SCB), and three SquareCB algorithms that use the best (Oracle), second-best (with test accuracy 84%), and worst feature maps (experts). Each plot is averaged over 100 LB problems. Figure D.1 shows that **1)** FS-SCB always performs between the best and second-best experts, **2)** the regret of FS-SCB that scales as $\sqrt{\log M}$ is close to RB-SCB (scales as $\sqrt{M}$) for small M, but gets much better as M grows, and **3)** RB-SCB has much higher variance than the other algorithms.

CIFAR-100 Dataset

CIFAR-100 dataset consists of 50000 training and 10000 test images of size 32×32, each belonging to one of 100 classes. We extracted the features of the images by fine tuning and taking the output of the second-to-last layer of the EfficientNet-B0 Network [159] and got the feature matrix of dimension 50000×1280. For all experts $i \in [M]$, we multiply this feature matrix with a Gaussian random matrix of dimension $1280 \times d_i$ for $d_i \in [2, 128]$ to get the d_i dimensional feature maps ϕ^i. These feature maps have accuracy between 5% (worst model) and 78% (best model). We set the best one as true model ϕ^{i_*}. For each class $s \in \mathcal{S} = \{0, \ldots, 99\}$, we fit a linear model, given the feature map ϕ^{i_*} and

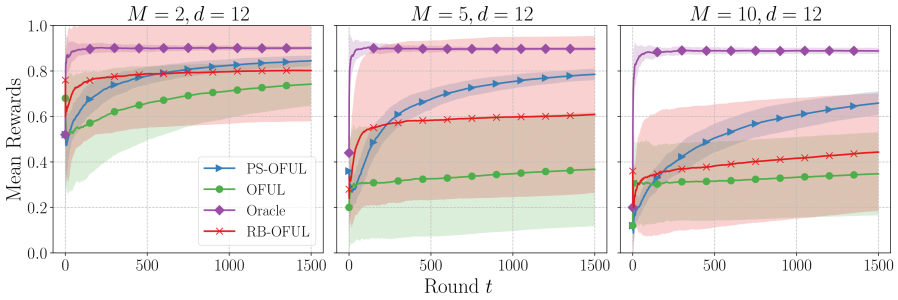

Figure D.3: Parameter selection on MNIST dataset, where $100M$ datasets of size 500 are used to define the balls. The results are averaged over 50 runs.

obtain parameters $\{\theta_s^{i_*}\}_{s=0}^{99}$. At the beginning of each LB task, we select a class $s_* \in \mathcal{S}$ uniformly at random and set its parameter to $\theta_{s_*}^{i_*}$. At each round $t \in [T]$, the learner is given an action set consists of 10 images, one from class s_* and the rest randomly selected from the other classes. The reward of each action a is defined as $\langle \phi^{i_*}(a), \theta_{s_*}^{i_*} \rangle + \eta_t \in [0, 1]$, where $\phi^{i_*}(a)$ is the application of the feature map ϕ^{i_*} to the image corresponding to action a and $\eta_t \sim \mathcal{U}[-0.5, 0.5]$ is the noise.

In Figure D.2, we compare the regret of our FS-SCB algorithm for different number of models M with a regret balancing algorithm that uses SquareCB baselines (RB-SCB) and aggregate them according to (D.5.1), and three SquareCB algorithms that use the best (Oracle), second-best (with test accuracy 55%), and worst feature maps (experts). Each plot is averaged over 100 LB problems. Figure D.2 shows that **1)** FS-SCB always performs close to the best and second-best experts, **2)** the regret of FS-SCB that scales as $\sqrt{\log M}$ is close to RB-SCB (scales as $\sqrt{M}$) for small M, but gets much better as M grows, and **3)** RB-SCB has much higher variance than the other algorithms.

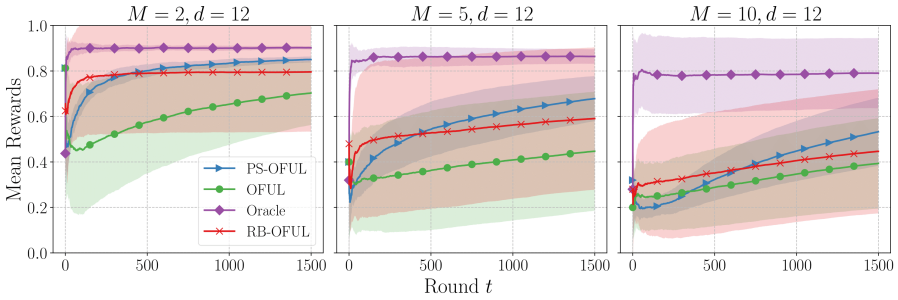

Figure D.4: Parameter selection on MNIST dataset, where $10M$ datasets of size 50 are used to define the balls. The results are averaged over 50 runs.

D.6.2 Parameter Selection

Image Classification on MNIST Dataset

MNIST dataset consists of 60000 test and 10000 training images of size 28×28, each belonging to one of 10 classes. We train a CNN with $d = 12$ neurons on second-to-last layer on MNIST dataset with 98% accuracy. We then select this d-dimensional layer as our feature map ϕ. To define our M models (balls), we sample $100M$ datasets of size 500. For each dataset, we randomly select a class $s_* \in [M]$, assign reward 1 to images from s_* and 0 to other images, and fit a linear model to it to obtain a parameter vector. Finally, we fit a Gaussian mixture model (GMM) with M components to these $100M$ parameter vectors and use the means and covariances of the resulting clusters as the center and radii of our M models (balls). At the beginning of each LB task, we select a class $s_* \in [M]$ uniformly at random. At each round $t \in [T]$, the learner is given an action set consists of 10 images, one from class s_* and the rest randomly selected from the other classes. The learner receives a reward from Ber(0.9) if it selects the image from class s_*, and from Ber(0.1), otherwise.

In Figure D.3, we compare the mean reward of PS-OFUL for different number of models M with a regret balancing algorithm that uses OFUL baselines (RB-OFUL) [147], OFUL (individual learning), and BIAS-OFUL [143] with bias being the center of the

true model (Oracle). Figure D.3 shows **1)** the good performance of PS-OFUL, **2)** the performance of PS-OFUL gets better than RB-OFUL as M grows ($\sqrt{\log M}$ vs. $\sqrt{M}$ scaling), **3)** the large variance of RB-OFUL, especially in comparison to PS-OFUL, and finally **4)** the advantage of transfer (PS-OFUL) over individual (OFUL) learning.

Impact of the model estimates: In order to demonstrate the impact of the accuracy of the model center estimates as well as the radii of the balls, we defined a less accurate set of M models (balls) using $10M$ datasets of size 50 (as opposed to $100M$ datasets of size 500). In Figure D.4, we compare the mean reward of PS-OFUL for different number of models M with RB-OFUL, OFUL, and BIAS-OFUL.

Bibliography

[1] "Expect More EV Charging Stations as States Tap into Federal Dollars, [Online]." https://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/2022/10/11/expect-more-ev-charging-stations-as-states-tap-into-federal-dollars. Accessed: October 2022.

[2] "Long Queue at Tesla Supercharge Point , [Online]." https://www.reuters.com/article/factcheck-tesla-supercharge/fact-check-video-of-queue-at-tesla-supercharge/-point-was-captured-on-thanksgiving-in-2019-idUSL1N2YM10J. Accessed: July 2022.

[3] A. Moradipari, S. Bae, M. Alizadeh, E. M. Pari, and D. Isele, *Predicting parameters for modeling traffic participants*, in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 703–708, IEEE, 2022.

[4] M. Khodayar, L. Wu, and Z. Li, *Electric vehicle mobility in transmission-constrained hourly power generation scheduling*, *Smart Grid, IEEE Transactions on* **4** (June, 2013) 779–788.

[5] X. Xi and R. Sioshansi, *Using price-based signals to control plug-in electric vehicle fleet charging*, *IEEE Transactions on Smart Grid* **5** (May, 2014) 1451–1464.

[6] R. Sioshansi, *Or forum-modeling the impacts of electricity tariffs on plug-in hybrid electric vehicle charging, costs, and emissions*, *Operations Research* **60** (2012), no. 3 506–516.

[7] M. Alizadeh, H. T. Wai, A. Goldsmith, and A. Scaglione, *Retail and wholesale electricity pricing considering electric vehicle mobility*, *IEEE Transactions on Control of Network Systems* **PP** (2018), no. 99 1–1.

[8] W. Wei, L. Wu, J. Wang, and S. Mei, *Network equilibrium of coupled transportation and power distribution systems*, *IEEE Transactions on Smart Grid* **9** (Nov, 2018) 6764–6779.

[9] J. Hu, S. You, M. Lind, and J. Østergaard, *Coordinated charging of electric vehicles for congestion prevention in the distribution grid*, *IEEE Transactions on Smart Grid* **5** (March, 2014) 703–711.

[10] D. Wang, H. Wang, J. Wu, X. Guan, P. Li, and L. Fu, *Optimal aggregated charging analysis for pevs based on driving pattern model*, in *2013 IEEE Power Energy Society General Meeting*, pp. 1–5, July, 2013.

[11] D. Tang and P. Wang, *Nodal impact assessment and alleviation of moving electric vehicle loads: From traffic flow to power flow*, *IEEE Transactions on Power Systems* **31** (Nov, 2016) 4231–4242.

[12] T. Chen, B. Zhang, H. Pourbabak, A. Kavousi-Fard, and W. Su, *Optimal routing and charging of an electric vehicle fleet for high-efficiency dynamic transit systems*, *IEEE Transactions on Smart Grid* **9** (July, 2018) 3563–3572.

[13] D. Goeke and M. Schneider, *Routing a mixed fleet of electric and conventional vehicles*, *European Journal of Operational Research* **245** (2015), no. 1 81 – 99.

[14] T. Wang, C. G. Cassandras, and S. Pourazarm, *Optimal motion control for energy-aware electric vehicles*, *Control Engineering Practice* **38** (2015) 37 – 45.

[15] Q. Guo, S. Xin, H. Sun, Z. Li, and B. Zhang, *Rapid-charging navigation of electric vehicles based on real-time power systems and traffic data*, *IEEE Transactions on Smart Grid* **5** (July, 2014) 1969–1979.

[16] S. Pourazarm, C. G. Cassandras, and T. Wang, *Optimal routing and charging of energy-limited vehicles in traffic networks*, *International Journal of Robust and Nonlinear Control* **26** (2016), no. 6 1325–1350.

[17] H. Yang, S. Yang, Y. Xu, E. Cao, M. Lai, and Z. Dong, *Electric vehicle route optimization considering time-of-use electricity price by learnable partheno-genetic algorithm*, *IEEE Transactions on Smart Grid* **6** (March, 2015) 657–666.

[18] Y. Cao, S. Tang, C. Li, P. Zhang, Y. Tan, Z. Zhang, and J. Li, *An optimized ev charging model considering tou price and soc curve*, *IEEE Transactions on Smart Grid* **3** (March, 2012) 388–393.

[19] P. Fan, B. Sainbayar, and S. Ren, *Operation analysis of fast charging stations with energy demand control of electric vehicles*, *IEEE Transactions on Smart Grid* **6** (July, 2015) 1819–1826.

[20] H. Qin and W. Zhang, *Charging scheduling with minimal waiting in a network of electric vehicles and charging stations*, in *Proceedings of the Eighth ACM international workshop on Vehicular inter-networking*, pp. 51–60, ACM, 2011.

[21] A. Gusrialdi, Z. Qu, and M. A. Simaan, *Scheduling and cooperative control of electric vehicles' charging at highway service stations*, in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, pp. 6465–6471, IEEE, 2014.

[22] A. Moradipari, N. Tucker, and M. Alizadeh, *Mobility-aware electric vehicle fast charging load models with geographical price variations*, *IEEE Transactions on Transportation Electrification* **7** (2020), no. 2 554–565.

[23] I. Zenginis, J. Vardakas, N. Zorba, and C. Verikoukis, *Performance evaluation of a multi-standard fast charging station for electric vehicles*, *IEEE Transactions on Smart Grid* **9** (Sep., 2018) 4480–4489.

[24] I. S. Bayram, G. Michailidis, I. Papapanagiotou, and M. Devetsikiotis, *Decentralized control of electric vehicles in a network of fast charging stations*, in *Global Communications Conference (GLOBECOM), 2013 IEEE*, pp. 2785–2790, IEEE, 2013.

[25] C. Liu, M. Zhou, J. Wu, C. Long, and Y. Wang, *Electric vehicles en-route charging navigation systems: Joint charging and routing optimization*, *IEEE Transactions on Control Systems Technology* **27** (2019) 906–914.

[26] E. Bitar and S. Low, *Deadline differentiated pricing of deferrable electric power service*, in *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pp. 4991–4997, IEEE, 2012.

[27] M. Alizadeh, Y. Xiao, A. Scaglione, and M. van der Schaar, *Dynamic incentive design for participation in direct load scheduling programs*, *IEEE Journal of Selected Topics in Signal Processing* **8** (Dec, 2014) 1111–1126.

[28] A.-K. Katta and J. Sethuraman, *Pricing strategies and service differentiation in queues – a profit maximization perspective*, *Working paper* (2005).

[29] R. M. Bradford, *Pricing, routing, and incentive compatibility in multiserver queues*, *European Journal of Operational Research* **89** (1996), no. 2 226–236.

[30] S. Huang, Q. Wu, S. S. Oren, R. Li, and Z. Liu, *Distribution locational marginal pricing through quadratic programming for congestion management in distribution networks*, *IEEE Transactions on Power Systems* **30** (July, 2015) 2170–2178.

[31] R. Li, Q. Wu, and S. S. Oren, *Distribution locational marginal pricing for optimal electric vehicle charging management*, *IEEE Transactions on Power Systems* **29** (Jan, 2014) 203–211.

[32] R. N. Allan, R. Billinton, I. Sjarief, L. Goel, and K. S. So, *A reliability test system for educational purposes-basic distribution system data and results*, *IEEE Transactions on Power Systems* **6** (May, 1991) 813–820.

[33] "Open Access Same-time Information System (OASIS), [Online]." http://oasis.caiso.com/mrioasis/logon.do.

[34] Q. Wu, A. H. Nielsen, J. Ostergaard, S. T. Cha, F. Marra, Y. Chen, and C. Træholt, *Driving pattern analysis for electric vehicle (ev) grid integration study*, in *2010 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT Europe)*, pp. 1–6, Oct, 2010.

[35] "California ISO Supply and Renewables, [Online]." http://www.caiso.com/TodaysOutlook/Pages/supply.aspx. Accessed: Jun-2019.

[36] A. Moradipari and M. Alizadeh, *Pricing and routing mechanisms for differentiated services in an electric vehicle public charging station network*, *IEEE Transactions on smart grid* **11** (2019), no. 2 1489–1499.

[37] A. Moradipari and M. Alizadeh, *Pricing differentiated services in an electric vehicle public charging station network*, in *2018 IEEE Conference on Decision and Control (CDC)*, pp. 6488–6494, IEEE, 2018.

[38] P. Auer, N. Cesa-Bianchi, and P. Fischer, *Finite-time analysis of the multiarmed bandit problem*, *Machine Learning* **47** (2002), no. 2-3 235–256.

[39] S. Agrawal and N. Goyal, *Analysis of Thompson sampling for the multi-armed bandit problem*, in *Conference on Learning Theory*, pp. 39–1, 2012.

[40] V. Dani, T. Hayes, and S. M. Kakade, *Stochastic linear optimization under bandit feedback*, *21st Annual Conference on Learning Theory* (2008).

[41] P. Auer, N. Cesa-Bianchi, and P. Fischer, *Finite-time analysis of the multi-armed bandit problem*, *Machine learning* **47** (2002), no. 2-3 235–256.

[42] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, *Improved algorithms for linear stochastic bandits*, in *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

[43] M. Abeille, A. Lazaric, *et. al.*, *Linear thompson sampling revisited*, *Electronic Journal of Statistics* **11** (2017), no. 2 5165–5197.

[44] S. Agrawal and N. Goyal, *Thompson sampling for contextual bandits with linear payoffs*, in *International Conference on Machine Learning*, pp. 127–135, 2013.

[45] I. Usmanova, A. Krause, and M. Kamgarpour, *Safe convex learning under uncertain constraints*, in *Proceedings of Machine Learning Research* (K. Chaudhuri and M. Sugiyama, eds.), vol. 89 of *Proceedings of Machine Learning Research*, pp. 2106–2114, PMLR, 16–18 Apr, 2019.

[46] A. Kazerouni, M. Ghavamzadeh, Y. Abbasi, and B. Van Roy, *Conservative contextual linear bandits*, in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 3910–3919. Curran Associates, Inc., 2017.

[47] S. Daulton, S. Singh, V. Avadhanula, D. Dimmery, and E. Bakshy, *Thompson sampling for contextual bandit problems with auxiliary safety constraints*, *arXiv preprint arXiv:1911.00638* (2019).

[48] P. Rusmevichientong and J. N. Tsitsiklis, *Linearly parameterized bandits*, *Mathematics of Operations Research* **35** (2010), no. 2 395–411.

[49] S. Vakili, K. Liu, and Q. Zhao, *Deterministic sequencing of exploration and exploitation for multi-armed bandit problems*, *IEEE Journal of Selected Topics in Signal Processing* **7** (2013), no. 5 759–767.

[50] C. Tekin and M. van der Schaar, *Distributed online learning via cooperative contextual bandits*, *IEEE Transactions on Signal Processing* **63** (2015), no. 14 3700–3714.

[51] D. Kalathil, N. Nayyar, and R. Jain, *Decentralized learning for multiplayer multiarmed bandits*, *IEEE Transactions on Information Theory* **60** (2014), no. 4 2331–2345.

[52] C. Tekin and E. Turğay, *Multi-objective contextual multi-armed bandit with a dominant objective*, *IEEE Transactions on Signal Processing* **66** (2018), no. 14 3799–3813.

[53] C. Tekin and E. Turgay, *Multi-objective contextual bandits with a dominant objective*, in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2017.

[54] S. Shahrampour, M. Noshad, and V. Tarokh, *On sequential elimination algorithms for best-arm identification in multi-armed bandits*, *IEEE Transactions on Signal Processing* **65** (2017), no. 16 4281–4292.

[55] C. Gan, R. Zhou, J. Yang, and C. Shen, *Cost-aware cascading bandits*, *IEEE Transactions on Signal Processing* **68** (2020) 3692–3706.

[56] G. Liu and L. Lai, *Action-manipulation attacks against stochastic bandits: Attacks and defense*, *IEEE Transactions on Signal Processing* **68** (2020) 5152–5165.

[57] G. Liu and L. Lai, *Action-manipulation attacks on stochastic bandits*, in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3112–3116, 2020.

[58] Z. Wang, R. Zhou, and C. Shen, *Regional multi-armed bandits with partial informativeness*, *IEEE Transactions on Signal Processing* **66** (2018), no. 21 5705–5717.

[59] S. Vakili and Q. Zhao, *Risk-averse online learning under mean-variance measures*, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1911–1915, 2015.

[60] A. Cassel, S. Mannor, and A. Zeevi, *A general approach to multi-armed bandits under risk criteria*, *arXiv preprint arXiv:1806.01380* (2018).

[61] A. Sani, A. Lazaric, and R. Munos, *Risk-aversion in multi-armed bandits*, in *Advances in Neural Information Processing Systems*, pp. 3275–3283, 2012.

[62] S. Vakili and Q. Zhao, *Risk-averse multi-armed bandit problems under mean-variance measure*, *IEEE Journal of Selected Topics in Signal Processing* **10** (2016), no. 6 1093–1111.

[63] O.-A. Maillard, *Robust risk-averse stochastic multi-armed bandits*, in *International Conference on Algorithmic Learning Theory*, pp. 218–233, Springer, 2013.

[64] A. Aswani, H. Gonzalez, S. S. Sastry, and C. Tomlin, *Provably safe and robust learning-based model predictive control*, *Automatica* **49** (2013), no. 5 1216–1226.

[65] T. Koller, F. Berkenkamp, M. Turchetta, and A. Krause, *Learning-based model predictive control for safe exploration*, in *2018 IEEE Conference on Decision and Control (CDC)*, pp. 6059–6066, IEEE, 2018.

[66] Y. Sui, A. Gotovos, J. W. Burdick, and A. 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, pp. 997–1005, JMLR.org, 2015.

[67] Y. Sui, J. Burdick, Y. Yue, *et. al.*, *Stagewise safe bayesian optimization with gaussian processes*, in *International Conference on Machine Learning*, pp. 4788–4796, 2018.

[68] C. J. Ostafew, A. P. Schoellig, and T. D. Barfoot, *Robust constrained learning-based nmpc enabling reliable mobile robot path tracking*, *The International Journal of Robotics Research* **35** (2016), no. 13 1547–1563.

[69] 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*, pp. 1424–1431, Dec, 2014.

[70] S. Amani, M. Alizadeh, and C. Thrampoulidis, *Linear stochastic bandits under safety constraints*, in *Advances in Neural Information Processing Systems*, pp. 9252–9262, 2019.

[71] K. Khezeli and E. Bitar, *Safe linear stochastic bandits*, *arXiv preprint arXiv:1911.09501* (2019).

[72] A. Pacchiano, M. Ghavamzadeh, P. Bartlett, and H. Jiang, *Stochastic bandits with linear constraints*, in *International Conference on Artificial Intelligence and Statistics*, pp. 2827–2835, PMLR, 2021.

[73] A. Moradipari, C. Thrampoulidis, and M. Alizadeh, *Stage-wise conservative linear bandits*, *Advances in neural information processing systems* **33** (2020) 11191–11201.

[74] T. Chen, A. Gangrade, and V. Saligrama, *Strategies for safe multi-armed bandits with logarithmic regret and risk*, *arXiv preprint arXiv:2204.00706* (2022).

[75] T. Chen, A. Gangrade, and V. Saligrama, *A doubly optimistic strategy for safe linear bandits*, *arXiv preprint arXiv:2209.13694* (2022).

[76] W. Thompson, *On the likelihood that one unknown probability exceeds another in view of the evidence of two samples*, *Biometrika* **25** (1933), no. 3/4 285–294.

[77] D. Russo and B. Van Roy, *Learning to optimize via posterior sampling*, *Mathematics of Operations Research* **39** (2014), no. 4 1221–1243.

[78] E. Kaufmann, N. Korda, and R. Munos, *Thompson sampling: An asymptotically optimal finite-time analysis*, in *International Conference on Algorithmic Learning Theory*, pp. 199–213, Springer, 2012.

[79] I. Osband and B. Van Roy, *Bootstrapped thompson sampling and deep exploration*, *arXiv preprint arXiv:1507.00300* (2015).

[80] D. Russo and B. Van Roy, *An information-theoretic analysis of thompson sampling*, *The Journal of Machine Learning Research* **17** (2016), no. 1 2442–2471.

[81] S. Dong and B. Van Roy, *An information-theoretic analysis for thompson sampling with many actions*, in *Advances in Neural Information Processing Systems*, pp. 4157–4165, 2018.

[82] S. Dong, T. Ma, and B. Van Roy, *On the performance of thompson sampling on logistic bandits*, *arXiv preprint arXiv:1905.04654* (2019).

[83] A. Gopalan and S. Mannor, *Thompson sampling for learning parameterized markov decision processes*, in *Conference on Learning Theory*, pp. 861–898, 2015.

[84] A. Gopalan, S. Mannor, and Y. Mansour, *Thompson sampling for complex online problems*, in *International Conference on Machine Learning*, pp. 100–108, 2014.

[85] A. Moradipari, S. Amani, M. Alizadeh, and C. Thrampoulidis, *Safe linear thompson sampling with side information*, *IEEE Transactions on Signal Processing* **69** (2021) 3755–3767.

[86] A. Moradipari, M. Alizadeh, and C. Thrampoulidis, *Linear thompson sampling under unknown linear constraints*, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3392–3396, IEEE, 2020.

[87] A. Moradipari, S. Shahsavari, A. Esmaeili, and F. Marvasti, *Using empirical covariance matrix in enhancing prediction accuracy of linear models with missing information*, in *2017 International conference on sampling theory and applications (SampTA)*, pp. 446–450, IEEE, 2017.

[88] N. Tucker, A. Moradipari, and M. Alizadeh, *Multi-armed bandit approaches for real-time electricity pricing with grid reliability constraints*, .

[89] A. Moradipari, S. Amani, M. Aliradeh, and C. Thrampoulidis, *Safe linear bandits*, in *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–1, 2021.

[90] S. A. H. Aqajari, E. K. Naeini, M. A. Mehrabadi, S. Labbaf, N. Dutt, and A. M. Rahmani, *pyeda: An open-source python toolkit for pre-processing and feature extraction of electrodermal activity*, *Procedia Computer Science* **184** (2021) 99–106.

[91] R. Trimananda, S. A. H. Aqajari, J. Chuang, B. Demsky, G. H. Xu, and S. Lu, *Understanding and automatically detecting conflicting interactions between smart home iot applications*, in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 1215–1227, 2020.

[92] S. A. H. Aqajari, R. Cao, E. K. Naeini, M.-D. Calderon, K. Zheng, N. Dutt, P. Liljeberg, S. Salanterä, A. M. Nelson, and A. M. Rahmani, *Pain assessment tool with electrodermal activity for postoperative patients: method validation study*, *JMIR mHealth and uHealth* **9** (2021), no. 5 e25258.

[93] A. H. A. Zargari, S. A. H. Aqajari, H. Khodabandeh, A. M. Rahmani, and F. Kurdahi, *An accurate non-accelerometer-based ppg motion artifact removal technique using cyclegan*, *arXiv preprint arXiv:2106.11512* (2021).

[94] M. A. Mehrabadi, S. A. H. Aqajari, I. Azimi, C. A. Downs, N. Dutt, and A. M. Rahmani, *Detection of covid-19 using heart rate and blood pressure: Lessons learned from patients with ards*, in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2140–2143, IEEE, 2021.

[95] S. A. H. Aqajari, R. Cao, A. H. A. Zargari, and A. M. Rahmani, *An end-to-end and accurate ppg-based respiratory rate estimation approach using cycle generative adversarial networks*, in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 744–747, IEEE, 2021.

[96] S. A. H. Aqajari, E. K. Naeini, M. A. Mehrabadi, S. Labbaf, A. M. Rahmani, and N. Dutt, *Gsr analysis for stress: Development and validation of an open source tool for noisy naturalistic gsr data*, *arXiv preprint arXiv:2005.01834* (2020).

[97] M. A. Mehrabadi, S. A. H. Aqajari, A. H. A. Zargari, N. Dutt, and A. M. Rahmani, *Novel blood pressure waveform reconstruction from photoplethysmography using cycle generative adversarial networks*, *arXiv preprint arXiv:2201.09976* (2022).

[98] R. Cao, S. A. H. Aqajari, E. K. Naeini, and A. M. Rahmani, *Objective pain assessment using wrist-based ppg signals: A respiratory rate based method*, in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1164–1167, IEEE, 2021.

[99] A. M. Rahmani, N. Dutt, K. Zheng, A. Nelson, P. Liljeberg, S. Salantera, M. Jiang, A. Anzanpour, E. Syrjala, R. Mieronkoski, *et. al.*, *Pain assessment method and apparatus for patients unable to self-report pain*, Nov. 14, 2019. US Patent App. 16/406,739.

[100] J.-A. Lee, S. A. H. Aqajari, E. Ju, P. Kehoe, L. Gibbs, and A. Rahmani, *Home-visit intervention to reduce stress of underserved family caregivers for persons with dementia*, *Innovation in Aging* **5** (2021), no. Supplement_1 154–154.

[101] P. Sarangi, S. Shahsavari, and P. Pal, *Robust doa and subspace estimation for hybrid channel sensing*, in *2020 54th Asilomar Conference on Signals, Systems, and Computers*, pp. 236–240, IEEE, 2020.

[102] N. Zarmehi, S. Shahsavari, and F. Marvasti, *Comparison of uniform and random sampling for speech and music signals*, in *2017 International Conference on Sampling Theory and Applications (SampTA)*, pp. 552–555, IEEE, 2017.

[103] S. Shahsavari, J. Millhiser, and P. Pal, *Fundamental trade-offs in noisy super-resolution with synthetic apertures*, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4620–4624, IEEE, 2021.

[104] H. Qiao, S. Shahsavari, and P. Pal, *Super-resolution with noisy measurements: Reconciling upper and lower bounds*, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9304–9308, IEEE, 2020.

[105] S. Shahsavari, P. Sarangi, M. C. Hücümenoğlu, and P. Pal, *Ada-jsr: Sample efficient adaptive joint support recovery from extremely compressed measurement vectors*, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9077–9081, IEEE, 2022.

[106] S. Shahsavari, P. Sarangi, and P. Pal, *Beamspace esprit for mmwave channel sensing: Performance analysis and beamformer design*, *Front. Sig. Proc. 1: 820617. doi: 10.3389/frsip* (2022).

[107] S. Shahsavari, P. Sarangi, and P. Pal, *Kr-lista: Re-thinking unrolling for covariance-driven sparse inverse problems*, in *2021 55th Asilomar Conference on Signals, Systems, and Computers*, pp. 1403–1408, IEEE, 2021.

[108] D. Romero, P. Gerstoft, H. Givehchian, and D. Bharadia, *Spoofing attack detection in the physical layer with commutative neural networks*, *arXiv preprint arXiv:2211.04269* (2022).

[109] H. Givehchian, N. Bhaskar, E. R. Herrera, H. R. L. Soto, C. Dameff, D. Bharadia, and A. Schulman, *Evaluating physical-layer ble location tracking attacks on mobile devices*, in *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1690–1704, IEEE, 2022.

[110] A. Nikoofard, H. Givehchian, N. Bhaskar, A. Schulman, D. Bharadia, and P. P. Mercier, *Protecting bluetooth user privacy through obfuscation of carrier frequency offset*, *IEEE Transactions on Circuits and Systems II: Express Briefs* (2022).

[111] S. Shahsavari, J. Chen, and P. Pal, *Exploring the geometry of relu generative priors with applications in cellular mri*, in *2022 56th Asilomar Conference on Signals, Systems, and Computers*, IEEE, 2022.

[112] S. Koga, A. Asgharivaskasi, and N. Atanasov, *Active exploration and mapping via iterative covariance regulation over continuous se (3) trajectories*, in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2735–2741, IEEE, 2021.

[113] S. Koga, A. Asgharivaskasi, and N. Atanasov, *Active slam over continuous trajectory and control: A covariance-feedback approach*, in *2022 American Control Conference (ACC)*, pp. 5062–5068, IEEE, 2022.

[114] P. Yang, Y. Liu, S. Koga, A. Asgharivaskasi, and N. Atanasov, *Learning continuous control policies for information-theoretic active perception*, *arXiv preprint arXiv:2209.12427* (2022).

[115] D. T. Larsson, A. Asgharivaskasi, J. Lim, N. Atanasov, and P. Tsiotras, *Information-theoretic abstraction of semantic octree models for integrated perception and planning*, *arXiv preprint arXiv:2209.10035* (2022).

[116] A. Asgharivaskasi, S. Koga, and N. Atanasov, *Active mapping via gradient ascent optimization of shannon mutual information over continuous se (3) trajectories*, *arXiv preprint arXiv:2204.07623* (2022).

[117] Z. Wang, T. Rajabzadeh, N. Lee, and A. H. Safavi-Naeini, *Automated discovery of autonomous quantum error correction schemes*, *PRX Quantum* **3** (2022), no. 2 020302.

[118] T. Rajabzadeh, Z. Wang, N. Lee, T. Makihara, Y. Guo, and A. H. Safavi-Naeini, *Analysis of arbitrary superconducting quantum circuits accompanied by a python package: Sqcircuit*, *arXiv preprint arXiv:2206.08319* (2022).

[119] T. Rajabzadeh, C. J. Sarabalis, O. Atalar, and A. H. Safavi-Naeini, *Photonics-to-free-space interface in lithium niobate-on-sapphire*, in *CLEO: Science and Innovations*, pp. STu4J–6, Optica Publishing Group, 2020.

[120] N. Tucker, A. Moradipari, and M. Alizadeh, *Constrained thompson sampling for real-time electricity pricing with grid reliability constraints*, *IEEE Transactions on Smart Grid* (2020) 1–1.

[121] Y. Wu, R. Shariff, T. Lattimore, and C. Szepesvári, *Conservative bandits*, in *International Conference on Machine Learning*, pp. 1254–1262, 2016.

[122] Y. Mansour, A. Slivkins, and V. Syrgkanis, *Bayesian incentive-compatible bandit exploration*, in *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pp. 565–582, 2015.

[123] S. Katariya, B. Kveton, Z. Wen, and V. K. Potluru, *Conservative exploration using interleaving*, *arXiv preprint arXiv:1806.00892* (2018).

[124] A. Moradipari, M. Alizadeh, and C. Thrampoulidis, *Linear thompson sampling under unknown linear constraints*, in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3392–3396, 2020.

[125] A. Moradipari, S. Amani, M. Alizadeh, and C. Thrampoulidis, *Safe linear thompson sampling with side information*, *arXiv* (2019) arXiv–1911.

[126] A. Moradipari, M. Ghavamzadeh, and M. Alizadeh, *Collaborative multi-agent stochastic linear bandits*, *arXiv preprint arXiv:2205.06331* (2022).

[127] T. Lai and H. Robbins, *Asymptotically efficient adaptive allocation rules*, *Advances in Applied Mathematics* **6** (1985), no. 1 4–22.

[128] T. Lattimore and C. Szepesvari, *Bandit Algorithms*. Cambridge University Press, 2020.

[129] A. Moradipari, C. Silva, and M. Alizadeh, *Learning to dynamically price electricity demand based on multi-armed bandits*, in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 917–921, IEEE, 2018.

[130] P. Rusmevichientong and J. N. Tsitsiklis, *Linearly parameterized bandits*, *Mathematics of Operations Research* **35** (2010), no. 2 395–411.

[131] D. Foster and A. Rakhlin, *Beyond ucb: Optimal and efficient contextual bandits with regression oracles*, in *International Conference on Machine Learning*, pp. 3199–3210, 2020.

[132] A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire, *Corralling a band of bandit algorithms*, in *COLT*, 2017.

[133] Y. Abbasi-Yadkori, D. Pal, and C. Szepesvari, *Online-to-confidence-set conversions and application to sparse stochastic bandits*, in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, 2012.

[134] Y. Zhu and R. Nowak, *Pareto optimal model selection in linear bandits*, *arXiv:2102.06593* (2021).

[135] D. Foster, C. Gentile, M. Mohri, and J. Zimmert, *Adapting to misspecification in contextual bandits*, in *Advances in Neural Information Processing Systems 35*, pp. 11478–11489, 2020.

[136] D. Haussler, J. Kivinen, and M. K. Warmuth, *Sequential prediction of individual sequences under general loss functions*, *IEEE Trans. Inform. Theory* (1998) 1906–1925.

[137] N. Abe and P. Long, *Associative reinforcement learning using linear probabilistic concepts*, in *ICML*, 1999.

[138] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire, *The non-stochastic multi-armed bandit problem*, *SIAM Journal of Computing* (2002).

[139] O. Maillard and R. Munos, *Adaptive bandits: Towards the best history-dependent strategy*, in *AISTATS*, 2011.

[140] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, *The nonstochastic multiarmed bandit problem*, *SIAM journal on computing* **32** (2002), no. 1 48–77.

[141] S. Bubeck and N. Cesa-Bianchi, *Regret analysis of stochastic and nonstochastic multiarmed bandit problems*, *Foundations and Trends in Machine Learning* (2012).

[142] M. Papini, A. Tirinzoni, M. Restelli, A. Lazaric, and M. Pirotta, *Leveraging good representations in linear contextual bandits*, in *ICML*, 2021.

[143] L. Cella, A. Lazaric, and M. Pontil, *Meta-learning with stochastic linear bandits*, in *International Conference on Machine Learning*, pp. 1360–1370, 2020.

[144] A. Moradipari, M. Ghavamzadeh, T. Rajabzadeh, C. Thrampoulidis, and M. Alizadeh, *Multi-environment meta-learning in stochastic linear bandits*, *arXiv preprint arXiv:2205.06326* (2022).

[145] O. Maillard and S. Mannor, *Latent bandits*, in *ICML*, 2014.

[146] J. Hong, B. Kveton, M. Zaheer, Y. Chow, A. Ahmed, and C. Boutilier, *Latent bandits revisited*, in *NeurIPS*, 2020.

[147] Y. Abbasi-Yadkori, A. Pacchiano, and M. Phan, *Regret balancing for bandit and RL model selection*, *arXiv:2006.05491* (2020).

[148] A. Pacchiano, C. Dann, C. Gentile, and P. Bartlett, *Regret bound balancing and elimination for model selection in bandits and rl*, *arXiv:2012.13045* (2020).

[149] J. Hong, B. Kveton, M. Zaheer, M. Ghavamzadeh, and C. Boutilier, *Thompson sampling with a mixture prior*, in *AISTATS*, 2022.

[150] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, *Proceedings of the IEEE* **86** (1998), no. 11 2278–2324.

[151] A. Krizhevsky *et. al.*, *Learning multiple layers of features from tiny images*, .

[152] M. Tan and Q. V. Le, *Efficientnetv2: Smaller models and faster training*, *arXiv preprint arXiv:2104.00298* (2021).

[153] A. Moradipari, B. Turan, Y. Abbasi-Yadkori, M. Alizadeh, and M. Ghavamzadeh, *Feature and parameter selection in stochastic linear bandits*, in *International Conference on Machine Learning*, pp. 15927–15958, PMLR, 2022.

[154] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[155] J. A. Tropp, *User-friendly tail bounds for sums of random matrices*, *Foundations of computational mathematics* **12** (2012), no. 4 389–434.

[156] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge University Press, 2006.

[157] A. Pacchiano, M. Phan, Y. Abbasi-Yadkori, A. Rao, J. Zimmert, T. Lattimore, and C. Szepesvari, *Model selection in contextual stochastic bandit problems*, *arXiv preprint arXiv:2003.01704* (2020).

[158] A. Cutkosky, C. Dann, A. Das, C. Gentile, A. Pacchiano, and M. Purohit, *Dynamic balancing for model selection in bandits and rl*, in *International Conference on Machine Learning*, pp. 2276–2285, PMLR, 2021.

[159] M. Tan and Q. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, in *International Conference on Machine Learning*, pp. 6105–6114, PMLR, 2019.