RISK AND SAFETY IN ONLINE LEARNING AND OPTIMIZATION: THEORY AND APPLICATIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Kia Khezeli

May 2020

RISK AND SAFETY IN ONLINE LEARNING AND OPTIMIZATION: THEORY AND APPLICATIONS

Kia Khezeli, Ph.D.

Cornell University 2020

This dissertation focuses on risk and safety considerations in the design and analysis of online learning algorithms for sequential decision-making problems under uncertainty. The particular motivating application for the mathematical models and methods developed in this dissertation is demand response programs. Demand response programs denote the general family of mechanisms designed to improve the efficiency and the reliability of electric power systems by affecting the demand of residential customers.

First, we design a risk-sensitive online learning algorithm for linear models. In particular, we consider the setting in which an electric power utility seeks to curtail its peak electricity demand by offering a fixed group of customers a uniform price for reductions in consumption relative to their predetermined baselines. The underlying demand curve, which describes the aggregate reduction in consumption in response to the offered price, is assumed to be affine and subject to unobservable random shocks. Assuming that both the parameters of the demand curve and the distribution of the random shocks are initially unknown to the utility, we investigate the extent to which the utility might dynamically adjust its offered prices to maximize its cumulative risk-sensitive payoff over a finite number of T days. In order to do so effectively, the utility must design its pricing policy to balance the trade-off between the need to learn the unknown demand model (exploration) and maximize its payoff (exploitation)

over time. We propose a semi-greedy pricing policy, and show that its expected regret defined as the risk-sensitive payoff loss over T days, relative to an oracle pricing policy that knows the underlying demand model, is no more than $O(\sqrt{T}\log(T))$. Moreover, the proposed pricing policy is shown to yield a sequence of prices that converge to the oracle optimal prices in the mean square sense.

Second, we develop an online learning algorithm for linear models subject to stagewise safety constraints. More specifically, we introduce the safe linear stochastic bandit framework — a generalization of linear stochastic bandits — where, in each stage, the learner is required to select an arm with an expected reward that is no less than a predetermined (safe) threshold with high probability. We assume that the learner initially has knowledge of an arm that is known to be safe, but not necessarily optimal. Leveraging on this assumption, we introduce a learning algorithm that systematically combines known safe arms with exploratory arms to safely expand the set of safe arms over time, while facilitating safe greedy exploitation in subsequent stages. In addition to ensuring the satisfaction of the safety constraint at every stage of play, the proposed algorithm is shown to exhibit an expected regret that is no more than $O(\sqrt{T\log(T)})$ after T stages of play.

Third, we extend our methodology developed for linear models to design an online learning algorithm with near-optimal performance for a more general class of nonparametric smooth reward models. Specifically, we adopt the perspective of an aggregator, which seeks to coordinate its *purchase* of demand reductions from a fixed group of residential electricity customers, with its *sale* of the aggregate demand reduction in a two-settlement wholesale energy market. The aggregator procures reductions in demand by offering its customers

a uniform price for reductions in consumption relative to their predetermined baselines. Prior to its realization of the aggregate demand reduction, the aggregator must also determine how much energy to sell into the two-settlement energy market. In the day-ahead market, the aggregator commits to a forward contract, which calls for the delivery of energy in the real-time market. The underlying aggregate demand curve, which relates the aggregate demand reduction to the aggregator's offered price, is assumed to be unknown and subject to unobservable, random shocks. Assuming that both the demand curve and the distribution of the random shocks are initially unknown to the aggregator, we investigate the extent to which the aggregator might dynamically adapt its offered prices and forward contracts to maximize its expected profit over a time window of T days. Specifically, we design a dynamic pricing and contract offering policy that resolves the aggregator's need to learn the unknown demand model with its desire to maximize its cumulative expected profit over time. In particular, the proposed pricing policy is proven to incur an expected regret over T days that is no greater than $O(\sqrt{T}\log^2(T))$.

BIOGRAPHICAL SKETCH

Kia Khezeli received the B.Sc. degree in Electrical Engineering from Sharif University of Technology in 2012, and the M.Sc. degree in Electrical and Computer Engineering from McMaster University in 2014. He earned the M.Sc. and Ph.D. degrees in Electrical and Computer Engineering from Cornell University in 2018 and 2020, respectively.

His research interests include applied probability and statistics, statistical machine learning, stochastic optimization, and information theory. He has worked on developing mathematical models and methods for sequential decision-making problems under uncertainty with practical applications such as demand response programs in electric power systems.

He was a recipient of the Cornell ECE Outstanding Ph.D. TA Award in 2020. He was also a recipient of Irwin and Joan Jacobs Fellowship from the School of Electrical and Computer Engineering at Cornell University, the Outstanding Thesis Award from the department of Electrical and Computer Engineering at McMaster University, and the National Elite Foundation Fellowship from Sharif University of Technology.

To my parents, *Fariba* and *Kiumars*.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to my advisor Professor Eilyan Bitar for his guidance, support, and mentorship throughout my Ph.D. studies. Professor Bitar's clarity of thought, creativity, deep technical knowledge, and philosophical approach towards research has been a constant inspiration to me. He also helped me with great care in improving my presentation and writing skills, for which I will always be grateful.

I would like to thank my dissertation committee Professors Qing Zhao and Shane Henderson for all their helpful guidance and support in preparing this dissertation. I would like to thank Professors David Delchamps and Kevin Tang who I had the pleasure of being their teaching assistant, for their mentorship and their support. I would also like to thank Professor Ziv Goldfeld for broadening my perspective of research in the intersection of information theory and machine learning. He inspired me with his deep understanding of probability and statistics. Lastly, I would like to thank my former advisor at McMaster University, Professor Jun Chen for his support throughout my graduate studies.

I would like to thank my colleagues and friends at Cornell, Raphael Louca, Weixuan Lin, Daniel Munoz Alvarez, Polina Alexeenko, Kolbeinn Karlsson, Subhonmensh Bose, and Shih-Hao Tseng, for making my graduate studies more enjoyable.

I consider myself incredibly fortunate to have great friends who have always supported me in life. I would like to give a special mention to Omid Javidbakht, Poolad Imany, Babak Yazdanpanah, Kasra Ghaemi, Milad Taghavi, Caroline Motzer, Rouzbeh Kananizadeh, Ali Makhdoumi, Sattar Vakili, Karen Khatami, Pedram Madadkar, Kaveh Moussakhani, Reza Ghaemi, Hugh Bullen, and Heidi Kaila.

Most importantly, I would like thank my parents, Arman, Mina, and Roozbeh for all their encouragement, love, and support. Without my parents' unwavering support, unconditional love, and sacrifices this dissertation would not have been possible.

Kia Khezeli

Ithaca, NY
May 2020

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The focus of this dissertation is on the design and analysis of provably safe online learning algorithms for sequential decision-making problems under uncertainty. In this class of problems, a learner interacts with an uncertain and partially unknown environment over a finite (or an infinite) number of stages with the objective of maximizing a cumulative reward function. The general framework for the sequential decision-making problems is depicted in Figure 1.1. There are a number of practical applications that fit within this sequential decision-making paradigm. These include power systems operation, clinical trials, online advertisement, robotic systems, and stock markets. The motivating application for several problems studied in this dissertation pertains to the design of demand response programs in modern electric power systems.

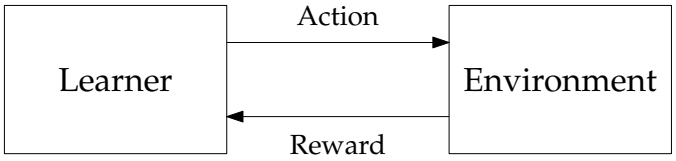


Figure 1.1: Sequential decision-making problem.

Demand response programs refer to programs operated by utility companies to affect the electric power consumption of consumers (e.g., residential customers) with the objective of improving the efficiency and the reliability of electric power systems. In other words, demand response programs are mechanisms designed to utilize the inherent flexibility of demand side resources to provide specific services facilitating the operation of electric power systems. For instance, an electric power utility (or another third party entity) manually ad-

justs flexible loads in response to wholesale energy prices, e.g., by moving deferrable loads off peak. Electric vehicles are an example of such deferrable loads that due to their considerable capacity and flexibility have been the target of several demand response programs, some of which are currently operational, e.g., OptimizeEV administered by NYSEG [72]. Thermostatically Controlled Loads (TLC)s, e.g., air conditioning units, are another category of flexible demand resources that have been the focus of demand response programs. The control mechanism and the incentives provided to customers for participation vary from one demand response program to another. Despite such differences, many of these programs fit within the framework of sequential decision-making problem as illustrated in Figure 1.2.

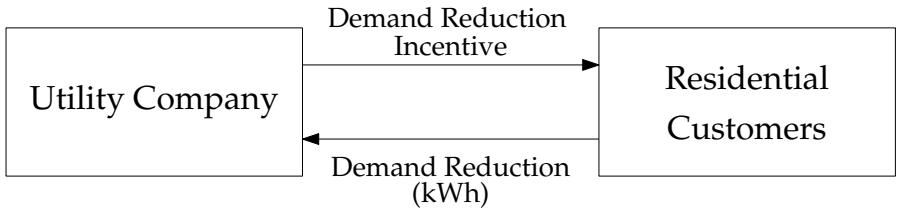


Figure 1.2: Modeling demand response program as a sequential decision-making problem.

In Chapters 2 and 4, we consider a class of DR programs in which an electric power utility seeks to elicit a reduction in the aggregate electricity demand of a fixed group of customers, during peak demand periods as shown in Figure 1.3. More specifically, in this class of demand response programs known as Peak-Time Rebate (PTR) programs, an electric power utility elicits demand reduction from a group of residential customers by offering a non-discriminatory price for demand reduction from customers' baseline consumption. The underlying demand function, which models the reduction in consumption of customers in

response to the DR price is assumed to be initially unknown to the utility. Due to this ignorance, the utility faces a challenge as to how to set the price for demand reduction. We take an online learning approach from the perspective of the utility. More specifically, we design an online learning algorithm to sequentially adjust the offered prices with the objective of maximizing a cumulative reward function. There are several reward functions that one may consider for this particular application. In Chapter 2, we consider the objective of maximizing the cumulative risk-sensitive revenue, which is defined as the revenue that the utility is guaranteed to receive with a user-specified probability. In Chapter 4, we consider the objective of maximizing the cumulative expected profit from selling the aggregate demand reduction in wholesale energy markets.

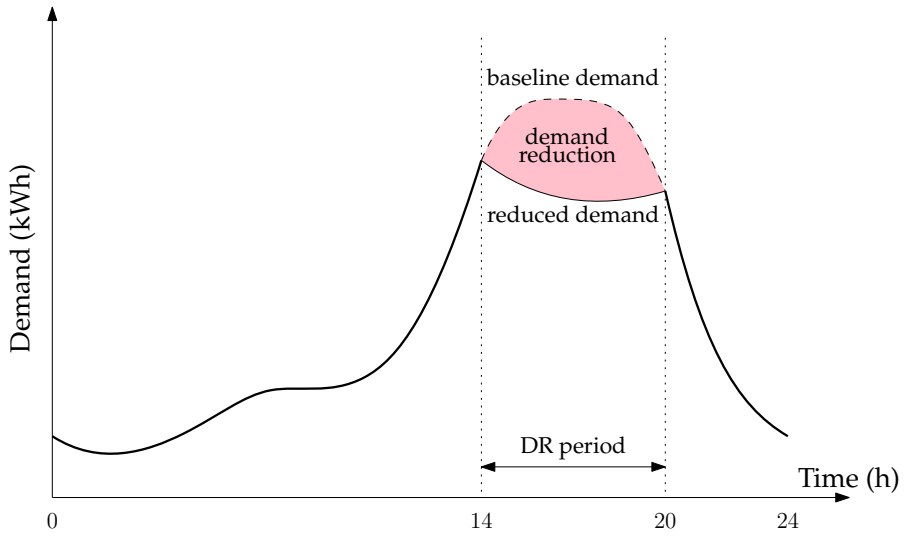


Figure 1.3: Demand reduction during peak demand hours in a demand resposne program.

1.1 Challenges

1.1.1 Unknown Environments

The most basic challenge that the learner faces is due to its ignorance about the environment or the reward function. More specifically, the learner faces a dilemma in deciding upon which actions to take at each stage. On the one hand, the learner can *exploit* the information at hand. More precisely, the learner can estimate the underlying model using the information gathered in preceding stages, and, then take the *greedy* action that is optimal assuming the correctness of the estimated model. On the other hand, the learner can *explore* by taking an action that reveals information about the underlying environment (e.g., a randomly chosen action). This information, in turn, can be utilized to improve the learner's decision-making ability in subsequent stages. Naturally, there exists a trade-off between exploration and exploitation as taking the greedy action may not elicit new information. Balancing this trade-off is critical in achieving the objective of maximizing the cumulative reward. There are, however, other considerations such as safety that are of significant importance for practical applications.

1.1.2 Safety

One of the primary challenges of implementing algorithms developed for the aforementioned class of problems in real-world applications (e.g., robotic systems, clinical trials, electric power systems) pertains to *safety* [39]. In such applications, taking an unsafe action may cause an irreversible damage to the un-

derlying system. Consequently, the learner must incorporate safety considerations in the design of online learning algorithms. There are several real-world incidents in which failure to ensure safety of the learning methods resulted in catastrophic outcomes. In 2016, a Microsoft AI chatbot started twitting offensive remarks in less than 24 hours of its implementation[1]. In 2018, IBM's algorithm that recommends personalized cancer treatments for patients allegedly recommended unsafe and dangerous treatments[2]. Although the notion of safety has different implications across different fields (e.g., avoiding discriminatory and offensive language on social networks, and not jeopardizing patients' health in personal healthcare recommendations) such incidents call for the careful consideration of safety in designing online learning algorithms.

1.2 Algorithm Design Considerations

1.2.1 Failure of A Greedy Approach

In this section, we provide a more detailed account of the failure of greedy approaches as it pertains to the trade-off between exploration and exploration introduced in Section 1.1.1. For the clarity of exposition, in this section, we restrict our focus to parameterized models. That is, we assume that the reward function is characterized by an unknown parameter $\theta^\star$. Had the learner known the reward parameter $\theta^\star$, then, there was no need for exploration as the learner could have taken the optimal action at every stage of play. A natural greedy approach

[1]https://en.wikipedia.org/wiki/Tay_(bot)

[2]https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

for the learner at each stage is to first estimates the reward parameter using the history of observations, and then, take an action that is optimal assuming that the estimated parameter is correct.

This intuitive greedy approach, which is also known as the *myopic* or *certainty equivalent* approach[3] has been extensively studied in the literature. Historically, such approaches were incorrectly believed to be *asymptotically efficient*. In particular, in 1976, Anderson and Taylor [4] proposed the Least Squares Certainty Equivalence (LSCE) control law for a scalar linear system of the form

$$Y_t = \langle X_t, \theta^\star \rangle + \varepsilon_t,$$

where Y_t, X_t, and ε_t model the output, action, and noise at stage t, respectively. They consider the objective of driving the output Y_t to a prespecified level $y^\star$, i.e., maximizing the cumulative reward function defined as $-\sum_{t=1}^{T}(Y_t - y^\star)^2$. Using numerical simulations, they conjecture that the actions chosen under their proposed method converge to the optimal action with probability 1. However, their conjecture was later shown to be incorrect by Lai and Robbins [56]. More specifically, Lai and Robbins show that the LSCE control law with a positive probability takes a sub-optimal action in all stages $t \geq 3$.

The convergence of the sequence of parameter estimates $\widehat{\theta}_t$ to an uninformative value different from the reward parameter $\theta^\star$ with positive probability, i.e., $\mathbf{P}\left(\lim_{t\to\infty} \widehat{\theta}_t \neq \theta^\star\right) > 0$ is referred to by *incomplete learning* in the literature. For a more detailed analysis of the certainty equivalence approach and incomplete learning, we refer the reader to [51] and the references therein.

[3]This approach stems from the certainty equivalence principle, which refers to "the procedure to obtain control policies for stochastic systems by considering the optimal control policies for the related deterministic systems where the random variables are replaced by their expected values" [5].

We close this section by describing *indeterminate equilibria*, a concept that is closely related to incomplete learning. The indeterminate equilibria denotes the set of parameters θ for which the observation under the respective greedy actions will subsequently confirms the belief that θ is the underlying model parameter. We now clarify this concept with a detailed example. Consider the problem of dynamic pricing of a single good with an unknown demand function with the objective of maximizing the cumulative expected revenue of the seller. Assume that d_t, the demand at stage t is of the form

$$d_t = -\alpha^\star p_t + \beta^\star + \varepsilon_t,$$

where $\alpha^\star > 0$ and $\beta^\star > 0$ are unknown model parameters and ε_t is a zero-mean random variable modeling the demand shock. It is straight forward to observe that the optimal price $p^\star$ is given by $p^\star = \beta^\star/(2\alpha^\star)$. Indeterminate equilibria, depicted in Figure 1.4 is the set of parameters (α, β) such that $-\alpha^\star p + \beta^\star = -\alpha p + \beta$ for $p = \beta/(2\alpha)$. Under the greedy approach, the sequence of parameters may converge to an uninformative value belonging to the indeterminate equilibria. That is, subsequent observations confirm the correctness of an "incorrect" belief.

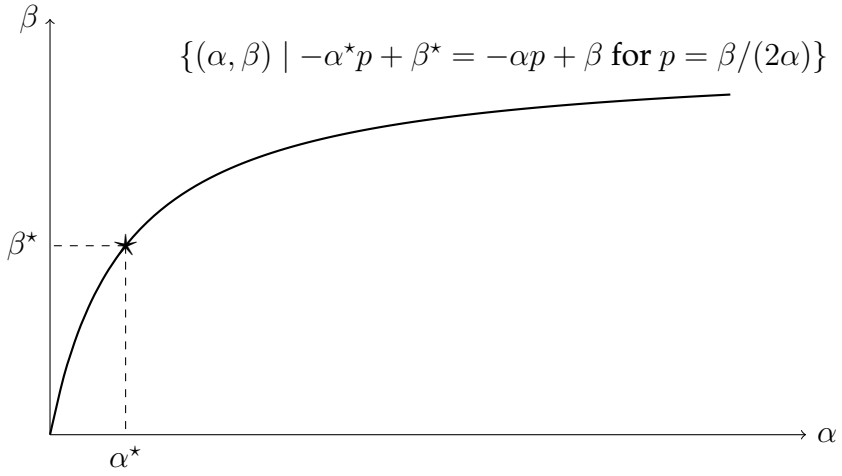


Figure 1.4: Indeterminate equilibria for a linear model.

1.2.2 Exploration vs. Exploitation

The performance of an online learning algorithm is determined by its ability to balance the trade-off between exploration and exploitation. On the one hand, insufficient exploration may result in sub-optimal actions in subsequent stages due to lack of information about the environment. On the other hand, excessive exploration may negatively impact the cumulative reward by taking sub-optimal exploratory actions more than necessary. There are several approaches introduced in the literature to balance this trade-off in a way that over time, the cumulative reward converges to the optimal value (i.e., the cumulative reward gained by taking the optimal action at every stage). We briefly describe some of the well-known approaches in this active area of research. We will provide a more detailed review of the literature related to models studied in this dissertation in respective subsequent chapters.

Thompson Sampling

Thompson Sampling (TS), introduced in early 1930's by W. R. Thompson [88], is perhaps the first method designed to balance the trade-off between exploration and exploitation. Under this approach, the learner assumes a prior distribution on the reward of each action (or on the parameter of the reward). At each stage, the learner selects an action randomly according to its probability of being the optimal action.

The trade-off between exploration and exploitation is implicit under Thompson Sampling. That is, at each stage, after observing the reward of the chosen action, the posterior distribution is updated. Depending on this observation, the

probability of taking an exploitative action (e.g., the certainty equivalent action associated with the maximum likelihood estimate of the parameter) is increased or decreased in the proceeding stage.

Optimism in the Face of Uncertainty

In 1985, Lai and Robbins [57] introduced the principle of Optimism in the Face of Uncertainty (OFU). Under this approach, at each stage, the learner construct upper confidence bounds (UCB) on the reward of each action. The learner, then, plays the most optimistic action (i.e., the action with the largest upper confidence bound on its reward).

The UCB of the reward of an action is a measure of both its reward and uncertainty. More precisely, the UCB of the reward an action being large implies either that the estimated reward is large (exploitation) or the reward is highly uncertain (exploration).

Information Directed Sampling

The Information Directed Sampling (IDS) approach [74] quantifies the trade-off between exploration and exploitation by defining an information ratio. The information ratio of each action measures the cost that the learner incurs (i.e., stagewise regret) per bit of information gained from taking the said action. At each stage, the learner selects an action that minimizes this information ratio.

1.2.3 Safety

TS, OFU, and IDS introduced in section 1.2.2 have been tailored to various configurations of sequential decision-making problems. In particular, they are shown to perform near-optimally for several variations of the bandit problem [7, 1, 2, 3, 74, 53] and reinforcement learning [45, 41, 8]. However, the mentioned approaches may not be directly applicable in safety-critical applications as they may take unsafe actions during the learning process. As one may expects, different notions of risk and safety require distinct approaches to algorithm design. In what follows, we summarize the particular safety and risk considerations studied in this dissertation, our contributions, and the organization of the dissertation.

1.3 Summary of Contributions and Dissertation Organization

In the majority of the dissertation, we consider linear reward models, i.e., the reward associated with each action is a linear function of the action with an unknown parameter. In Chapter 2, we consider the objective of maximizing a risk-sensitive reward function, which is closely related to the concept of Value-at-Risk (VaR). In Chapter 3, we consider the objective of maximizing expected reward subject to a more restrictive safety measure. More precisely, we model safety as a stagewise probabilistic constraint on the reward of the chosen action.

Although linear functions may be too simplistic to model a complex environment, studying them can provide insight on how to develop algorithms for more complex systems. In particular, in the final Chapter of the dissertation, we

utilize misspecified linear proxies to develop an algorithm with near-optimal performance for a more general family of smooth nonparametric reward models. Specifically, the algorithm estimates local linearizations of the demand model through a careful data selection procedure. This careful data selection procedure is needed to account for the potential model misspecification due to linearization. The algorithm then takes semi-greedy actions with respect to these estimated linear functions.

1.3.1 Risk-Sensitive Online Learning of Linear Models

In Chapter 2, we study the risk-sensitive dynamic pricing problem with application to modern electric power systems. More specifically, we consider the problem of pricing demand response programs and model the demand reduction from customers as an affine function of the demand response price subject to additive random demand shocks. We assume that, a priori, the learner neither knows the demand model parameters nor the distribution of the demand shock. The objective of the learner, alias the electric power utility administrating the demand response program, is to maximize her cumulative risk-sensitive revenue. The risk-sensitive revenue is defined as the revenue that the utility is subject to receive with probability $1 - \alpha$ where $\alpha \in (0, 1)$ is the utility's risk tolerance.

Contributions

The consideration of risk in the reward function (the revenue of the utility) creates the need to learn the distribution of the demand shock as well as the de-

mand model parameters. We propose a pricing policy, which carefully applies perturbations to the sequence of certainty equivalent prices, under which the utility learns both the demand model parameters and the distribution of the demand shock. We show that our proposed method exhibits near-optimal expected regret, which is guaranteed to be no more than $O(\sqrt{T}\log(T))$. Interestingly, we show that the variations in the sequence of wholesale electricity prices eliminates the need to generate exogenous exploration in the sequence of demand response prices. This, in turn, results in a remarkable improvement of the expected regret exhibited by the myopic policy. More precisely, the expected regret under the myopic policy is no more than $O(\log^2(T))$ when the sequence of wholesale electricity prices varies over time. It is worth noting that such variation in the sequence of wholesale electricity prices naturally occurs in wholesale electricity markets across the United States.

1.3.2 Safe Online Learning of Linear Models

In Chapter 3, we introduce the safe linear stochastic bandit framework — a generalization of linear stochastic bandit — where, in each stage, the learner is required to select an arm with an expected reward that is no less than a predetermined (safe) threshold with high probability.

Contributions

We propose a new learning algorithm that is tailored to the safe linear bandit framework. The proposed algorithm is shown to exhibit near-optimal expected regret, while guaranteeing the satisfaction of the proposed safety constraint at

every stage of play. To the best of our knowledge our work is the first of its kind that proposes an algorithm with stagewise safety guarantee that exhibits near-optimal regret for the stochastic linear bandit framework.

1.3.3 Online Learning of Nonparametric Models

We adopt the perspective of an aggregator, which seeks to coordinate its purchase of demand reductions from a fixed group of residential electricity customers, with its sale of the aggregate demand reduction in a two-settlement wholesale energy market. The aggregator procures reductions in demand by offering its customers a uniform price for reductions in consumption relative to their predetermined baselines. Prior to its realization of the aggregate demand reduction, the aggregator must also determine how much energy to sell into the two-settlement energy market.

Contributions

We generalize the linear model studied in Chapters 2 and 3 to a class of smooth nonparametric models. We propose an online learning policy that coordinates the purhcase of demand response from customers with its sale in the wholesale energy market. We show that the upper bound on the expected regret of our proposed policy is no more than $O(\sqrt{T}\log^2(T))$. The proposed method generalizes the approach developed by Besbes and Zeevi [12] for the dynamic pricing problem to the two-settlement market model.

CHAPTER 2

RISK-SENSITIVE ONLINE LEARNING OF LINEAR MODELS

2.1 Introduction

The ability to implement residential *demand response* (DR) programs at scale has the potential to substantially improve the efficiency and reliability of electric power systems. In this chapter, we consider a class of DR programs in which an electric power utility seeks to elicit a reduction in the aggregate electricity demand of a fixed group of customers, during peak demand periods. The class of DR programs we consider rely on non-discriminatory, price-based incentives for demand reduction. That is to say, each participating customer is remunerated for her reduction in electricity demand according to a uniform price determined by the utility.

There are several challenges a utility faces in implementing such programs, the most basic of which is the prediction of how customers will adjust their aggregate demand in response to different prices – the so-called aggregate demand curve. The extent to which customers are willing to forego consumption, in exchange for monetary compensation, is contingent on variety of idiosyncratic and stochastic factors – the majority of which are initially unknown or not directly measurable by the utility. The utility must, therefore, endeavor to learn the behavior of customers over time through observation of aggregate demand reductions in response to its offered prices for DR. At the same time, the utility must set its prices for DR in such a manner as to promote increased earnings over time. As we will later establish, such tasks are inextricably linked, and give rise to a trade-off between *learning* (exploration) and *earning* (exploitation)

in pricing demand response over time.

Contribution and Related Work: We consider the setting in which the electric power utility is faced with a demand curve that is affine in price, and subject to unobservable, additive random shocks. Assuming that both the parameters of the demand curve and the distribution of the random shocks are initially unknown to the utility, we investigate the extent to which the utility might dynamically adjust its offered prices for demand curtailment to maximize its cumulative risk-sensitive payoff over a finite number of T days. We define the utility's payoff on any given day as the largest return the utility is guaranteed to receive with probability no less than $1 - \alpha$. Here, $\alpha \in (0, 1)$ encodes the utility's sensitivity to risk. In this chapter, we propose a causal pricing policy, which resolves the trade-off between the utility's need to learn the underlying demand model and maximize its cumulative risk-sensitive payoff over time. More specifically, the proposed pricing policy is shown to exhibit an expected payoff loss over T days – relative to an oracle that knows the underlying demand model – which is at most $O(\sqrt{T}\log(T))$. Moreover, the proposed pricing policy is shown to yield a sequence of offered prices, which converges to the sequence of oracle optimal prices in the mean square sense.

There is a related stream of literature in operations research and adaptive control [12, 30, 50, 56, 29], which considers a similar setting in which a monopolist endeavors to sell a product over multiple time periods – with the aim of maximizing its cumulative expected revenue – when the underlying demand curve (for that product) is unknown and subject to exogenous shocks. What distinguishes our formulation from this prevailing literature is the explicit treatment of risk-sensitivity in the optimization criterion we consider, and the subse-

quent need to design pricing policies that not only learn the underlying demand curve, but also learn the shock distribution.

Focusing explicitly on demand response applications, there are several related papers in the literature, which formulate the problem of eliciting demand response under uncertainty within the framework of multi-armed bandits [87, 48, 44, 96]. In this setting, each arm represents a customer or a class of customers. Taylor and Mathieu [87] show that, in the absence of exogenous shocks on load curtailment, the optimal policy is indexable. Kalathil and Rajagopal [48] consider a similar multi-armed bandit setting in which a customer's load curtailment is subject to an exogenous shock, and attenuation due to fatigue resulting from repeated requests for reduction in demand over time. They propose a policy, which guarantees that the T-period regret is bounded from above by $O(\sqrt{T \log T})$. There is a related stream of literature, which treats the problem of pricing demand response under uncertainty using techniques from online learning [40, 47, 70, 79]. Perhaps closest to the setting considered in this chapter, Jia et al. [47] consider the problem of pricing demand response when the underlying demand function is unknown, affine, and subject to normally distributed random shocks. With the aim of maximizing the utility's expected surplus, they propose a stochastic approximation-based pricing policy, and establish an upper bound on the T-period regret that is of the order $O(\log T)$. There is another stream of literature, which considers an auction-based approach to the procurement of demand response [13, 14, 63, 67, 75, 98, 85]. In such settings, the primary instrument for analysis is game-theoretic in nature.

Organization: The rest of the chapter is organized as follows. In Section 2.2, we develop the demand model and formulate the utility's pricing problem for

demand response. In Section 2.3, we outline a scheme for demand model learning. In Section 2.4, we propose a pricing policy and analyze its performance. We investigate the behavior of the proposed pricing policy with a numerical case study in Section 2.6. All mathematical proofs are presented in Appendix Chapter A.

2.2 Model

2.2.1 Demand Response Model

We consider a class of demand response (DR) programs in which an electric power utility seeks to elicit a reduction in peak electricity demand from a fixed group of N customers over multiple time periods (e.g., days) indexed by $t = 1, 2, \ldots$. The class of DR programs we consider rely on uniform price-based incentives for demand reduction.[1] Specifically, prior to each time period t, the utility broadcasts a single price p_t (\$/kWh), to which each participating customer i responds with a reduction in demand D_{it} (kWh) – thus entitling customer i to receive a payment in the amount of $p_t D_{it}$.[2]

We model the response of each customer i to the posted price p_t at time t according to a linear demand function given by

$$D_{it} = a_i p_t + b_i + \varepsilon_{it}, \quad \text{for } i = 1, \ldots, N,$$

[1] This class of DR programs falls within the more general category of programs that rely on *peak time rebates* (PTR) as incentives for demand reduction [33].

[2] A customer's reduction in demand is measured against a predetermined baseline. The question as to how such baselines might be reliably inferred is a challenging and active area of research [19, 22, 24, 25, 68]. Expanding our model to make endogenous the calculation of customer baselines is left as a direction for future research.

where $a_i \in \mathbb{R}$ and $b_i \in \mathbb{R}$ are model parameters *unknown to the utility*, and ε_{it} is an unobservable demand shock, which we model as a random variable with zero mean.[3] *Its distribution is also unknown to the utility*. We define the aggregate response of customers at time t as $D_t := \sum_{i=1}^{N} D_{it}$, which satisfies

$$D_t = ap_t + b + \varepsilon_t. \tag{2.1}$$

Here, the aggregate model parameters and shock are defined as $a := \sum_{i=1}^{N} a_i$, $b := \sum_{i=1}^{N} b_i$, and $\varepsilon_t := \sum_{i=1}^{N} \varepsilon_{it}$. To simplify notation in the sequel, we write the deterministic component of aggregate demand as $\lambda(p, \theta) := ap + b$, where $\theta := (a, b)$ denotes the aggregate demand function parameters.

We assume throughout the chapter that $a \in [\underline{a}, \overline{a}]$ and $b \in \left[0, \overline{b}\right]$, where the model parameter bounds are assumed to be known and satisfy $0 < \underline{a} \leq \overline{a} < \infty$ and $0 \leq \overline{b} < \infty$. Such assumptions are natural, as they ensure that the price elasticity of aggregate demand is strictly positive and bounded, and that reductions in aggregate demand are guaranteed to be nonnegative in the absence of demand shocks. We also assume that the sequence of shocks $\{\varepsilon_t\}$ are independent and identically distributed random variables, in addition to the following technical assumption.

Assumption 1. The aggregate demand shock ε_t has a bounded range $[\underline{\varepsilon}, \overline{\varepsilon}]$, and a cumulative distribution function F, which is bi-Lipschitz over this range. Namely, there exists a real constant $L \geq 1$, such that for all $x, y \in [\underline{\varepsilon}, \overline{\varepsilon}]$, it holds that

$$\frac{1}{L} |x - y| \leq |F(x) - F(y)| \leq L |x - y| .$$

There is a large family of distributions respecting Assumption 1 including uniform and doubly truncated normal distributions. Moreover, the assumption

[3] We note that the assumption that ε_{it} be zero-mean is without loss of generality.

that the aggregate demand shock takes bounded values is natural, given the inherent physical limitation on the range of values that demand can take. And, technically speaking, the requirement that F be bi-Lipschitz is stated to ensure Lipschitz continuity of its inverse, which will prove critical to the derivation of our main results. Finally, we note that the electric power utility need not know the parameters specified in Assumption 1, beyond the assumption of their boundedness.

Remark 1 (On the Linearity Assumption)**.** While the assumption of linearity in the underlying demand model might appear restrictive at first glance, there are several sensible arguments in support of its adoption. First, the assumption of linearity is routinely employed in the revenue management and pricing literature [11, 47, 46, 50, 83, 86], as it serves to facilitate theoretical analyses, thereby bringing to light key features of the problem and its solution structure. More practically, if the range of allowable prices is sufficiently limited, then it is reasonable to assume that the underlying (possibly nonlinear) demand function is well approximated by an affine function over that range. And, in the specific context of pricing for DR programs, it is reasonable to expect that the electric power utility, being a regulated company, will face restrictions on the range of prices that it can offer to customers. Finally, there are recent results in the revenue management literature [12], which demonstrate how the assumption of a linear demand model might be dynamically *adapted* to price in environments where the true demand function is nonlinear. In Chapter 4, we generalize and adopt such techniques to the two-sided optimization problem of selling uncertain demand response resources in wholesale energy markets.

2.2.2 Utility Model and Pricing Policies

We consider a setting in which the utility seeks to reduce its peak electricity demand over multiple days, indexed by t. Accordingly, we let w_t (\$/kWh) denote the wholesale price of electricity during peak demand hours on day t. And, we let f (\$/kWh) denote the retail price of electricity, i.e., the fixed price that customers are charged for their electricity consumption. For the remainder of the chapter, it will be convenient to work with the difference between the wholesale and retail prices of electricity on each day t, which we denote by $c_t := w_t - f$. We assume throughout the chapter that $c_t \in [0, \overline{c}]$ for all days t, where $0 \leq \overline{c} < \infty$.[4] In addition, we assume that c_t is known to the utility prior to its determination of the DR price p_t in each period t. Upon broadcasting a price p_t to its customer base, and realizing an aggregate demand reduction D_t, the utility derives a net reduction in its peak electricity cost in the amount of $(c_t - p_t)D_t$. Henceforth, we will refer to the net savings $(c_t - p_t)D_t$ as the *revenue* derived by the utility in period t.

The utility is assumed to be *sensitive to risk*, in that it would like to set the price for DR in each period t to maximize the revenue it is guaranteed to receive with probability no less than $1 - \alpha$. Clearly, the parameter $\alpha \in (0, 1)$ encodes the degree to which the utility is sensitive to risk. Accordingly, we define the *risk-sensitive revenue* derived by the utility in period t given a posted price p_t as

$$r_\alpha(p_t) := \sup \{x \in \mathbb{R} : \mathbf{P}((c_t - p_t)D_t \geq x) \geq 1 - \alpha\}. \tag{2.2}$$

The risk measure specified in (2.2) is closely related to the standard concept of

[4] Implicit in this requirement is the assumption that $f \leq w_t \leq \overline{c} + f$ for all days t. The lower bound on w_t implies that the utility will only call for a demand reduction on those days in which the wholesale market manifests in prices that exceed the fixed retail price for electricity. The upper bound on w_t implies the enforcement of a *price cap* in the wholesale market.

value at risk commonly used in mathematical finance. Conditioned on a fixed price p_t, one can reformulate the expression in (2.2) as

$$r_\alpha(p_t) = (c_t - p_t)(\lambda(p_t, \theta) + F^{-1}(\alpha)), \tag{2.3}$$

where $F^{-1}(\alpha) := \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}$ denotes the α-quantile of the random variable ε_t. It is immediate to see from the simplified expression in (2.3) that $r_\alpha(p_t)$ is strictly concave in p_t. Let p_t^* denote the *oracle optimal price*, which maximizes the risk-sensitive revenue in period t. Namely,

$$p_t^* := \operatorname{argmax} \{r_\alpha(p_t) : p_t \in \mathbb{R}\}.$$

The optimal price is readily derived from the corresponding first order optimality condition, and is given by

$$p_t^* = \frac{c_t}{2} - \frac{b + F^{-1}(\alpha)}{2a}.$$

Notice that the optimal price may be negative if the wholesale price c_t is small or the risk parameter α is large. However, it is natural to assume that the utility seeks to maximize the revenue it is guaranteed to receive with high probability, i.e., choose a small α. Hence, in practice, it is unlikely to observe such negative optimal prices. We define the *oracle risk-sensitive revenue* accumulated over T time periods as

$$R_T^* := \sum_{t=1}^{T} r_\alpha(p_t^*).$$

The term oracle is used, as R_T^* equals the maximum risk-sensitive revenue achievable by the utility over T periods if it were to have *perfect knowledge* of the demand model.

In the setting considered in this chapter, we assume that both the demand model parameters $\theta = (a, b)$ and the shock distribution F are *unknown to the utility* at the outset. As a result, the utility must attempt to learn them over time by

observing aggregate demand reductions in response to offered prices. Namely, the utility must endeavor to learn the demand model, while simultaneously trying to maximize its risk-sensitive returns over time. As we will later see, such task will naturally give rise to a trade-off between *learning* (exploration) and *earning* (exploitation) in pricing demand response over time. First, we describe the space of feasible pricing policies.

We assume that, prior to its determination of the DR price in period t, the utility has access to the entire history of prices and demand reductions until period $t-1$. We, therefore, define a *feasible pricing policy* as an infinite sequence of functions $\pi := (p_1, p_2, \dots)$, where each function in the sequence is allowed to depend only on the past history. More precisely, we require that the function p_t be measurable according to the σ-algebra generated by the history of past decisions and demand observations $(p_1, \dots, p_{t-1}, D_1, \dots, D_{t-1})$ for all $t \geq 2$, and that p_1 be a deterministic constant. The *expected risk-sensitive revenue* generated by a feasible pricing policy π over T time periods is defined as

$$R_T^\pi := \mathbb{E}^\pi \left[\sum_{t=1}^{T} r_\alpha(p_t) \right],$$

where expectation is taken with respect to the demand model (2.1) under the pricing policy π.

2.2.3 Performance Metric

We evaluate the performance of a feasible pricing policy π according to the T-period *regret*, which we define as

$$\Delta_T^\pi := R_T^* - R_T^\pi.$$

Naturally, pricing policies yielding a small regret are preferred, as the oracle risk-sensitive revenue R_T^* stands as an upper bound on the expected risk-sensitive revenue R_T^π achievable by any feasible pricing policy π. Ultimately, we seek a pricing policy whose T-period regret is sublinear in the horizon T. Such a pricing policy is said to have *no-regret*.

Definition 1 (No Regret Pricing)**.** A feasible pricing policy π is said to exhibit *no-regret* if $\lim_{T\to\infty} \Delta_T^\pi / T = 0$.

Implicit in the goal of designing a no-regret policy is that the sequence of prices that it generates should converge to the oracle optimal price sequence.

2.3 Demand Model Learning

Clearly, the ability to price with no-regret will rely centrally on the rate at which the unknown parameters, θ, and quantile function, $F^{-1}(\alpha)$, can be learned from the market data. In what follows, we describe a basic approach to learning the demand model using the method of least squares estimation.

2.3.1 Parameter Estimation

Given the history of past prices and demand observations $(p_1, \ldots, p_t, D_1, \ldots, D_t)$ through period t, define the *least squares estimator* (LSE) of θ as

$$\theta_t := \arg\min \left\{ \sum_{k=1}^{t} (D_k - \lambda(p_k, \vartheta))^2 \ : \ \vartheta \in \mathbb{R}^2 \right\},$$

for time periods $t = 1, 2, \ldots$. The LSE at period t admits an explicit expression of the form

$$\theta_t = \left(\sum_{k=1}^{t} \begin{bmatrix} p_k \\ 1 \end{bmatrix} \begin{bmatrix} p_k \\ 1 \end{bmatrix}^\top \right)^{-1} \left(\sum_{k=1}^{t} \begin{bmatrix} p_k \\ 1 \end{bmatrix} D_k \right), \tag{2.4}$$

provided the indicated inverse exists. It will be convenient to define the 2×2 matrix

$$\mathscr{J}_t := \sum_{k=1}^{t} \begin{bmatrix} p_k \\ 1 \end{bmatrix} \begin{bmatrix} p_k \\ 1 \end{bmatrix}^\top = \begin{bmatrix} \sum_{k=1}^{t} p_k^2 & \sum_{k=1}^{t} p_k \\ \sum_{k=1}^{t} p_k & t \end{bmatrix}.$$

Utilizing the definition of the aggregate demand model (2.1), in combination with the expression in (2.4), one can obtain the following expression for the parameter estimation error:

$$\theta_t - \theta = \mathscr{J}_t^{-1} \left(\sum_{k=1}^{t} \begin{bmatrix} p_k \\ 1 \end{bmatrix} \varepsilon_k \right). \tag{2.5}$$

Remark 2 (The Role of Price Dispersion)**.** The expression for the parameter estimation error in (2.5) reveals how consistency of the LSE is reliant upon the asymptotic spectrum of the matrix $\mathscr{J}_t$. Namely, the minimum eigenvalue of $\mathscr{J}_t$, must grow unbounded with time, in order that the parameter estimation error converge to zero in probability. In [50, Lemma 2], the authors establish a sufficient condition for such growth. Specifically, they prove that the minimum eigenvalue of $\mathscr{J}_t$ is bounded from below (up to a multiplicative constant) by the *sum of squared price deviations* defined as

$$J_t := \sum_{k=1}^{t} (p_k - \overline{p}_t)^2,$$

where $\overline{p}_t := (1/t) \sum_{k=1}^{t} p_k$. The result is reliant on the assumption that the underlying pricing policy π yields a bounded sequence of prices $\{p_t\}$. An important

consequence of such a result is that it reveals the explicit role that *price dispersion* (i.e., exploration) plays in facilitating consistent parameter estimation.

Finally, given the underlying assumption that the unknown model parameters θ belong to a compact set defined $\Theta := [\underline{a}, \overline{a}] \times [0, \overline{b}]$, one can improve upon the LSE at time t by projecting it onto the set Θ. Accordingly, we define the *truncated least squares estimator* as

$$\widehat{\theta}_t := \arg\min \{\|\vartheta - \theta_t\|_2 : \vartheta \in \Theta\}. \tag{2.6}$$

Clearly, we have that $\|\widehat{\theta}_t - \theta\|_2 \leq \|\theta_t - \theta\|_2$. In the following section, we describe an approach to estimating the underlying quantile function using the parameter estimator defined in (2.6).

2.3.2 Quantile Estimation

Building on the parameter estimator specified in Equation (2.6), we construct an estimator of the unknown quantile function $F^{-1}(\alpha)$ according to the empirical quantile function associated with the demand estimation residuals. Namely, in each period t, define the sequence of *residuals* associated with the estimator $\widehat{\theta}_t$ as

$$\widehat{\varepsilon}_{k,t} := D_k - \lambda(p_k, \widehat{\theta}_t),$$

for $k = 1, \ldots, t$. Define their *empirical distribution* as

$$\widehat{F}_t(x) := \frac{1}{t} \sum_{k=1}^{t} \mathbb{1}\{\widehat{\varepsilon}_{k,t} \leq x\},$$

and their corresponding *empirical quantile function* as $\widehat{F}_t^{-1}(\alpha) := \inf\{x \in \mathbb{R} : \widehat{F}_t(x) \geq \alpha\}$ for all $\alpha \in (0, 1)$. It will be useful in the sequel to express the empirical quantile function in terms of the order statistics associated with sequence of

residuals. Essentially, the *order statistics* $\widehat{\varepsilon}_{(1),t}, \ldots, \widehat{\varepsilon}_{(t),t}$ are defined as a permutation of $\widehat{\varepsilon}_{1,t}, \ldots, \widehat{\varepsilon}_{t,t}$ such that $\widehat{\varepsilon}_{(1),t} \leq \widehat{\varepsilon}_{(2),t} \leq \cdots \leq \widehat{\varepsilon}_{(t),t}$. With this concept in hand, the empirical quantile function can be equivalently expressed as

$$\widehat{F}_t^{-1}(\alpha) = \widehat{\varepsilon}_{(i),t}, \tag{2.7}$$

where the index i is chosen such that $\frac{i-1}{t} < \alpha \leq \frac{i}{t}$. It is not hard to see that $i = \lceil t\alpha \rceil$. Using Equation (2.7), one can relate the quantile estimation error to the parameter estimation error according to the following inequality

$$|\widehat{F}_t^{-1}(\alpha) - F^{-1}(\alpha)| \leq |F_t^{-1}(\alpha) - F^{-1}(\alpha)| + \left(1 + p_{(i),t}^2\right)^{1/2} \|\widehat{\theta}_t - \theta\|_2, \tag{2.8}$$

where $p_{(i),t}$ is defined as the ith order statistic of the sequence of prices $p_1, \ldots, p_t$. Here, F_t^{-1} is defined as the empirical quantile function associated with the sequence of demand shocks $\varepsilon_1, \ldots, \varepsilon_t$. Their empirical distribution is defined as

$$F_t(x) := \frac{1}{t} \sum_{k=1}^{t} \mathbb{1}\{\varepsilon_k \leq x\}.$$

The inequality in (2.8) reveals that consistency of the quantile estimator (2.7) is reliant upon consistency of the both the *parameter estimator* and the *empirical quantile function* defined in terms of the sequence of demand shocks. Consistency of the former is established in Lemma 1 under a suitable choice of a pricing policy, which we specify in Equation (2.11). Consistency of the latter is established in what follows under any feasible pricing policy. More specifically, as the empirical quantile function F_t^{-1} (unlike $\widehat{F}_t^{-1}$) only depends on the sequence of random shocks, the rate at which it converges to F^{-1} is not determined by the choice of the pricing policy.

Proposition 1. Let $\mu_1 := 2/(L^2 \log(2))$. It holds that

$$\mathbf{P}\left(|F_t^{-1}(\alpha) - F^{-1}(\alpha)| > \gamma\right) \leq 2\exp(-\mu_1 \gamma^2 t) \tag{2.9}$$

for all $\gamma > 0$ and $t \geq 2$.

Proposition 1 is similar in nature to [31, Lemma 2], which provides a bound on the rate at which the empirical distribution function converges to the true cumulative distribution function in probability. The combination of Assumption 1 with [31, Lemma 2] enables the derivation of the upper bound in Proposition 1.

2.4 Design of Pricing Policies

Building on the approach to demand model learning in Section 2.3, we construct a DR pricing policy, which is guaranteed to exhibit *no-regret*.

2.4.1 Myopic Policy

We begin with a description of a natural approach to pricing, which interleaves the model estimation scheme defined in Section 2.3 with a *myopic* approach to pricing. That is to say, at each stage $t+1$, the utility estimates the demand model parameters and quantile function according to (2.6) and (2.7), respectively, and sets the price according to

$$\widehat{p}_{t+1} = \frac{c_{t+1}}{2} - \frac{\widehat{b}_t + \widehat{F}_t^{-1}(\alpha)}{2\widehat{a}_t}. \tag{2.10}$$

Under this pricing policy, the utility essentially treats its model estimate in each period as if it is correct, and disregards the subsequent impact of its choice of price on its ability to accurately estimate the demand model in future time periods. A danger inherent to a myopic approach to pricing such as this is that the resulting price sequence may fail to elicit information from demand at a rate,

which is fast enough to enable consistent model estimation. As a result, the model estimates may converge to incorrect values. Such behavior is well documented in the literature [30, 50, 56], and is commonly referred to as *incomplete learning*. In Section 2.6, we provide a numerical example, which demonstrates the occurrence of incomplete learning under the myopic pricing policy (2.10).

2.4.2 Perturbed Myopic Policy

In order to prevent the possibility of incomplete learning, we propose a pricing policy that is guaranteed to elicit information from demand at a sufficient rate through carefully designed perturbations to the myopic pricing policy (2.10). The pricing policy we propose is defined as

$$p_{t+1} = \begin{cases} \widehat{p}_{t+1}, & t \text{ odd} \\ \widehat{p}_t + \frac{1}{2}(c_{t+1} - c_t) + \rho\delta_{t+1}, & t \text{ even}, \end{cases} \tag{2.11}$$

where $\rho \geq 0$ is a user specified positive constant, and

$$\delta_t := \text{sgn}\,(c_t - c_{t-1}) \cdot t^{-1/4}.$$

We refer to the policy (2.11) as the *perturbed myopic policy*.[5]

The perturbed myopic policy differs from the myopic policy in two important ways. First, the model parameter estimate, $\widehat{\theta}_t$, and quantile estimate, $\widehat{F}_t^{-1}(\alpha)$, are updated at every other time step. Second, to enforce sufficient price exploration, an offset is added to the myopic price at every other time step. Roughly speaking, the sequence of myopic price offsets $\{\rho\delta_t\}$ is chosen to decay at a rate, which is slow enough to ensure consistent model learning, but not so

[5]In defining the sign function, we require that $\text{sgn}(0) = 1$.

slow as to preclude a sub-linear growth rate for regret. In Section 2.5, we will show that the combination of these features is enough to ensure consistent parameter estimation and a sub-linear growth rate for the T-period regret, which is bounded from above by $O(\sqrt{T}\log(T))$.

Remark 3 (On the Perturbation Order)**.** We briefly describe the rationale behind the selection of the order of the perturbation sequence as $\delta_t = O(t^{-1/4})$. First, notice from Equation (2.12) that the regret incurred by any feasible pricing policy is equal to the sum of the squared pricing errors generated by the policy. Combining this expression with the upper bound on the absolute pricing error induced by the perturbed myopic policy in (2.14), it becomes clear to see the conflicting effects that the perturbation sequence has on regret. On the one hand, an increase in the order of the perturbation sequence will tend to reduce the growth rate of regret by increasing the rate at which the parameter estimation error $\|\widehat{\theta}_t - \theta\|_2$ converges to zero. On the other hand, an increase in the order of the perturbation sequence will tend to have the counterproductive effect of increasing the growth rate of regret by increasing the rate at which the deliberate pricing errors $\rho|\delta_t|$ accumulate. A tradeoff, therefore, emerges in selecting the order of the perturbation sequence. In Appendix A.2, we show that among all perturbation sequences that are polynomial in t, perturbation sequences of the order $O(t^{-1/4})$ are optimal in the sense of minimizing the asymptotic order of our upper bound on regret (ignoring logarithmic factors).

2.5 A Bound on Regret

Given the demand model considered in this chapter, one can express the T-period regret as

$$\Delta_T^{\pi} = a \sum_{t=1}^{T} \mathbb{E}^{\pi} \left[(p_t - p_t^*)^2 \right], \tag{2.12}$$

under any pricing policy π. It becomes apparent, upon examination of Equation (2.12), that the rate at which regret grows is directly proportional to the rate at which pricing errors accumulate. We, therefore, proceed in deriving a bound on the rate at which the absolute pricing error $|p_t - p_t^*|$ converges to zero in probability, under the perturbed myopic policy.

First, it is not difficult to show that, under the perturbed myopic policy (2.11), the absolute pricing error incurred in each even time period t is upper bounded by

$$|p_{t+1} - p_{t+1}^*| \leq \kappa_1 \|\widehat{\theta}_{t-1} - \theta\|_2 \; + \; \kappa_2 |\widehat{F}_{t-1}^{-1}(\alpha) - F^{-1}(\alpha)| \; + \; \rho |\delta_{t+1}|, \tag{2.13}$$

where $\kappa_1 := (\underline{a}^2 + (\overline{b} + \overline{\varepsilon})^2)^{1/2} / (2\underline{a}^2)$ and $\kappa_2 := 1/(2\underline{a})$. The pricing error incurred during odd time periods t is similarly bounded, sans the explicit dependency on the myopic price perturbation. The upper bound in (2.13) is intuitive as it consists of three terms: the parameter estimation error, the quantile estimation error, and the myopic price perturbation – each of which represents a rudimentary source of pricing error.

One can further refine the upper bound in (2.13), by leveraging on the fact that, under the perturbed myopic policy, the generated sequence of prices is uniformly bounded. That is to say, $|p_t| \leq \overline{p}$ for all time periods t, where

$$\overline{p} := \frac{1}{2} \max \left\{ \overline{c} - \frac{\underline{\varepsilon}}{\underline{a}} \, , \; \overline{c} - \frac{\underline{\varepsilon}}{\overline{a}} \, , \; \frac{\overline{b} + \overline{\varepsilon}}{\underline{a}} \right\}.$$

Combining this fact with the previously derived upper bound on the quantile estimation error in (2.8), we have that

$$|p_{t+1} - p^*_{t+1}| \leq \kappa_3 \|\widehat{\theta}_{t-1} - \theta\|_2 \ + \ \kappa_2 |F^{-1}_{t-1}(\alpha) - F^{-1}(\alpha)| \ + \ \rho|\delta_{t+1}|, \tag{2.14}$$

for even time periods t, where $\kappa_3 := \kappa_1 + \kappa_2(1 + \overline{p}^2)^{1/2}$.

Consistency of the perturbed myopic policy depends on the asymptotic behavior of each term in (2.14). The price perturbation converges to zero by construction, and consistency of the empirical quantile function is established in Proposition 1. The following Lemma establishes a bound on the mean squared parameter estimation error under the perturbed myopic policy (2.11).

Lemma 1 (Consistent Parameter Estimation)**.** There exists a finite positive constant μ_2 such that, under the perturbed myopic policy (2.11),

$$\mathbb{E}\left[\|\widehat{\theta}_t - \theta\|^2\right] \leq \frac{\mu_2}{\rho^2} \frac{\log(t)}{\sqrt{t}},$$

for all $t \geq 3$ and $\rho > 0$.

The following Theorem establishes an $O(\sqrt{T}\log(T))$ upper bound on the T-period regret.

Theorem 1 (Sub-linear Regret)**.** The T-period regret incurred by the perturbed myopic policy (2.11) satisfies

$$\Delta^{\pi}(T) \leq C_0 + C_1\sqrt{T}\log(T) + C_2\log(T), \tag{2.15}$$

for all $T \geq 3$. Here, C_0, C_1, and C_2 are finite positive constants.[6]

[6]We refer the reader to Equations (A.17) -(A.19) for the exact specification of the coefficients C_0, C_1, and C_2.

In the process of proving Theorem 1, we also show that the perturbed myopic policy generates a sequence of market prices $\{p_t\}$ that converges to the oracle optimal price sequence $\{p_t^*\}$ in the mean square sense. More formally, we have the following corollary.

Corollary 1 (Price Consistency)**.** The sequence of prices $\{p_t\}$ generated by the perturbed myopic policy (2.11) satisfies

$$\lim_{t\to\infty} \mathbb{E}\left[(p_t - p_t^*)^2\right] = 0,$$

where $\{p_t^*\}$ denotes the oracle optimal price sequence.

2.5.1 The Exploratory Effect of Wholesale Price Variation

Thus far in this chapter, we have made no assumption on the nature of variation in the sequence of wholesale electricity prices $\{w_t\}$. In particular, all of the previously stated results hold for any sequence of time-varying wholesale electricity prices. This includes the special case in which the wholesale price of electricity is constant across time, i.e., $w_t = w$ for all time periods t. It is, however, natural to inquire as to how the degree of variation in the sequence of wholesale prices might impact the performance of the pricing policies considered in this chapter.

First, it is straightforward to see from Equation (2.10) that variation in the sequence of wholesale prices induces equivalent variation in the sequence of myopic prices. Such variation in the myopic price sequence is most naturally interpreted as a form of *costless exploration*. In the following result, we establish a sufficient condition on the variation of wholesale prices, which eliminates the need for external perturbations to the myopic price sequence (i.e., setting

$\rho = 0$), while guaranteeing an upper bound on the resulting T-period regret that is $O(\log^2(T))$. The redundancy of exogenous exploration in the presence of sufficient variation in the sequence of observations generated by the environment is also independently discovered in the revenue management literature [9].

Theorem 2 (Logarithmic Regret)**.** Assume that there exists a finite positive constant $\sigma > 0$ such that

$$|w_t - w_{t-1}| \geq \sigma, \tag{2.16}$$

for all time periods t.[7] It follows that the T-period regret incurred by the perturbed myopic policy (2.11), with $\rho = 0$, satisfies

$$\Delta^{\pi}(T) \leq M_0 + \frac{M_2}{\sigma^2} + M_1 \log(T) + \frac{M_2}{\sigma^2} \log^2(T), \tag{2.17}$$

for all $T \geq 3$. Here, M_0, M_1, and M_2 are finite positive constants[8], which are independent of the parameter σ.

Several comments are in order. First, under the additional assumption of persistent wholesale price variation (2.16), we establish in Theorem 2 an improvement upon the original order of regret stated in Theorem 1 from $O(\sqrt{T}\log(T))$ to $O(\log^2(T))$. However, as one might expect, the magnitude of the upper bound on regret in (2.17) scales in a manner that is inversely proportional to σ^2. As a result, the upper bound on the T-period regret goes to infinity as σ goes to zero, and, therefore, provides little useful information when σ is small.

[7] Note that Assumption (2.16) in Theorem 2 implies that $|c_t - c_{t-1}| \geq \sigma$.

[8] We refer the reader to Equations (A.20) -(A.22) for the exact specification of the coefficients M_0, M_1, and M_2.

2.6 Simulations

We conduct a numerical analysis to compare the performance of the myopic policy (2.10) against the perturbed myopic policy (2.11) over a time horizon of $T = 10^4$. We set the tuning parameter $\rho = 0.19$. We consider the setting in which there are $N = 1000$ customers participating in the DR program. For each customer i, we select a_i uniformly at random from the interval $[0.04, 0.20]$, and independently select b_i according an exponential distribution (with mean equal to 0.01) truncated over interval $[0, 0.1]$. Parameters are drawn independently across customers.[9] For each customer i, we take the demand shock to be distributed according to a normal distribution with zero-mean and standard deviation equal to 0.04, truncated over the interval $[-0.4, 0.4]$. We consider a utility with risk sensitivity equal to $\alpha = 0.1$. In other words, the utility seeks to maximize the revenue it is guaranteed to receive with probability no less than 0.9. Finally, we set the retail price of electricity to $f = 0.17$ (\$/kWh), and set the wholesale price of electricity to $w_t = 1.67$ (\$/kWh) for all days t. Such values are consistent with the average residential retail and peak wholesale prices of electricity in the state of New York in 2016 [91, 71].

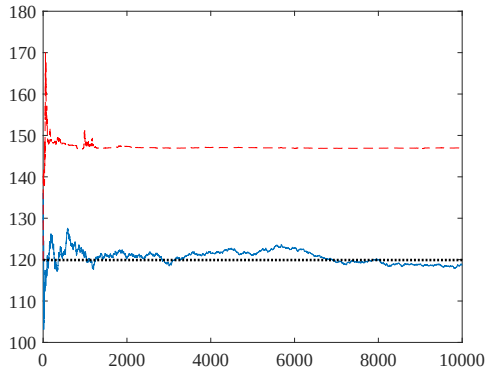
2.6.1 Discussion

Because the wholesale price of electricity is fixed over time, the parameter and quantile estimates represent the only source of variation in the sequence of prices generated by the myopic policy. Due to the combined structure of the

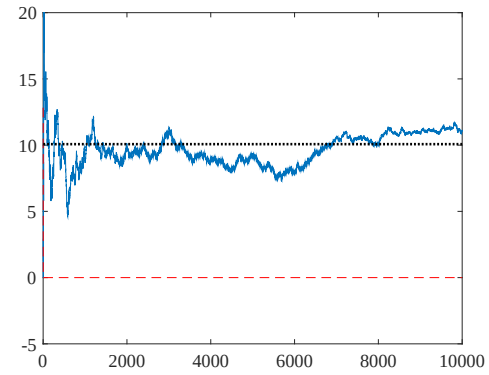
[9] It is worth noting that the range of parameter values $a_i \in [0.04, 0.20]$ considered in this numerical study is consistent with the range of demand price elasticities observed in several real-time pricing programs conducted in the United States [90, 34].

myopic policy and the least squares estimator, the value of each new demand observation rapidly diminishes over time, which, in turn, manifests in a rapid convergence of the sequence of prices generated under the myopic policy. The resulting lack of exploration in the sequence of myopic prices results in incomplete learning, which is seen in Figure 2.1. Namely, the sequence of myopic prices converges to a value, which substantially differs form the oracle optimal price. As a consequence, the myopic policy incurs a T-period regret that grows linearly with the horizon T, as is observed in Figure 2.2.

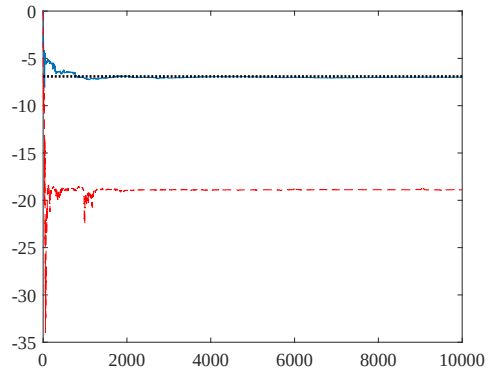
On the other hand, the sequence of perturbations $\{\rho\delta_t\}$ generate enough variation in the sequence of prices generated by the perturbed myopic policy to ensure consistent model estimation, as is seen in Figures 2.1a and 2.1b. This, in turn, results in convergence of the sequence of posted prices to the oracle optimal price. This, combined with the fact that the price offset $\rho\delta_t$ vanishes at a sufficiently fast rate, ensures sublinearity in the growth rate of the corresponding T-period regret, as is observed in Figure 2.2.



(a) Sequences of demand parameter $\widehat{a}_t$.

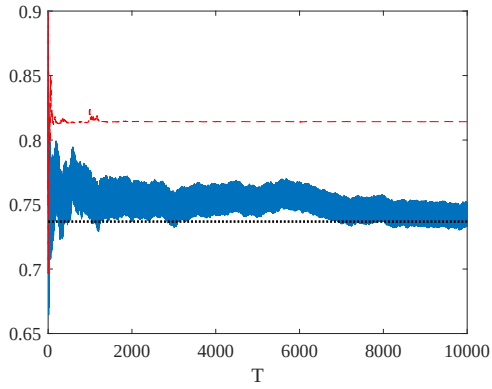


(b) Sequences of demand parameter $\widehat{b}_t$.

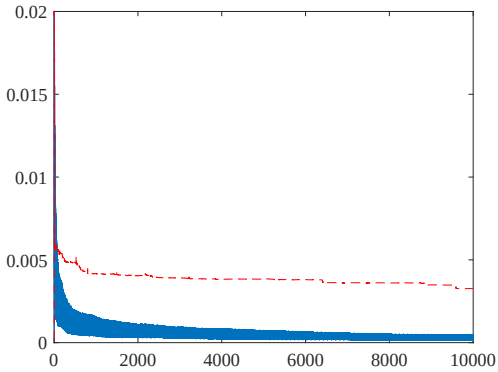


(c) Sequences of quantile function $\widehat{F}_t^{-1}(\alpha)$.

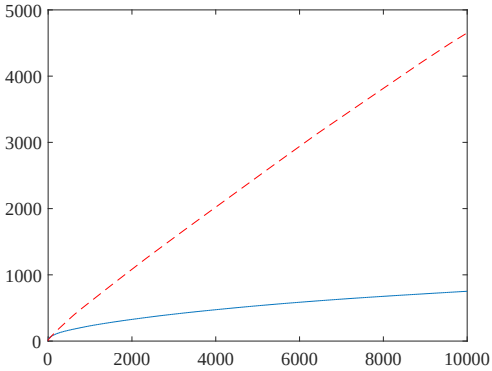
Figure 2.1: (a)-(b) Sample paths of the parameter estimates, and (c) sample path of the shock quantile estimates under the *myopic policy* (---), the *perturbed myopic policy* (——), and the *oracle policy* (·····).



(a) Sequences of posted prices.



(b) Mean squared pricing error.



(c) Regret.

Figure 2.2: (a) Sample path of posted prices, (b) mean squared pricing error, and (c) regret under the *myopic policy* (- - -), the *perturbed myopic policy* (——), and the *oracle policy* (·····).

CHAPTER 3

SAFE ONLINE LEARNING OF LINEAR MODELS

3.1 Introduction

We investigate the role of safety in constraining the design of learning algorithms within the classical framework of linear stochastic bandits [27, 73, 1]. Specifically, we introduce a family of *safe linear stochastic bandit problems* where—in addition to the typical goal of designing learning algorithms that minimize regret—we impose a constraint requiring that an algorithm's stagewise expected reward remains above a predetermined safety threshold with high probability at every stage of play. In the proposed framework, we assume that a "safe" baseline arm is initially known, and consider a class of safety thresholds that are defined as fixed cutbacks on the expected reward of the known baseline arm. Accordingly, an algorithm that is deemed to be safe cannot induce stagewise rewards that dip below the baseline reward by more than a fixed amount. Critically, the assumption of a known baseline arm—and the limited capacity for exploration implied by the class of safety thresholds considered—can be leveraged on to initially guide the exploration of allowable arms by playing combinations of the baseline arm and exploratory arms in a manner that expands the set of safe arms over time, while simultaneously preserving safety at every stage of play.

There are a variety of real-world applications that might benefit from the design of stagewise-safe online learning algorithms [52, 60, 81]. Most prominently, clinical trials have long been used as a motivating application for the multi-armed bandit [10] and linear bandit [27] frameworks. However, as pointed out

by [95]: "Despite this apparent near-perfect fit between a real-world problem and a mathematical theory, the MABP has yet to be applied to an actual clinical trial." One could argue that the ability to provide a learning algorithm that is guaranteed to be stagewise safe has the potential to facilitate the utilization of bandit models and algorithms in clinical trials. More concretely, consider the possibility of using the linear bandit framework to model the problem of optimizing a combination of d candidate treatments for a specific health issue. In this context, an "arm" represents a mixture of treatments, the "unknown reward vector" encodes the effectiveness of each treatment, and the "reward" represents a patient's response to a chosen mixture of treatments. In terms of the safety threshold, it is natural to select the "baseline arm" to be the (possibly suboptimal) combination of treatments possessing the largest reward known to date. As it is clearly unethical to prescribe a treatment that may degrade a patient's health, the stagewise safety constraint studied in this chapter can be interpreted as a requirement that a patient's response to a chosen treatment must be arbitrarily close to that of the baseline treatment, if not better.

3.1.1 Contributions

In this chapter, we propose a new learning algorithm that is tailored to the safe linear bandit framework. The proposed algorithm, which we call the *Safe Exploration and Greedy Exploitation* (SEGE) algorithm, is shown to exhibit near-optimal expected regret, while guaranteeing the satisfaction of the proposed safety constraint at every stage of play. Initially, the SEGE algorithm performs safe exploration by combining the baseline arm with a random exploratory arm that is constrained by an "exploration budget" implied by the stagewise safety con-

straint. Over time, the proposed algorithm systematically expands the family of safe arms in this manner to include new safe arms with expected rewards that exceed the baseline reward level. Exploitation under the SEGE algorithm is based on the certainty equivalence principle. That is, the algorithm constructs an "estimate" of the unknown reward parameter, and selects an arm that is optimal for the given parameter estimate. The SEGE algorithm only plays the certainty equivalent (i.e., greedy) arm when it is safe—a condition that is determined according to a lower confidence bound on its expected reward. Moreover, the proposed algorithm balances the trade-off between exploration and exploitation by controlling the rate at which information is accumulated over time, as measured by the growth rate of the minimum eigenvalue of the so-called information matrix.[1] More specifically, the SEGE algorithm guarantees that the minimum eigenvalue of the information matrix grows at a rate ensuring that the expected regret of the algorithm is no greater than $O(\sqrt{T}\log(T))$ after T stages of play. This regret rate is near optimal in light of $\Omega(\sqrt{T})$ lower bounds previously established in the linear stochastic bandit literature [27, 73].

3.1.2 Related Literature

There is an extensive literature on linear stochastic bandits. For this setting, several algorithms based on the principle of Optimism in the Face of Uncertainty (OFU) [27, 73, 1] or Thompson Sampling [3] have been proposed. Although such algorithms are known to be near-optimal under various measures of regret, they may fail in the safe linear bandit framework, as their (unconstrained)

[1] We note that a closely related class of learning algorithms, which explicitly control the rate of information gain in this manner, have been previously studied in the context of dynamic pricing algorithms for revenue maximization [30, 50].

approach to exploration may result in a violation of the stagewise safety constraints considered in this chapter.

In the context of multi-armed bandits, there is a related stream of literature that focuses on the design of "risk-sensitive" learning algorithms by encoding risk in the performance objectives according to which regret is measured [18, 28]. Typical risk measures that have been studied in the multi-armed bandit literature include Mean-Variance [77, 94], Value-at-Risk [93], and Conditional Value-at-Risk [36]. Although such risk-sensitive algorithms are inclined to exhibit reduced volatility in the cumulative reward that is received over time, they are not constrained in a manner that explicitly limits the stagewise risk of the reward processes that they induce.

Closer to the setting studied in this chapter is the conservative bandit framework [97, 49, 38], which incorporates explicit safety constraints on the reward process induced by the learning algorithm. However, in contrast to the stagewise safety constraints considered in this chapter, conservative bandits encode their safety requirements in the form of constraints on the cumulative rewards received by the algorithm. Along a similar line of research, [82] investigate the design of learning algorithms for risk-constrained contextual bandits that balance a tradeoff between cumulative constraint violation and regret. Given the cumulative nature of the safety constraints considered by the aforementioned algorithms, they cannot be directly applied to the stagewise safe linear bandit problem considered in this chapter. In Section 3.6.3, we provide a simulation-based comparison between the SEGE algorithm and the Conservative Linear Upper Confidence Bound (CLUCB) algorithm [49] to more clearly illustrate the potential weaknesses and strengths of each approach.

We close this section by mentioning another closely related body of work in the online learning literature that investigates the design of stagewise-safe algorithms for a more general class of smooth reward functions [81, 80, 92]. Although the proposed algorithms are shown to respect stagewise safety constraints that are similar in spirit to the class of safety constraints considered in this chapter, they lack formal upper bounds on their cumulative regret.

3.1.3 Organization

The remainder of the chapter is organized as follows. We introduce pertinent notation in Section 3.2. In Section 3.3, we define the safe linear stochastic bandit problem. In Section 3.4, we introduce the Safe Exploration and Greedy Exploitation (SEGE) algorithm. We present our main theoretical findings in Section 3.5, and close the chapter with a simulation study of the SEGE algorithm in Section 3.6. All mathematical proofs are presented in the Appendix Chapter B.

3.2 Notation

We denote the standard Euclidean norm of a vector $x \in \mathbb{R}^d$ by $\|x\|$ and define its weighted Euclidean norm as $\|x\|_S = \sqrt{x^\top S x}$ where $S \in \mathbb{R}^{d\times d}$ is a given symmetric positive semidefinite matrix. We denote the inner product of two vectors $x, y \in \mathbb{R}^d$ by $\langle x, y\rangle = x^\top y$. For a square matrix $A \in \mathbb{R}^{d\times d}$, we denote its minimum and maximum eigenvalues by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively.

3.3 Problem Formulation

In this section, we introduce the safe linear stochastic bandit model considered in this chapter. Before doing so, we review the standard model for linear stochastic bandits on which our formulation is based.

3.3.1 Linear Bandit Model

Linear stochastic bandits belong to a class of sequential decision-making problems in which a learner (i.e., decision-maker) seeks to maximize an unknown linear function using noisy observations of its function values that it collects over multiple stages. More precisely, at each stage $t = 1, 2, \ldots$, the learner is required to select an arm (i.e., action) X_t from a compact set $\mathcal{X} \subset \mathbb{R}^d$ of allowable arms, which is assumed to be an ellipsoid of the form

$$\mathcal{X} = \left\{ x \in \mathbb{R}^d \mid (x - \bar{x})^\top H^{-1} (x - \bar{x}) \leq 1 \right\}, \tag{3.1}$$

where $\bar{x} \in \mathbb{R}^d$ and $H \in \mathbb{R}^{d \times d}$ is a symmetric and positive definite matrix. In response to the particular arm played at each stage t, the learner observes a reward Y_t that is induced by the stochastic linear relationship:

$$Y_t = \langle X_t, \theta^* \rangle + \eta_t. \tag{3.2}$$

Here, the noise process $\{\eta_t\}_{t=1}^{\infty}$ is assumed be a sequence of independent and zero-mean random variables, and, critically, the reward parameter $\theta^* \in \mathbb{R}^d$ is assumed to be fixed and unknown. This a priori uncertainty in the reward parameter gives rise to the need to balance the exploration-exploitation trade-off in adaptively guiding the sequence of arms played in order to maximize the expected reward accumulated over time.

Admissible Policies and Regret.

We restrict the learner's decisions to those which are causal in nature. That is to say, at each stage t, the learner is required to select an arm based only on the history of past observations $H_t = (X_1, Y_1, \ldots, X_{t-1}, Y_{t-1})$, and on an external source of randomness encoded by a random variable U_t. The random process $\{U_t\}_{t=1}^{\infty}$ is assumed to be independent across time, and independent of the random noise process $\{\eta_t\}_{t=1}^{\infty}$. Formally, an *admissible policy* is a sequence of functions $\pi = \{\pi_t\}_{t=1}^{\infty}$, where each function π_t maps the information available to the learner at each stage t to a feasible arm $X_t \in \mathcal{X}$ according to $X_t = \pi_t(H_t, U_t)$.

The performance of an admissible policy after T stages of play is measured according to its *expected regret*,[2] which equals the difference between the expected reward accumulated by the optimal arm and the expected reward accumulated by the given policy after T stages of play. Formally, the expected regret of an admissible policy is defined as

$$R_T = \sum_{t=1}^{T} \langle X^*, \theta^* \rangle - \mathbb{E}\left[\sum_{t=1}^{T} \langle X_t, \theta^* \rangle \right], \tag{3.3}$$

where expectation is taken with respect to the distribution induced by the underling policy, and $X^* \in \mathcal{X}$ denotes the *optimal arm* that maximizes the expected reward at each stage of play given knowledge of the reward parameter θ^*, i.e.,

$$X^* = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \langle x, \theta^* \rangle. \tag{3.4}$$

At a minimum, we seek policies exhibiting an expected regret that is sublinear in the number of stages played T. Such policies are said to have *no-regret* in the sense that $\lim_{T \to \infty} R_T/T = 0$. To facilitate the design and theoretical analysis of such policies, we adopt a number of technical assumptions, which are

[2] It is worth noting, that in the context of linear stochastic bandits, *expected regret* is equivalent to *expected pseudo-regret* due to the additive nature of the noise process [1].

standard in the literature on linear stochastic bandits, and are assumed to hold throughout the chapter.

Assumption 2. The unknown reward parameter is bounded according to $\|\theta^*\| \leq S$, where $S > 0$ is a known constant.

Assumption 2 will prove essential to the design of policies that *safely explore* the parameter space in a manner ensuring that the expected reward stays above a predetermined (safe) threshold with high probability at each stage of play. We refer the reader to Definition 2 for a formal definition of the particular safety notion considered in this chapter.

Assumption 3. Each element of $\{\eta_t\}_{t=1}^{\infty}$ is assumed to be σ_η-sub-Gaussian, where $\sigma_\eta \geq 0$ is a fixed constant. That is,

$$\mathbb{E}\left[\exp(\gamma\eta_t)\right] \leq \exp\left(\gamma^2\sigma_\eta^2/2\right)$$

for all $\gamma \in \mathbb{R}$ and $t \geq 1$.

Assumptions 2 and 3, together with the class of admissible policies considered in this chapter, enable the utilization of existing results that provide an explicit characterization of confidence ellipsoids for the unknown reward parameter based on a ℓ_2-regularized least-squares estimator [1]. Such confidence regions play a central role in the design of no-regret algorithms for the linear stochastic bandits [27, 73, 1].

3.3.2 Safe Linear Bandit Model

In what follows, we introduce the framework of *safe linear stochastic bandits* studied in this chapter. Loosely speaking, an admissible policy is said to be *safe* if

the expected reward $\mathbb{E}[Y_t \mid X_t] = \langle X_t, \theta^* \rangle$ that it induces at each stage t is guaranteed to stay above a given reward threshold with high probability.[3] More formally, we have the following definition.

Definition 2 (Stagewise Safety Constraint)**.** Let $b \in \mathbb{R}$ and $\delta \in [0, 1]$. An admissible policy π—or equivalently the arm X_t that it induces—is defined to be (δ, b)-*safe at stage* t if

$$\mathbf{P}\left(\langle X_t, \theta^* \rangle \geq b\right) \geq 1 - \delta, \tag{3.5}$$

where the probability is calculated according to the distribution induced by the policy π.

The stagewise safety constraint requires that the expected reward at stage t exceed the *safety threshold* $b \in \mathbb{R}$ with probability no less than $1 - \delta$, where $\delta \in [0, 1]$ encodes the *maximum allowable risk* that the learner is willing to tolerate.

Clearly, without making additional assumptions, it is not possible to design policies that are guaranteed to be safe according to (3.5) given arbitrary safety specifications. We circumvent this obvious limitation by giving the learner access to a *baseline arm* with a known lower bound on its expected reward. We formalize this assumption as follows.

Assumption 4 (Baseline Arm)**.** We assume that the learner knows a deterministic baseline arm $X_0 \in \mathcal{X}$ satisfying

$$\langle X_0, \theta^* \rangle \geq b_0,$$

where $b_0 \in \mathbb{R}$ is a known lower bound on its expected reward.

[3] To simplify the exposition, we will frequently refer to $\mathbb{E}[Y_t \mid X_t]$—the expected reward conditioned on the arm X_t—as the *expected reward*, unless it is otherwise unclear from the context.

We note that it is straightforward to construct a baseline arm satisfying Assumption 4 by leveraging on the assumed boundedness of the unknown reward parameter as specified by Assumption 2. In particular, any arm $X_0 \in \mathcal{X}$ and its corresponding "worst-case" reward given by $b_0 = \min_{\|\theta\| \leq S} \langle X_0, \theta \rangle = -S\|X_0\|$ are guaranteed to satisfy Assumption 4.

With Assumption 4 in hand, the learner can leverage on the baseline arm to initially guide its exploration of allowable arms by playing combinations of the baseline arm and carefully designed exploratory arms in a manner that safely expands the set of safe arms over time. Plainly, the ability to safely explore in the vicinity of the baseline arm is only possible under stagewise safety constraints defined in terms of safety thresholds satisfying $b < b_0$. Under such stagewise safety constraints, the difference in rewards levels $b_0 - b$ can be interpreted as a stagewise "exploration budget" of sorts, as it reflects the maximum relative loss in expected reward that the learner is willing to tolerate when playing arms that deviate from the baseline arm. Naturally, the larger the exploration budget, the more aggressively can the learner explore. With the aim of designing safe learning algorithms that leverage on this simple idea, we will restrict our attention to stagewise safety constraints that are specified in terms of safety thresholds satisfying $b < b_0$.

Before proceeding, we briefly summarize the framework of *safe linear stochastic bandits* considered in this chapter. Given a baseline arm satisfying Assumption 4, the learner is initially required to fix a safety threshold that satisfies $b < b_0$. At each subsequent stage $t = 1, 2, \ldots$, the learner must select a risk level $\delta_t \in [0, 1]$ and a corresponding arm $X_t \in \mathcal{X}$ that is (δ_t, b)-safe. The learner aims to design an admissible policy that minimizes its expected regret, while

simultaneously ensuring that all arms played satisfy the stagewise safety constraints. In the following section, we propose a policy that is guaranteed to both exhibit no-regret and satisfy the safety constraint at every stage of play.

Relationship to Conservative Bandits.

We briefly discuss the relationship between the safety constraints considered in this chapter and the conservative bandit framework orginally studied by [97] in the context of multi-armed bandits, and subsequently extended to the setting of linear bandits by [49]. In contrast to the stagewise safety constraints considered in this chapter, conservative bandits encode their safety requirements in the form of constraints on the cumulative expected rewards received by a policy. Specifically, given a baseline arm satisfying Assumption 4, an admissible policy is said to respect the safety constraint defined in [49] if

$$\mathbf{P}\left(\sum_{k=1}^{t}\langle X_k, \theta^*\rangle \geq (1-\alpha)\sum_{k=1}^{t} b_0,\ \forall\, t \geq 1\right) \geq 1-\delta, \tag{3.6}$$

where $\delta \in [0,1]$ and $\alpha \in (0,1)$. Here, the parameter α encodes the maximum fraction of the cumulative baseline rewards that the learner is willing to forgo over time. In this context, smaller values of α imply greater levels of conservatism (safety). It is straightforward to show that conservative performance constraints of the form (3.6) are a special case of the class of stagewise safety constraints considered in Definition 2. In particular, if we set the safety threshold according to $b = (1-\alpha)b_0$, and let $\{\delta_t\}_{t=1}^{\infty}$ be any summable sequence of risk levels satisfying $\sum_{t=1}^{\infty}\delta_t \leq \delta$, then any admissible policy that is (δ_t, b)-safe for each stage $t \geq 1$ also satisfies the conservative performance constraint (3.6).

3.4 A Safe Linear Bandit Algorithm

In this section, we propose a new algorithm, which we call the *Safe Exploration and Greedy Exploitation* (SEGE) algorithm, that is guaranteed to be safe in every stage of play, while exhibiting a near-optimal expected regret. Before proceeding with a detailed description of the proposed algorithm, we briefly summarize the basic elements underpinning its design. Initially, the SEGE algorithm performs safe exploration by playing convex combinations of the baseline arm and random exploratory arms in a manner that satisfies Definition 2. Through this process of exploration, the SEGE algorithm is able to expand the family of safe arms to incorporate new arms that are guaranteed to outperform the baseline arm with high probability. Among all safe arms available to the algorithm at any given stage of play, the arm with the largest lower confidence bound on its expected reward is used as the basis for safe exploration. The SEGE algorithm performs exploitation by playing the certainty equivalent (greedy) arm based on a ℓ_2-regularized least-squares estimate of the unknown reward parameter. The SEGE algorithm only plays the greedy arm when it is safe, i.e., when a lower confidence bound on its expected reward exceeds the given safety threshold. Critically, the proposed algorithm balances the trade-off between exploration and exploitation by explicitly controlling the growth rate of the so-called information matrix (cf. Eq. (3.8)) in a manner that ensures that the expected regret of the SEGE algorithm is no greater than $O(\sqrt{T}\log(T))$ after T stages of play. The pseudocode for the SEGE algorithm is presented in Algorithm 1.

In the following section, we introduce a regularized least-squares estimator that will serve as the foundation for the proposed learning algorithm.

3.4.1 Regularized Least Squares Estimator

The ℓ_2-regularized least-squares estimate of the unknown reward parameter θ^* based on the information available to the algorithm up until and including stage t is defined as

$$\widehat{\theta}_t = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \sum_{k=1}^{t} (Y_k - \langle X_k, \theta \rangle)^2 + \lambda \|\theta\|^2 \right\}.$$

Here, $\lambda > 0$ denotes a user-specified regularization parameter. It is straightforward to show that

$$\widehat{\theta}_t = V_t^{-1} \sum_{k=1}^{t} X_k Y_k, \tag{3.7}$$

where

$$V_t = \lambda I + \sum_{k=1}^{t} X_k X_k^\top. \tag{3.8}$$

Throughout the chapter, we will frequently refer to the matrix V_t as the *information matrix* at each stage t.

The following result taken from [1, Theorem 2] provides an ellipsoidal characterization of a confidence region for the unknown reward parameter based on the regularized least-squares estimator (3.7). It is straightforward to verify that the conditions of [1, Theorem 2] are satisfied under the standing assumptions of this chapter.

Theorem 3. For any admissible policy and $\delta \in (0, 1)$, it holds that

$$\mathbf{P}\left(\theta^* \in \mathcal{C}_t(\delta),\ \forall t \geq 1\right) \geq 1 - \delta,$$

where the confidence set $\mathcal{C}_t(\delta)$ is defined as

$$\mathcal{C}_t(\delta) = \left\{ \theta \in \mathbb{R}^d \ :\ \|\widehat{\theta}_t - \theta\|_{V_t} \leq r_t(\delta) \right\}. \tag{3.9}$$

Here, $r_t(\delta)$ is defined as

$$r_t(\delta) = \sigma_\eta \sqrt{d \log\left(\frac{1 + tL^2/\lambda}{\delta}\right)} + \sqrt{\lambda} S, \tag{3.10}$$

where $L = \max_{x \in \mathcal{X}} \|x\|$.

In the following section, we propose a method for safe exploration using the characterization of the confidence ellipsoids introduced in Theorem 3.

3.4.2 Safe Exploration

We now describe a novel approach to "safe exploration" that will be utilized in the design of the proposed learning algorithm. At each stage $t \geq 1$, given a risk level δ_t, the SEGE algorithm constructs a safe exploration arm (X_t^{SE}) as a convex combination of a (δ_t, b_0)-safe arm (X_t^{S}) and a random exploratory arm (U_t), i.e.,

$$X_t^{\mathsf{SE}} = (1 - \rho) X_t^{\mathsf{S}} + \rho U_t. \tag{3.11}$$

Qualitatively, the user-specified parameter $\rho \in (0, 1)$ controls the balance between safety and exploration. Figure 3.1 provides a graphical illustration of the set of all safe exploration arms induced by a given safe arm X_t^{S} according to (3.11).

The random exploratory arm process $\{U_t\}_{t=1}^{\infty}$ is generated according to

$$U_t = \bar{x} + H^{1/2} \zeta_t, \tag{3.12}$$

where the random process $\{\zeta_t\}_{t=1}^{\infty}$ is assumed to be a sequence of independent, zero-mean, and symmetric random vectors. For each element of the sequence, we require that $\|\zeta_t\| = 1$ almost surely and $\sigma_\zeta^2 = \lambda_{\min}(\mathrm{Cov}\,(\zeta_t)) > 0$. Additionally, we define $\sigma^2 = \lambda_{\min}(\mathrm{Cov}\,(U_t))$. The parameters σ and ρ both determine

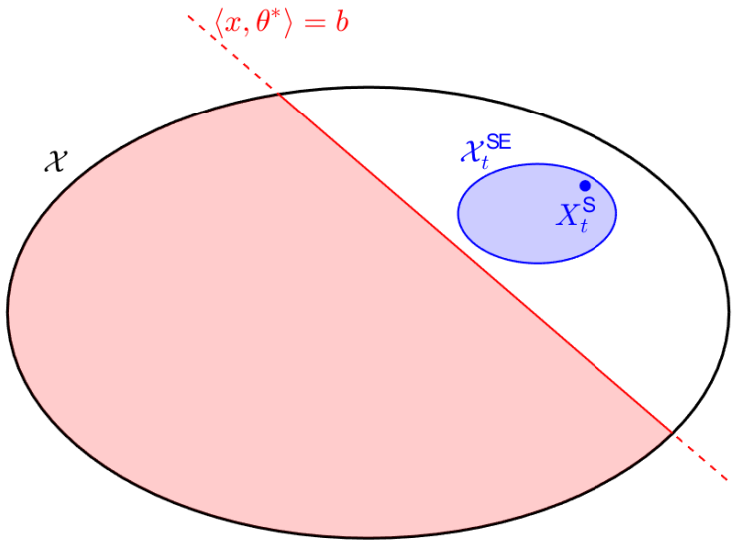


Figure 3.1: The figure illustrates the effect of the safety constraint on the learner's decision making ability. The shaded blue ellipse $\mathcal{X}_t^{\mathsf{SE}}$ depicts the set of all safe exploration arms constructed using the safe arm X_t^{S} under the SEGE algorithm, i.e., $\mathcal{X}_t^{\mathsf{SE}} = \{(1-\rho)X_t^{\mathsf{S}} + \rho x \mid \rho \in (0, \bar{\rho}], x \in \partial\mathcal{X}\}$. The red shaded area depicts the set of unsafe arms. The black ellipse (and its interior) depicts the entire set of allowable arms.

how aggressively the algorithm can explore the set of allowable arms. However, exploration that is too aggressive may result in a violation of the stagewise safety constraint. In the following Lemma, we establish an upper bound on ρ such that for all choices of $\rho \in (0, \bar{\rho}]$, the arm X_t^{SE} is guaranteed to be safe for any $\sigma \geq 0$.

Lemma 2. Let $\rho \in (0, \bar{\rho}]$ where $\bar{\rho} > 0$ is defined as

$$\bar{\rho} = \min \left\{ 1, \frac{b_0 - b}{2S\sqrt{\lambda_{\max}(H)}} \right\}. \tag{3.13}$$

Then, for every stage $t \geq 1$, the safe exploration arm X_t^{SE} defined in Equation (3.11) is (δ, b)-safe for any $\delta \in [0, 1]$.

As the SEGE algorithm expands its set of safe arms over time, it attempts to increase the stagewise efficiency with which it safely explores by exploring in

the vicinity of the safe arm with the largest lower confidence bound on its expected reward. More specifically, at each stage t, the SEGE algorithm constructs a confidence set $\mathcal{C}_{t-1}(\delta_t)$ according to Equation (3.9). With this confidence set in hand, the proposed algorithm calculates a lower confidence bound (LCB) on the expected reward of each arm $x \in \mathcal{X}$ according to

$$\text{LCB}_t(x) = \min_{\theta \in \mathcal{C}_{t-1}(\delta_t)} \langle x, \theta \rangle.$$

It is straightforward to show that the lower confidence bound defined above admits the closed-form expression:

$$\text{LCB}_t(x) = \langle x, \widehat{\theta}_{t-1} \rangle - r_t(\delta_t) \|x\|_{V_{t-1}^{-1}}.$$

We define the LCB arm (X_t^{LCB}) to be the arm with the largest lower confidence bound on its expected reward among all allowable arms. It is given by:

$$X_t^{\mathsf{LCB}} = \operatorname*{argmax}_{x \in \mathcal{X}} \ \text{LCB}_t(x). \tag{3.14}$$

Clearly, the LCB arm is guaranteed to be (δ_t, b_0)-safe if $\text{LCB}_t(X_t^{\mathsf{LCB}}) \geq b_0$. In this case, the SEGE algorithm relies on the LCB arm for safe exploration, as its expected reward is *potentially* superior to the baseline arm's expected reward.[4] Putting everything together, the SEGE algorithm sets the safe arm (X_t^{S}) at each stage t according to:

$$X_t^{\mathsf{S}} = \begin{cases} X_t^{\mathsf{LCB}}, & \text{if } \ \text{LCB}_t(X_t^{\mathsf{LCB}}) \geq b_0, \\ X_0, & \text{otherwise}. \end{cases} \tag{3.15}$$

Before closing this section, it is important to note that the LCB arm (3.14) can be calculated in polynomial time by solving a second-order cone program.

[4] It is important to note that the condition $\text{LCB}_t(X_t^{\mathsf{LCB}}) \geq b_0$ does not guarantee superiority of the LCB arm to the baseline arm, as b_0 is only assumed to be a lower bound on the baseline arm's expected reward.

This is in stark contrast to the non-convex optimization problem that needs to be solved when computing the UCB arm (i.e., the arm with the largest upper confidence bound on the expected reward)—a problem that has been shown to be NP-hard in general [27].

3.4.3 Safe Greedy Exploitation

We now describe a novel approach to "safe exploration" that will be utilized in the design of the proposed learning algorithm. Exploitation under the SEGE algorithm relies on the certainty equivalence principle. That is, the algorithm first estimates the unknown reward parameter according to Equation (3.7). Then, the algorithm chooses an arm that is optimal for the given parameter estimate. Given the ellipsoidal structure of the set of allowable arms, the optimal arm X^* can be calculated as

$$X^* = \bar{x} + \frac{H\theta^*}{\|\theta^*\|_H}. \tag{3.16}$$

Similarly, the certainty equivalent (greedy) arm can be calculated as

$$X_t^{\mathsf{CE}} = \bar{x} + \frac{H\widehat{\theta}_{t-1}}{\|\widehat{\theta}_{t-1}\|_H}, \tag{3.17}$$

where $\widehat{\theta}_{t-1}$ is the regularized least-squares estimate of the unknown reward parameter, as defined in Equation (3.7).

It is important to note that the SEGE algorithm only plays the greedy arm (3.17) when the lower confidence bound on its expected reward is greater than or equal to the safety threshold b. This ensures that the greedy arm is only played when it is safe.

Algorithm 1 SEGE Algorithm

Input: X_0, b_0, $\mathcal{X}$, $S > 0$, $c > 0$, $\nu \in (0,1)$, $\lambda > 0$, $b < b_0$, $\rho \in (0, \bar{\rho}]$, $\delta_t \in [0,1]\ \forall t \geq 1$
for $t = 1, 2, 3, \ldots$ **do**
{*Parameter Estimation*}
    Set $\widehat{\theta}_{t-1}$ according to Eq. (3.7)
    Set $\mathcal{C}_{t-1}(\delta_t)$ according to Eq. (3.9)
{*Safe Greedy Exploitation*}
    if $\text{LCB}_t(X_t^{\mathsf{CE}}) \geq b$ **and** $\lambda_{\min}(V_{t-1}) \geq c(t-1)^{\nu}$
        Set $X_t = X_t^{\mathsf{CE}}$ according to Eq. (3.17)
{*Safe Exploration*}
    else
        Set $X_t = X_t^{\mathsf{SE}}$ according to Eq. (3.11)
    end if
    Observe $Y_t = \langle X_t, \theta^* \rangle + \eta_t$
end for

3.5 Theoretical Results

We now present our main theoretical results showing that the SEGE algorithm exhibits near optimal regret for a large class of risk levels (cf. Theorem 5), in addition to being safe at every stage of play (cf. Theorem 4). As an immediate corollary to Theorem 5, we establish sufficient conditions under which the SEGE algorithm is also guaranteed to satisfy the conservative bandit constraint (3.6), while preserving the upper bound on regret in Theorem 5 (cf. Corollary 2).

Theorem 4 (Stagewise Safety Guarantee)**.** The SEGE algorithm is (δ_t, b)-safe at each stage, i.e.,

$$\mathbf{P}\left(\langle X_t, \theta^* \rangle \geq b\right) \geq 1 - \delta_t$$

for all $t \geq 1$.

The ability to enforce safety in the sequence of arms played is not surprising given the assumption of a known baseline arm that is guaranteed to be safe

at the outset. However, given the potential suboptimality of the baseline arm, a naïve policy that plays the baseline arm at every stage will likely incur an expected regret that grows linearly with the number of stages played T. In constrast, we show, in Theorem 5, that the SEGE algorithm for an appropriate choice of parameters c and ν exhibits an expected regret that is no greater than $O(\sqrt{T}\log(T))$ after T stages—a regret rate that is near optimal given existing $\Omega(\sqrt{T})$ lower bounds on regret [27, 73].

Theorem 5 (Upper Bound on Expected Regret)**.** Fix $\overline{\delta} \in (0, 1]$, $\overline{\nu} \in (0, 1)$, and $K \geq 0$. Let $\{\delta_t\}_{t=1}^{\infty}$ be any sequence of risk levels satisfying

$$\delta_t \geq \overline{\delta} e^{-Kt^{\overline{\nu}}} \quad \text{for all } t \geq 1. \tag{3.18}$$

Fix $c > 2dKL^2\sigma_\eta^2/(b_0 - b)^2$ and $\nu \in [\overline{\nu}, 1)$. Then, there exist finite positive constants C_1 and C_2 such that the expected regret under the SEGE algorithm is upper bounded as

$$R_T \leq C_1 T^{\nu} + C_2 \log(T) T^{1-\nu} \tag{3.19}$$

for all $T \geq 1$.

In what follows, we provide a high-level sketch of the proof of Theorem 5. The complete proof is presented in Appendix B.3. We bound the expected regret incurred during the safe exploration and the greedy exploitation stages separately. First, we show that the stagewise expected regret incurred when playing the greedy arm is proportional to the mean squared parameter estimation error. We then employ Theorem 3 to show that, conditioned on the event $\{\lambda_{\min}(V_t) \geq ct^{\nu}\}$, the mean squared parameter estimation error at each stage t is no greater than $O(\log(t)/t^{\nu})$. It follows that the cumulative expected regret incurred during the exploitation stages is no more than $O(\log(T)T^{1-\nu})$ after T

stages of play. Now, in order to upper bound the expected regret accumulated during the safe exploration stages, it suffices to upper bound the expected number of safe exploration stages, since the stagewise regret can be upper bounded by a finite constant under any admissible policy. We show that the expected number of safe exploration stages is no more than $O(T^{\nu})$ after T stages of play for any sequence of risk levels that does not decay faster than the rate specified in (3.18).

Several comments are in order. In Theorem 5, we establish a sufficient condition on the sequence of risk levels for which we can guarantee a near-optimal expected regret. Qualitatively, in order to satisfy more stringent safety constraints in subsequent stages, the learner needs to gain more information during preceding stages. More precisely, the information gain, measured by the minimum eigenvalue of the information matrix, controls the structure of the uncertainty ellipsoid, which, in turn, determines the lower confidence bound on the reward of each arm. That is to say, the LCB of the reward of an arm is an increasing function of the information gain. The sub-exponential condition (3.18) establishes a limit on the decay rate of the sequence of risk levels for which the learner can safely and near-optimally balance the trade-off between exploration and exploitation. Establishing a necessary condition on the rate at which the sequence of risk levels can decay to ensure near-optimal regret is an interesting direction for future research.

For sub-exponentially decaying sequence of risk levels according to condition (3.18) with $\overline{\nu} \leq 1/2$, the SEGE algorithm is guaranteed to exhibit a near-optimal expected regret for appropriate choices of parameters c and ν. More

precisely, for parameters $\nu = 1/2$ and c that satisfies

$$c > \frac{2dKL^2\sigma_\eta^2}{(b_0 - b)^2}, \tag{3.20}$$

the expected regret under SEGE algorithm is upper bounded by $O(\sqrt{T}\log(T))$ using Theorem 5. The choice of $\nu = 1/2$ ensures that $\lambda_{\min}(V_t) = O(\sqrt{t})$, which, in turn, balances the trade-off between exploration and exploitation near-optimally. It is worth noting that the lower bound on parameter c in (3.20) is proportional to d and σ_η^2. That is, there is a need to gain more information (i.e., ensure larger $\lambda_{\min}(V_t)$) in higher dimensional problems or environments with larger noise variances. Moreover, the lower bound (3.20) is proportional to risk level parameter K and inversely proportional to the exploration budget $(b_0 - b)^2$. Hence, the learner needs to gain more information when facing more strict safety constraints.

We close this section with a result establishing sufficient conditions under which the SEGE algorithm is guaranteed to satisfy the conservative performance constraint (3.6), in addition to being stagewise safe, while satisfying an upper bound on its expected regret that matches that of the CLUCB algorithm [49, Theorem 5]. Corollary 2 is stated without proof, as it is an immediate consequence of Theorems 4 and 5.

Corollary 2 (Conservative Performance Guarantee)**.** Let $\delta \in (0, 1)$. Assume, in addition to the standing assumptions of Theorem 5, that $\{\delta_t\}_{t=1}^{\infty}$ is a summable sequence satisfying $\sum_{t=1}^{\infty} \delta_t \leq \delta$. Then, the SEGE algorithm satisfies the conservative performance constraint (3.6), and exhibits an expected regret that is upper bounded by $O(\sqrt{T}\log(T))$ for all $T \geq 1$.

3.6 Simulation Results

In this section, we conduct a simple numerical study to illustrate the qualitative features of the SEGE algorithm and compare it with the CLUCB algorithm introduced by [49].

3.6.1 Simulation Setup

Model Parameters.

We consider a linear bandit with a two-dimensional input space ($d = 2$), and restrict the set of allowable arms $\mathcal{X}$ to be closed disk of radius $r = 1$ centered at $\bar{x} = (1, 1)$. The true reward parameter is taken to be $\theta^* = (0.6, 0.8)$, and the upper bound on its norm is set to $S = 1$. We select a baseline arm at random from the set of allowable arms as $X_0 = (1.2, 1.9)$, and set the baseline expected reward to $b_0 = \langle X_0, \theta^* \rangle = 2.24$. We set the safety threshold to $b = 0.8 \times b_0$. The observation noise process $\{\eta_t\}_{t=1}^{\infty}$ is assumed to be an IID sequence of zero-mean Normal random variables with standard deviation $\sigma_\eta = 1$.

SEGE Algorithm.

We set the parameters of the SEGE algorithm to $c = 0.5$, $\lambda = 0.1$, and $\rho = \bar{\rho} = 0.224$. We generate the random exploration process according to $U_t = \bar{x} + \zeta_t$, where $\{\zeta_t\}_{t=1}^{\infty}$ is a sequence of IID random variables that are uniformly distributed on the unit circle. To enable a direct comparison between the SEGE and CLUCB algorithms, we restrict our attention to a summable sequence of

risk levels that satisfy the conditions of Corollary 2. Specifically, we set the sequence of risk levels to $\delta_t = 6\overline{\delta}/(\pi^2 t^2)$ for all stages $t \geq 1$, where $\overline{\delta} = 0.1$.

CLUCB Algorithm.

We note that the implementation of the CLUCB algorithm requires the repeated solution of a non-convex optimization problem in order to compute UCB arms. To circumvent this intractable calculation, we approximate the continuous set of arms $\mathcal{X}$ by a finite set of arms $\widehat{\mathcal{X}}$ that correspond to a uniform discretization of the boundary of $\mathcal{X}$. The error induced by this approximation is negligible, as $\max_{x \in \mathcal{X}} \langle x, \theta^* \rangle - \max_{x \in \widehat{\mathcal{X}}} \langle x, \theta^* \rangle \leq 3 \times 10^{-3}$.

3.6.2 Performance of the SEGE Algorithm

We first discuss the transient behavior and performance of the SEGE algorithm. As one might expect, the SEGE algorithm initially relies on the baseline arm for safe exploration as depicted in Figure 3.3a. Over time, as the algorithm accumulates information, it is able to gradually expand the set of safe arms as shown in Figure 3.2. This expansion enables the algorithm to increase the stagewise efficiency with which it safely explores by selecting arms in the vicinity of the safe arm with the largest lower confidence bounds on their expected rewards. In turn, the SEGE algorithm is able to exploit the information gained to play the greedy with increasing frequency over time. As a result, the growth rate of regret diminishes over time as depicted in Figure 3.3c. Critically, Figure 3.3a also shows that the SEGE algorithm maintains stagewise safety throughout each of the 250 independent experiments.

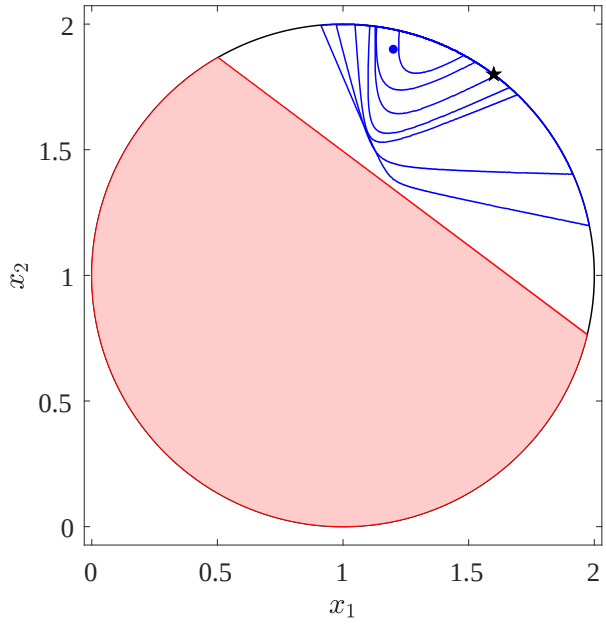
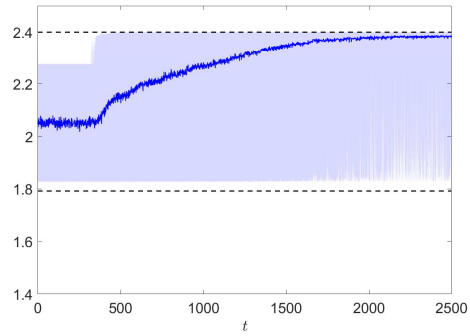


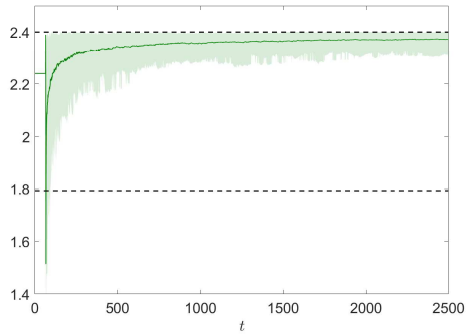
Figure 3.2: The blue curves depict the gradual expansion of the set of safe arms $\{x \in \mathcal{X} \mid \text{LCB}_t(x) \geq b\}$ over time under the SEGE algorithm for $t = 250, 500, 1000, 2000, 5000, 10000$, and 50000. The blue dot depicts the baseline arm X_0, the black star depicts the optimal arm X^*, and the red shaded area depicts the set of unsafe arms.

3.6.3 Comparison with the CLUCB Algorithm

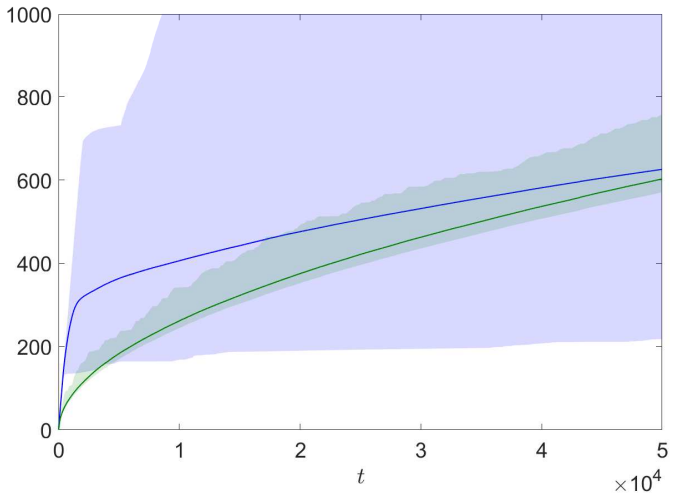
Unlike the SEGE algorithm, the CLUCB algorithm is seen to violate the stagewise safety constraint at an early stage in the learning process as depicted in Figure 3.3b. The violation of the stagewise safety constraint by the CLUCB algorithm is not surprising as it is only guaranteed to respect the conservative performance constraint (3.6). The SEGE algorithm, on the other hand, is guaranteed to satisfy the conservative performance constraint, in addition to being stagewise safe (cf. Corollary 2). However, as one might expect, the more stringent safety guarantee of the SEGE algorithm comes at a cost. Specifically, the regret under the SEGE algorithm initially grows more rapidly than the regret incurred by the CLUCB algorithm, as shown in Figure 3.3c. However, over time



(a) Stagewise expected reward under the SEGE algorithm.



(b) Stagewise expected reward under the CLUCB algorithm.



(c) Cumulative regret of the SEGE algorithm (blue) and the CLUCB algorithm (green).

Figure 3.3: These figures illustrate the empirical performance of the SEGE and CLUCB algorithms. The solid lines depict empirical means and the shaded regions depict empirical ranges computed from 250 independent simulations.

the growth rate of regret of the SEGE algorithm slows down as information accumulates and the need for safe exploration diminishes enabling the algorithm to play the greedy arm more frequently.

CHAPTER 4

ONLINE LEARNING OF NONPARAMETRIC MODELS

In many practical applications, linearly parameterized models are not flexible enough to capture the characteristics of the underlying system. However, one may be able to utilize the insight gained from the analysis of linearly parameterized models to develop methods that can be applied to more complex models. In particular, in this chapter, we rely on the methodology developed in Chapter 2 to design an online learning algorithm for a more general class of nonparametric smooth functions, which is of practical importance in modern electric power networks.

4.1 Introduction

The large scale utilization of demand response (DR) resources has the potential to substantially improve the reliability and efficiency of electric power systems. Accordingly, several state and federal mandates have been established to facilitate the integration of demand response resources into wholesale electricity markets. For example, FERC Order 719 mandates that Independent System Operators (ISOs) permit the direct sale of energy produced by DR resources into wholesale electricity markets [35]. However, as individual residential customers often posses insufficient capacity to participate in such markets directly, there emerges the need for an intermediary, or *aggregator*, with the ability to coordinate the demand response of large numbers of residential customers for direct sale into the wholesale electricity market. Such is consistent with the growing multitude of ISO and utility-run DR programs, which require that aggregated DR resources have a minimum load curtailment capability. For example, the

Proxy Demand Resource (PDR) program operated by the California ISO has minimum capacity requirement of 100 kW, while the Day-Ahead Demand Response Program (DADRP) operated by the New York ISO has a more stringent capacity requirement of two MW.

In this chapter, we adopt the perspective of an aggregator, which seeks to coordinate its *purchase* of an aggregate demand reduction from a fixed group of residential electricity customers, with its *sale* of the aggregate demand reduction into a two-settlement wholesale energy market.[1] Formally, this amounts to a two-sided optimization problem, which requires the aggregator to balance the cost it incurs in procuring a reduction in demand from participating customers against the revenue it derives from its sale of the (a priori uncertain) demand reduction into the wholesale energy market.

More specifically, we consider the setting in which the aggregator purchases demand reductions from its customers using a non-discriminatory, posted price mechanism. That is to say, each participating customer is payed for her reduction in electricity demand according to a uniform per-unit energy price determined by the aggregator. Pricing mechanisms of this form fall within the more general category of DR programs that rely on peak time rebates (PTR) as incentives for demand reduction. Prior to its realization of the aggregate demand reduction, the aggregator must also determine how much energy to sell into the two-settlement energy market. In the day-ahead (DA) market, the aggregator commits to a forward energy contract, which calls for delivery of the contracted energy in the real-time (RT) market. If the realized reduction in demand exceeds (falls short of) the forward contract, then the difference is sold (bought) in the

[1] From the perspective of the wholesale electricity market, the provisioning of a measurable reduction in demand from an aggregator is equivalent to an increase in supply.

RT market. Therefore, in order to maximize its profit, the aggregator must co-optimize the DR price it offers its customers with the forward contract that it commits to in the wholesale energy market, as the former determines its ability to deliver the latter.

There are a variety of challenges that the aggregator faces in operating such DR programs. *The most basic challenge is the prediction of how customers will adjust their aggregate demand in response to different DR prices*, i.e., the aggregate demand curve. If the offered price is too low, consumers may be unwilling to curtail their demand; if the offered price is too high, the aggregator pays too much and gets more reduction than is needed. As the aggregator is initially ignorant to the customers' aggregate demand curve, the aggregator must attempt to learn a model of customer behavior over time through repeated observations of demand reductions in response to the DR prices that it offers. Simultaneously, the aggregator must jointly adjust its DR prices and forward contract offerings in such a manner as to facilitate profit maximization over time. As we will later show, such tasks are intimately related, and give rise to a fundamental trade-off between the need to *learn* (explore) and *earn* (exploit).

Contribution: In this chapter, we study the setting in which the aggregator is faced with an aggregate demand curve that is unknown, and subject to unobservable, additive random shocks. We do not make any parametric assumption on the aggregate demand curve. Specifically, we assume that both the demand curve and the probability distribution of the random shocks are fixed, but *initially unknown* to the aggregator. Faced with such ignorance, we explore the extent to which the aggregator might dynamically adapt its posted DR prices and offered contracts to maximize its expected profit over a time frame of T

days. Specifically, we design a causal pricing and contract offering policy that resolves the aggregator's need to learn the unknown demand model with its desire to maximize its cumulative expected profit over time. The proposed pricing policy is proven to exhibit *regret* (relative to an oracle) over T days that is at most $O(\log(T)\sqrt{T})$. In addition, the proposed policy is proven to generate a sequence of posted DR prices and forward contracts that converge to the oracle optimal DR price and forward contract in the mean square sense. Our method generalizes the approach introduced by Besbes and Zeevi [12]. Specifically, they consider the problem of maximizing the expected revenue of a seller of a single product with unknown demand function. In contrast to their setting, in our two-sided optimization problem, there is a need to learn the underlying distribution of the demand shocks (in addition to the demand function).

Related Work: There is a large body of literature in power systems concerned with the aggregation and coordination of flexible demand-side resources to optimize certain economic objectives that an aggregator might encounter in wholesale energy or ancillary service markets. In such settings, the aggregator will typically exercise control over the consumption of participating demand-side resources using either (1) a *direct load control* mechanism whereby the aggregator can directly regulate the consumption of participating load resources according to a pre-specified contract [14, 21, 23, 32, 43, 55, 66, 69, 78, 84, 98]; or (2) an *indirect load control* mechanism whereby customers adjust their load in response to price signals or incentives offered by the aggregator (e.g., time-of-use pricing, peak time rebates, etc.) [15, 37, 47, 62, 61, 65, 76, 99].

The literature—as it relates to the problem of co-optimizing an aggregator's (two-sided) transactions between end-use customers and the wholesale

market—is much less developed. Campaigne et al. [17] consider a two-sided market model that is perhaps closest in nature to the one considered in this chapter. Specifically, the authors adopt a mechanism design approach to the procurement of load reductions from customers, where customers are rationed and remunerated according to their self-reported types.[2] In this chapter, we adopt a *posted price* approach to the procurement of demand reductions from customers. This is in sharp contrast to the mechanism design approach of [17], as it gives rise to the need to learn customers' types (i.e., demand functions) over time from measured data. From a practical standpoint, there are a variety of reasons as to why a posted price approach might be preferable to the mechanism design approach advocated by Campaigne et al. [17], not the least of which pertains to the simplicity and ease of implementation of posted pricing schemes. We refer the reader to [59] for a detailed discussion surrounding the advantages and disadvantages of such an approach in the context of online marketplaces. To the best of our knowledge, this chapter is the first to analyze the use of a posted pricing scheme by an aggregator participating in such two-sided markets.

Organization: The remainder of the chapter is organized as follows. In Section 4.3, we formulate the aggregator's profit maximization problem. In Section 4.4, we propose an adaptive pricing and contract offering policy for the aggregator. In Section 4.5 , we provide a theoretical analysis that establishes a sublinear growth rate of the expected regret incurred by the proposed policy. In Section 4.6, we illustrate the performance of our proposed policy with a numerical case study.

[2]We refer the reader to [20, 26] for a related line of literature, which also employs a mechanism design approach to the procurement of demand reductions in such two-sided markets.

4.2 Notation

We denote by $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{R}$ the sets of natural, integer, and real numbers, respectively. Given a real number $x \in \mathbb{R}$, we denote $\lfloor x \rfloor := \max\{m \in \mathbb{Z} \mid m \leq x\}$ and $x_+ := \max\{0, x\}$. Given a function $h : \mathbb{R} \to \mathbb{R}$, we denote its first and second derivative with respect to its argument by h' and h'', respectively, i.e., $h'(x) = dh(x)/dx$ and $h''(x) = d^2h(x)/dx^2$.

4.3 Model

We adopt the perspective of an aggregator who seeks to purchase demand reductions from a fixed group of N customers for sale into a two-settlement wholesale energy market. The market is assumed to repeat over multiple time periods indexed by $t = 1, 2, \ldots$. Each time period can be viewed as the specific time-slot during the day in which the demand response is scheduled (e.g., peak-load hour on each day). The actions taken by the both aggregator and customers are described in detail in the following subsections, and concisely summarized in Table 4.1.

4.3.1 Two-Settlement Market Model

At the beginning of each day t, the aggregator commits to a forward contract for energy in the day-ahead (DA) market in the amount of Q_t (kWh). The forward contract is remunerated at the *DA energy price*. The forward contract calls for delivery in the real-time (RT) market. If the energy delivered by the aggregator

(i.e., the aggregate demand reduction) falls short of the forward contract, the aggregator must purchase the shortfall in the RT market at the *shortage price*. If the energy delivered exceeds the forward contract, the aggregator must sell the excess supply in the RT market at the *overage price*.[3] The wholesale energy prices may vary both during each day and from day to day. We denote the wholesale energy prices (measured in \$/kWh) averaged over the demand response time-slot on day t by:

- λ_t, DA energy price,
- ρ_t^+, RT overage price,
- ρ_t^-, RT shortage price.

We make several standard assumptions regarding the aggregator's actions and the determination of energy prices in the wholesale market. First, we assume that the aggregator's maximum demand curtailment capacity is small relative to the total volume of the energy market. Under this assumption, it is reasonable to assume that the aggregator cannot appreciably affect price. Accordingly, we assume that the aggregator behaves as a *price taker* in the DA and RT energy markets. Second, as the wholesale energy prices λ_t, ρ_t^+, and ρ_t^- are *not known* to the aggregator at the time of posting the DR price, which is prior to committing to a forward contract in the DA market, we model them as random variables whose expected values are denoted by

$$\mu_\lambda := \mathbb{E}\left[\lambda_t\right], \quad \mu_\rho^+ := \mathbb{E}\left[\rho_t^+\right], \quad \text{and} \quad \mu_\rho^- := \mathbb{E}\left[\rho_t^-\right]$$

[3] We note that this two-settlement market structure reflects existing market rules, which govern the behavior of aggregators in a variety of DR programs in operation today—including the day-ahead demand response program (DADRP) and the proxy demand resource (PDR) program administered by the New York ISO and the California ISO, respectively.

for each period t. Note that while we allow the *realizations* of wholesale energy prices to vary across time, we require that their *expected values* be time invariant. We make the following technical assumption in a similar manner to [17].

Assumption 5. The DA energy price satisfies $\mu_\lambda > 0$ and $\mu_\rho^+ < \mu_\lambda < \mu_\rho^-$.

Assumption 5 serves to facilitate clarity of exposition and analysis in the sequel, as it will preserve the concavity of the aggregator's expected profit function (4.2). Moreover, this assumption eliminates the possibility of perverse market outcomes in which the aggregator offers forward energy contracts with the explicit intention of deviating from the contract in the RT market.

4.3.2 Demand Response Model

In order to fulfill its forward contract commitment Q_t on day t, the aggregator must elicit an aggregate reduction in demand from its customers. It does so by broadcasting a uniform DR price $p_t \geq 0$ prior to observing the DA price, to which each customer i responds with a reduction in demand in the amount of D_{it} (kWh) in real time. This entitles each customer i to receive a payment of $p_t D_{it}$. We note that implicit in this model is the assumption that each customer's reduction in demand is measured against a *predetermined baseline.* The problem of accurately estimating baseline demand is a challenging and active area of research [19, 22, 25, 64]. The generalization of our model to accommodate the endogenous estimation of a priori uncertain customer baselines is left as a direction for future research.

Actor (Decision)	Description of actions on day t
Aggregator (p_t)	Prior to committing to a contract in the DA market, the aggregator posts a uniform price p_t for demand reduction to the participating customers.
Aggregator (Q_t)	In the DA market, the aggergrator commits to a forward energy contract Q_t, which calls for delivery over prespecified interval of time in the RT market.
Customers (D_t)	In the RT market, customers respond to the aggregator's offered price p_t by reducing their aggregate demand by an amount D_t.

Table 4.1: Description and timing of actions taken by the aggregator and customers.

We model the aggregate demand reduction $D_t := \sum_{i=1}^{N} D_{it}$ as

$$D_t = g(p_t) + \varepsilon_t, \tag{4.1}$$

where $g : \mathbb{R}_+ \to \mathbb{R}_+$ is a deterministic function representing the expected aggregate demand reduction given the DR price p_t, and ε_t is an unobservable random demand shock. We interchangeably refer to g, the expected aggregate demand reduction, as the *demand function* throughout.

We assume that both the demand function and the probability distribution function of the demand shock are initially *unknown to the aggregator*. We allow the expected demand function to be nonlinear, and do not make explicit parametric assumptions about its form. We assume that g is concave, strictly increasing, and twice continuously differentiable in addition to the following technical assumption.

Assumption 6. There exists a constant $\kappa \in [0, 1)$ such that

$$\frac{1}{2}\left(\frac{|g''(p)|g(p)}{g'(p)^2}\right) \leq \kappa$$

for all $p \in [0, \mu_\lambda]$.

There are several standard demand functions that satisfy the requirements of Assumption 6 including linear, exponential, and logit demand functions. Loosely speaking, Assumption 6 imposes a restriction on the curvature of the demand function. This restriction of the curvature, enables the utilization of misspecified linear models in the problem of optimizing the aggregator's profit. In other words, demand functions with smaller values of κ can be more accurately approximated with linear functions. In particular, the constant κ associated with the family of linear functions is $\kappa = 0$. Moreover, in the proof of Lemma 3, we show that Assumption 6 ensures convexity of $pg(p)$, which, in turn, is utilized to show concavity of the aggregator's profit (4.2).

We also assume that the sequence of aggregate demand shocks $\{\varepsilon_t\}$ are zero-mean, independent and identically distributed (IID) random variables, which are mutually independent from the wholesale energy prices $\{\lambda_t\}$, $\{\rho_t^+\}$, and $\{\rho_t^-\}$.

Assumption 7. The aggregate demand shock ε_t takes values in the interval $[\underline{\varepsilon}, \bar{\varepsilon}]$ for all $t \geq 1$. Moreover, its cumulative distribution function F is strictly increasing over this range.

The assumption that the aggregate demand shock takes bounded values is natural, given the physical limitation on the range of values that demand can take. We also note that we do not require the aggregator to have explicit knowledge of $\underline{\varepsilon}$ and $\bar{\varepsilon}$ specified in Assumption 7. It is worth noting that the rigorous guarantees provided in this chapter hold under a weaker assumption on the aggregate demand shock. In particular, the assumption of boundedness of the

demand shock can be relaxed to sub-Gaussianity.

4.3.3 Aggregator Profit

The expected profit derived by the aggregator during period t given a fixed forward contract Q_t and price p_t is determined by

$$\pi_t(Q_t, p_t) := \mathbb{E}\left[\lambda_t Q_t + \rho_t^+(D_t - Q_t)_+ - \rho_t^-(Q_t - D_t)_+ - p_t D_t\right],$$

where the expectation is taken with respect to the randomness in the wholesale energy prices and the demand shock. Given our previous assumption that the demand shocks $\{\varepsilon_t\}$ are mutually independent from the wholesale energy prices $\{\lambda_t\}$, $\{\rho_t^+\}$, and $\{\rho_t^-\}$, the expected profit function simplifies to

$$\pi_t(Q_t, p_t) = \mu_\lambda Q_t + \mu_\rho^+ \mathbb{E}\left[(D_t - Q_t)_+\right] - \mu_\rho^- \mathbb{E}\left[(Q_t - D_t)_+\right] - p_t g(p_t). \tag{4.2}$$

Here, expectation is taken with respect to the random demand shock ε_t.

We define the *oracle optimal contract and price* as

$$(Q^*, p^*) := \underset{(Q,p)\in\mathbb{R}^2}{\operatorname{argmax}}\ r_t(Q, p). \tag{4.3}$$

That is to say, (Q^*, p^*) denote the forward contract and DR price, which jointly maximize the aggregator's expected profit on day t given perfect knowledge of the demand model. It is straightforward to calculate the oracle optimal contract and price from the first-order optimality condition associated with problem (4.3), as the expected profit criterion (4.2) is guaranteed to be jointly *concave* in its arguments given Assumptions 5 and 6. The implicit equations determining the oracle optimal price and contract are given in the following lemma.

Lemma 3 (Oracle Optimal Policy)**.** The oracle optimal contract and price (Q^*, p^*) satisfy

$$p^* = \mu_\lambda - \frac{g(p^*)}{g'(p^*)}, \tag{4.4}$$

$$Q^* = g(p^*) + F^{-1}(\zeta), \tag{4.5}$$

where

$$\zeta := \frac{\mu_\lambda - \mu_\rho^+}{\mu_\rho^- - \mu_\rho^+}.$$

Moreover, the optimal price is unique and $p^* \in [0, \mu_\lambda]$.

Here, $F^{-1}(\zeta) := \inf\{x \in \mathbb{R} \mid F(x) \geq \zeta\}$ denotes the ζ-quantile of the random demand shock ε_t. Assumption 5 ensures that the price ratio ζ is a valid probability, i.e., $\zeta \in (0, 1)$. Several comments are in order. The structure of the oracle optimal contract resembles the optimal inventory control decision in the classical newsvendor problem in the revenue management literature [6]. It is also worth noting that Q^* can be interpreted as the minimum demand reduction that the aggregator is guaranteed to receive with probability at least $1 - \zeta$ under the oracle optimal price p^*. Moreover, the price ratio ζ is inversely proportional to the difference between the imbalance prices. Therefore, the aggregator will offer larger contracts as the difference between imbalance prices decreases.

We define the *oracle optimal profit* accumulated over T time periods as

$$\Pi_T^* := \sum_{t=1}^{T} \pi_t(Q^*, p^*).$$

We employ the term *oracle*, as Π_T^* equals the maximum expected profit that an aggregator is able to extract over T time periods given perfect knowledge of the demand model at the outset.

4.3.4 Policy Design and Regret

We consider the scenario in which the aggregator knows neither the demand function g or the aggregate shock distribution F at the outset. Accordingly, the aggregator must endeavor to learn these features directly from the demand response data that it collects over time in response to its posted DR prices. At the same time, the aggregator must dynamically adapt its sequence of posted DR prices (and forward contract offerings) to improve its profit over time. In what follows, we describe the space of feasible policies that the aggregator might use to guide its adaptation of contracts $\{Q_t\}$ and DR prices $\{p_t\}$ over time.

Prior to its determination of the contract Q_t and the price p_t at time t, the aggregator has access to the entire history of prices, contract offerings, and aggregate demand reductions, up to and including time period $t-1$. We define a *feasible policy* as an infinite sequence of functions $\gamma := ((Q_1, p_1), (Q_2, p_2), \ldots)$, where each function in the sequence is allowed to depend only on the past data available until that point in time. More formally, we require that the functions (Q_t, p_t) be measurable according to the σ-algebra generated by the history of offered contracts, prices, and demand observations, i.e.,

$$(Q_1, \ldots, Q_{t-1}, p_1, \ldots, p_{t-1}, D_1, \ldots, D_{t-1})$$

for all time periods $t \geq 2$. For the initial time period $t = 1$, we require that (Q_1, p_1) be a pair of deterministic constants, as the aggregator has yet to collect any information about demand.

The *expected profit* generated by a feasible policy γ over T time periods is

defined as

$$\Pi_T^\gamma := \mathbf{E}^\gamma \left[\sum_{t=1}^{T} \pi_t(Q_t, p_t) \right], \tag{4.6}$$

where the expectation is taken with respect to the demand model (4.1) under the policy γ. We measure the performance of a feasible policy γ over T time periods according to the *T-period expected regret*, which is defined as

$$R_T^\gamma := \Pi_T^* - \Pi_T^\gamma.$$

The T-period expected regret incurred by a feasible policy equals the difference between the oracle optimal profit and the expected profit incurred by that policy over T time periods. Clearly, policies that produce low expected regret are preferred, as the oracle optimal profit is an upper bound on the maximum expected profit achievable by any feasible policy. Accordingly, we seek the design of policies whose T-period expected regret grows sublinearly with the horizon T. Such policies are said to have *no-regret* in the long run, as their average expected regret $(1/T)R_T^\gamma$ is guaranteed to vanish asymptotically. More formally, we have the following definition.

Definition 3 (No-Regret Policy)**.** A feasible policy γ is said to have *no-regret* if $\lim_{T\to\infty} R_T^\gamma/T = 0$.

4.4 Perturbed Certainty Equivalent Policy

In this section, we propose a pricing and contract offering policy that is guaranteed to have no-regret. We refer to this policy as the Perturbed Certainty Equivalent (PCE) policy. The PCE policy is episodic in nature. At the outset of each episode, it constructs a linear estimate of the demand function using only

data gathered during the preceding episode and discarding data from earlier episodes. This episodic approach to data selection will prove useful to limit the estimation error due to the misspecification of the demand model by a linear function. The PCE policy is semi-greedy. In particular, as depicted in Figure 4.1, each episode is split into two phases, with the first phase being dedicated to greedy exploitation, and the latter phase being dedicated to exploration. More specifically, during the exploitation phase of the episode, a certainty equivalent (CE) price and contract is offered based on the linear estimate of the demand model. That is, the price and contract that are optimal assuming the correctness of the estimated linear model. During the exploration phase of the episode, the PCE policy adds deliberate perturbations to the CE price to generate exploration, which facilitates improvement of model estimation in the proceeding episode. We design the length of each episode and the magnitude of these perturbations to balance the trade-off between exploration and exploitation near-optimally. Specifically, we show that the expected regret under the PCE policy is guaranteed to be no more than $O(\log(T)\sqrt{T})$.

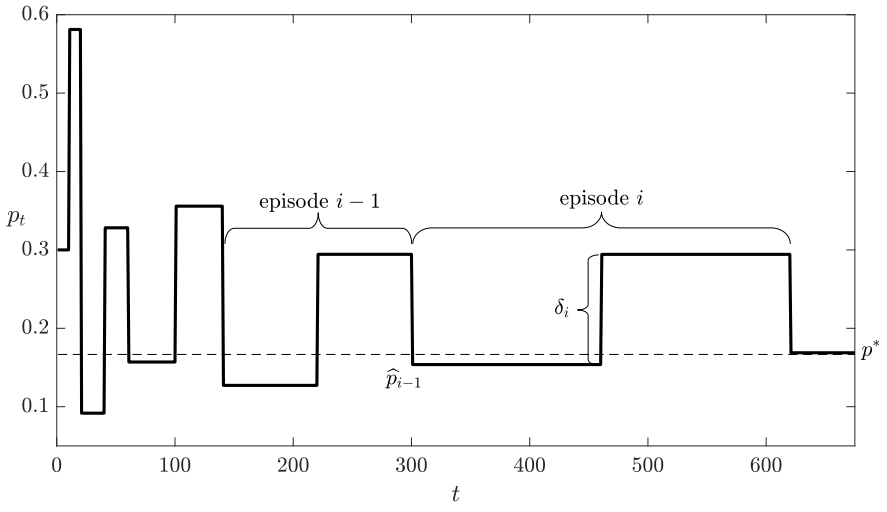


Figure 4.1: A sample path of the sequence of DR prices under the PCE policy.

4.4.1 Estimation via Linearization

We now introduce a linear approximation of the demand model that will be utilized in the PCE policy. More specifically, we define $\alpha(p)$ and $\beta(p)$ as the slope and the intercept of the linearization of the true demand function at a particular price p, respectively, i.e.,

$$\alpha(p) := g'(p),$$
$$\beta(p) := g(p) - g'(p)p.$$

The oracle optimal price (4.4) can be equivalently expressed in terms of the demand function linearization as

$$p^* = \frac{1}{2}\left(\mu_\lambda - \frac{\beta(p^*)}{\alpha(p^*)}\right). \tag{4.7}$$

Thus, in order to learn the oracle optimal price, it suffices to learn the parameters of the linearization of the demand function at the oracle optimal price.

As the underlying demand model may be nonlinear, incorporating the entire history of past observations to estimate the linearization of the demand model may result in estimation bias. Thus, in order to accurately estimate $\alpha(p)$ and $\beta(p)$ for a particular price p, only demand observations in response to prices that are in close proximity to p should be assimilated. The PCE policy streamlines this data selection by partitioning the time horizon into episodes in which the DR prices do not vary significantly. The policy then only uses observations taken during the preceding episode to estimate these parameters and discards observations during earlier episodes. More formally, we partition the natural numbers $\mathbb{N}$ into episodes $\mathcal{E}_i$ of length $2L_i$, i.e.,

$$\mathcal{E}_i := \{T_{i-1} + 1, \ldots, T_{i-1} + 2L_i\},$$

where T_i denotes the last time period of episode i. Recursively, we have that

$$T_i := T_{i-1} + 2L_i,$$

where $T_0 := 0$.

Given the history of demand and price observations during episode i, the *least squares estimate* (LSE) of the parameters of the linearization of the demand model is defined as

$$(\widehat{\alpha}_i, \widehat{\beta}_i) := \operatorname*{argmin}_{(\alpha,\beta)\in\mathbb{R}^2} \sum_{t\in\mathcal{E}_i} \left(D_t - (\alpha p_t + \beta)\right)^2. \tag{4.8}$$

In addition to estimating the demand model, we estimate the distribution of the demand shock as the oracle optimal contract depends on the quantile function of the demand shock. As the demand shock is unobservable, we rely on the sequence of residuals to estimate the quantile function. More precisely, let $\{\widehat{\varepsilon}_{it}\}$ be the sequence of residuals associated with the linear demand estimates $(\widehat{\alpha}_i, \widehat{\beta}_i)$, i.e.,

$$\widehat{\varepsilon}_{it} := D_t - (\widehat{\alpha}_i p_t + \widehat{\beta}_i) \text{ for } t = T_{i-1}+1, \ldots, T_{i-1}+L_i.$$

We denote by $\widehat{F}_i$, the empirical distribution function associated with the residuals, i.e., for all $x \in \mathbb{R}$ define $\widehat{F}_i(x) := (1/L_i)\sum_{t=T_{i-1}+1}^{T_{i-1}+L_i} \mathbb{1}\{\widehat{\varepsilon}_{it} \le x\}$. The empirical quantile function associated with the residuals is then given by

$$\widehat{F}_i^{-1}(\eta) := \inf\left\{x \in \mathbb{R} \mid \widehat{F}_i(x) \ge \eta\right\} \tag{4.9}$$

for all $\eta \in [0,1]$.

Then, the CE price and contract are defined as

$$\widehat{p}_i := \mathscr{P}\left(\frac{1}{2}\left(\mu_\lambda - \frac{\widehat{\beta}_{i-1}}{\widehat{\alpha}_{i-1}}\right)\right), \tag{4.10}$$

$$\widehat{Q}_i := \widehat{\alpha}_{i-1}\widehat{p}_i + \widehat{\beta}_{i-1} + \widehat{F}_{i-1}^{-1}(\zeta), \tag{4.11}$$

where $\mathscr{P} := \max\{0, \min\{\mu_\lambda, p\}\}$ is the projection operator onto interval $[0, \mu_\lambda]$. The projection of the price onto the interval $[0, \mu_\lambda]$ reduces the pricing error as $p^* \in [0, \mu_\lambda]$ from Lemma 3.

4.4.2 Price Exploration

Although it may seem natural to adopt the CE policy (4.10) and (4.11), it fails to elicit the information needed to accurately estimate the linearization of the demand function. More specifically, under this myopic approach, the estimated parameters may converge to a value that is different from the true model parameters. This phenomenon known as *incomplete learning* is well-documented in the adaptive control literature [16, 54, 56] and the revenue management literature [30, 50]. Sufficient exploration (excitation) in the sequence of prices is required to avoid incomplete learning, and ensuring that the policy exhibits no-regret.

The PCE policy generates exploration by adding perturbations to the CE price. More precisely, at each time period in episode i, the policy sets the price and contract according to

$$p_t := \begin{cases} \widehat{p}_{i-1}, & t \in \{T_{i-1}+1, \ldots, T_{i-1}+L_i\}, \\ \widehat{p}_{i-1} + \delta_i, & t \in \{T_{i-1}+L_i+1, \ldots, T_i\}, \end{cases} \tag{4.12}$$

$$Q_t := \widehat{Q}_{i-1}, \qquad t \in \{T_{i-1}+1, \ldots, T_i\}, \tag{4.13}$$

where $(\widehat{Q}_i, \widehat{p}_i)$ are the certainty equivalent contract and price defined as (4.11) and (4.10), and $(\widehat{Q}_0, \widehat{p}_0)$ are deterministic constants.

The length of each episode $L_i \in \mathbb{N}$ and the magnitude of the price perturbations $\delta_i \in \mathbb{R}_+$ play a critical role in balancing the trade-off between exploration

and exploitation. On the one hand, as the underlying demand curve may be non-linear, δ_i should be small to reduce the estimation bias. Moreover, to limit the regret incurred during the exploration phase of each episode, the perturbations must decay as the sequence of CE prices converge to the oracle optimal price. On the other hand, δ_i and L_i should be large enough to ensure sufficient exploration to accurately estimate the demand model, which will be used in the proceeding episode. A choice of L_i and δ_i that balances the trade-off between exploration and exploitation in a way that the policy exhibits near-optimal regret is given by

$$\delta_i = \delta_0 L_i^{-1/4}, \tag{4.14}$$

$$L_i = \lfloor L_0 \nu^i \rfloor, \tag{4.15}$$

where $\delta_0 > 0$, $L_0 \geq 1$, and $\nu > 1$ are user specified constants. Notice that by construction the length of episodes are increasing over time, i.e., $L_{i+1} > L_i$, and the size of price perturbations decays over time such that $\lim_{i\to\infty} \delta_i = 0$.

4.5 Theoretical Results

In this section, we establish a near-optimal upper bound on the expected regret incurred by the PCE policy. We first establish an upper bound on the expected regret of any feasible policy in terms of the pricing and contract offering errors relative to their oracle optimal counterparts.

Theorem 6. Let γ be a feasible policy. There exist finite positive constants C_0 and C_1 such that the T-period expected regret under γ is upper bounded by

$$R_T^\gamma \leq C_0 \mathbf{E}^\gamma \left[\sum_{t=1}^{T} (p_t - p^*)^2 \right] + C_1 \mathbf{E}^\gamma \left[\sum_{t=1}^{T} \left(Q_t - g(p_t) - (Q^* - g(p^*)) \right)^2 \right] \tag{4.16}$$

for all $T \geq 1$. Here, (Q^*, p^*) denote the oracle optimal contract and price.

We now apply Theorem 6 to establish an upper bound on the T-period expected regret under the PCE policy. In particular, we bound the expected regret in terms of the episode lengths L_i, price perturbations δ_i, CE pricing error $\widehat{p}_i - p^*$, and quantile estimation error $F_i^{-1}(\zeta) - F^{-1}(\zeta)$. Here, F_i^{-1} denotes the empirical quantile associated with the sequence of demand shocks, i.e.,

$$F_i^{-1}(\eta) := \inf \ \{x \in \mathbb{R} \mid F_i(x) \geq \eta\} \tag{4.17}$$

for all $\eta \in [0, 1]$ where F_i denotes the empirical distribution of the sequence of demand shocks defined as $\widehat{F}_i(x) = (1/L_i) \sum_{t=T_{i-1}+1}^{T_{i-1}+L_i} \mathbb{1}\{\varepsilon_t \leq x\}$ for all $x \in \mathbb{R}$. Note that we cannot utilize F_i^{-1} (instead of $\widehat{F}_i^{-1}$) in the design of a feasible policy as the sequence of demand shocks are unobservable as the demand function is unknown at the outset.

Corollary 3. There exist finite positive constants C_0, C_1, and C_2 such that the T-period expected regret under the PCE policy is upper bounded by

$$R_T^{\gamma} \leq C_2 \sum_{i=1}^{I_T} L_i \delta_i^2 + 3C_0 \sum_{i=1}^{I_T} L_i \mathbb{E}\left[(\widehat{p}_i - p^*)^2\right] + 3C_1 \sum_{i=1}^{I_T} L_i \mathbb{E}\left[\left(F_i^{-1}(\zeta) - F^{-1}(\zeta)\right)^2\right], \tag{4.18}$$

for all $T \geq 1$ where I_T denotes the number of episodes covering the horizon T, i.e., $I_T := \min\{i \in \mathbb{N} \mid T_i \geq T\}$.

To establish an upper bound on the expected regret of the PCE policy, we bound each term in Inequality (4.18) separately. We first establish an upper bound on the mean squared pricing error in Lemma 4.

Lemma 4 (Mean Squared Pricing Error)**.** Let parameter $\nu \in (1, 4/(1+\kappa^2)^2)$.

Then, there exists a finite positive constant C_3 such that the mean squared pricing error under the PCE policy is upper bounded by

$$\mathbb{E}\left[(\widehat{p}_i - p^*)^2\right] \leq C_3 \frac{i-1}{\sqrt{\nu^{i-1}}} \tag{4.19}$$

for all $i \geq 2$.

The proof of Lemma 4 follows from a similar argument as the proof of [12, Theorem 1]. To keep the chapter self-contained, we provide its proof in Appendix C.4.

The parameter ν controls the rate at which the length of episodes grow over time. The upper bound on this growth rate specified in Lemma 4, $4/(1+\kappa^2)^2$, is a decreasing function of the curvature parameter κ defined in Assumption 6. The CE price needs to be adjusted more frequently for demand models with larger curvature to account for the model misspecification due to the linearization of the demand function. Equivalently, the lengths of episodes should grow more slowly for demand functions with larger curvatures.

We now establish an upper bound on the mean squared quantile estimation error that is inversely proportional to the length of the episodes.

Lemma 5 (Mean Squared Quantile Error)**.** There exists a finite positive constant C_4 such that for all $\eta \in (0,1)$

$$\mathbb{E}\left[\left(F_i^{-1}(\eta) - F^{-1}(\eta)\right)^2\right] \leq C_4 \frac{1}{L_i}. \tag{4.20}$$

We provide a detailed proof of Lemma 4.20, which utilizes Hoeffding's inequality [42] in Appendix C.5.

Finally, we upper bound the expected regret under the PCE policy using Corollary 3 and intermediary Lemmas 4 and 5.

Theorem 7 (Upper Bound on Regret)**.** Let parameter $\nu \in (1, 4/(1+\kappa^2)^2)$. There exists a finite positive constant C_5 such that the T-period expected regret incurred by the PCE policy is upper bounded by

$$R_T^{\gamma} \leq C_5 \log(T)\sqrt{T} \tag{4.21}$$

for all $T \geq 1$.

The proof of Theorem 7 follows from applying the upper bounds on the mean squared pricing error (4.19) and the mean squared quantile error (4.20) to the upper bound on the expected regret in Corollary 3. We provide the detailed derivation of upper bound (4.21) in Appendix C.6.

4.6 Experiments

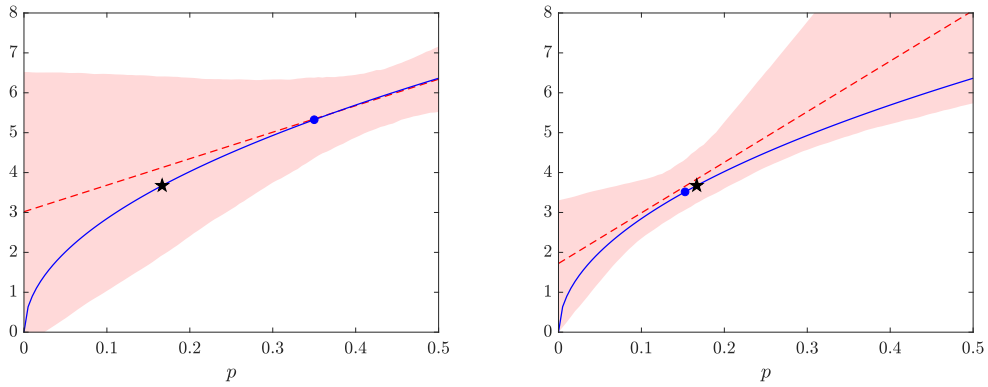
In this section, we we illustrate the behavior of the PCE policy for a nonlinear demand function using a synthetic example.

4.6.1 Model Parameters

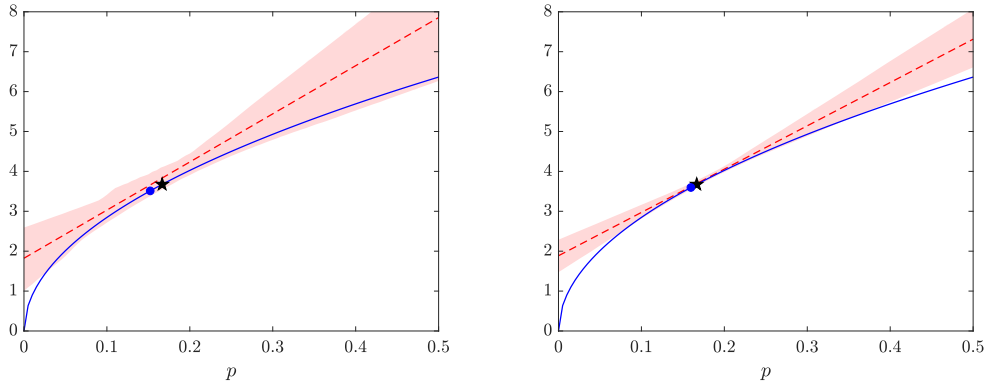
We assume that the aggregate demand function has the form

$$g(p) = 9\sqrt{p},$$

which satisfies Assumption 6 with $\kappa = 1/2$. This choice of demand function is consistent with the range of price elasticities observed in several real-time pricing programs operated in the United States [90, 34]. The estimated range of price elasticity for residential DR programs is $[0.04, 0.20]$. Under our model,



(a) Episode $i = 1$ (b) Episode $i = 6$



(c) Episode $i = 11$ (d) Episode $i = 16$

Figure 4.2: The figures depicts the demand function (in solid blue) and its estimated linearizations. The blue dot depicts $g(p_i)$ and the black star depicts the optimal demand $g(p^*)$. The red dashed lines depict the empirical mean of the estimated linearization of the demand functions at prices $\widehat{p}_i$, i.e., $\widehat{\alpha}_i p + \widehat{\beta}_i$, and the shaded areas depict their respective middle 80% empirical confidence interval computed using 1000 independent experiments.

assuming that the aggregator registers $N = 100$ customers in the program, this range of price elasticity is observed for DR prices in the range $[0.05, 1.26]$ \$/kWh. We assume that the sequence of demand shocks is an IID sequence of zero-mean normal random variables with variance equal to 2, truncated over the interval $[-4, 4]$. We set the mean of the DA energy price, the RT overage price, and the RT shortage price as $\mu_\lambda = 0.5$, $\mu_\rho^+ = 0.1$, and $\mu_\rho^- = 0.8$ (\$/kWh), respectively.

PCE Policy Parameters

We initialize the policy by choosing $\widehat{p}_0 = 0.7\mu_\lambda$ and $\widehat{Q}_0 = 0$. We choose the parameters of the PCE policy as $L_0 = 2$, $\delta_0 = 0.25$, and $\nu = 1.5$. Given these parameters the first episode consists of six time periods.

4.6.2 Discussion

Figure 4.2 illustrates improvement in two aspects of model estimation over time under the PCE policy. First, the estimation error of the parameters of the linearization of the demand model decreases over time. More precisely, one can observe that the empirical confidence region on $|(\alpha_i - \alpha(\widehat{p}_i)p - (\beta_i - \beta(\widehat{p}_i))|$ shrinks over episodes for all $p \in [0, 0.5]$. This improvement is due to increased exploration over episodes guaranteed by the choice of parameters δ_i and L_i of the form (4.14) and (4.15), respectively. That is, the exploration is guaranteed to increase over episodes as the rate at which the lengths of episodes grow is faster than the rate at which the magnitude of price perturbations decay.

Second, we observe that the sequence of CE prices converge to the oracle optimal price over time. Moreover, the CE pricing error together with the price perturbations decay at a sufficiently fast rate to ensure that the regret grows sublinearly as depicted in Figure 4.3.

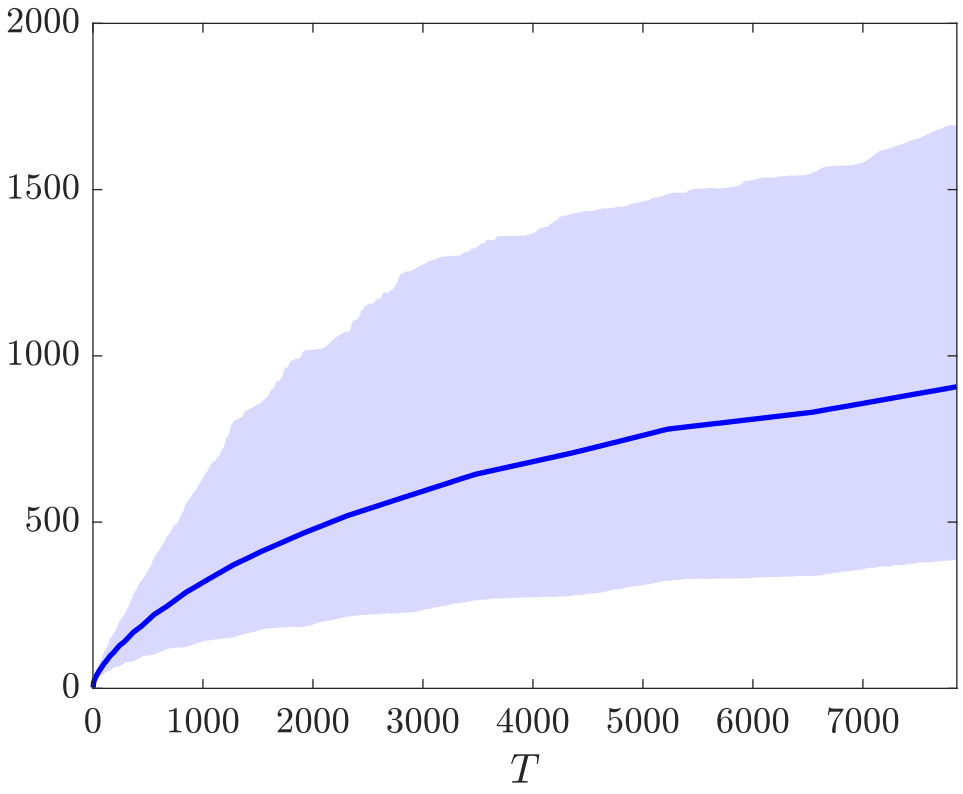


Figure 4.3: Regret under the PCE policy. The solid blue line depicts the empirical expected regret, and the shaded area depicts the middle 80% empirical confidence interval computed using 1000 independent experiments.

APPENDIX A

PROOFS OF RESULTS IN CHAPTER 2

In the following proofs, we consider a more general form of the perturbation as $\delta_t = \text{sgn}(c_t - c_{t-1}) \cdot t^{-r}$, where r is allowed to be an arbitrary constant in the interval $[0, 1/2)$. Ultimately, we will prove that a choice of $r = 1/4$ minimizes the asymptotic order of the upper bound on regret (ignoring logarithmic factors), which we establish in (A.15).

A.1 Proof of Lemma 1

The parameter estimation error derived in Equation (2.5) is given by

$$\theta_t - \theta = \mathscr{J}_t^{-1} \tilde{\varepsilon}_t,$$

where $\tilde{\varepsilon}_t$ is defined as

$$\tilde{\varepsilon}_t = \sum_{k=1}^{t} \begin{bmatrix} p_k \\ 1 \end{bmatrix} \varepsilon_k.$$

Using the Cauchy-Schwarz inequality and assuming that $\mathscr{J}_t$ is invertible, the 2-norm of parameter estimation error is bounded as follows.

$$\|\theta_t - \theta\|^2 = \|\mathscr{J}_t^{-1} \tilde{\varepsilon}_t\|^2 \leq \|\mathscr{J}_t^{-1/2}\|^2 \|\mathscr{J}_t^{-1/2} \tilde{\varepsilon}_t\|^2.$$

Using the definition of matrix norms, we get

$$\|\mathscr{J}_t^{-1/2}\|^2 = \left(\lambda_{\max}(\mathscr{J}_t^{-1/2})\right)^2 = \frac{1}{\lambda_{\min}(\mathscr{J}_t)},$$

where the operators $\lambda_{\max}$ and $\lambda_{\min}$ denote the largest and the smallest eigenvalues, respectively. In the following Lemma, we establish a lower bound on the minimum eigenvalue of $\mathscr{J}_t$ in terms of the price perturbations and the wholesale energy price variations.

Lemma 6. Under the perturbed myopic policy (2.11), it holds that

$$\lambda_{\min}(\mathscr{J}_t) \geq \frac{1}{1+\overline{p}^2} L_t \quad \text{a.s.}, \tag{A.1}$$

where L_t is defined as

$$L_t := \frac{1}{8}\left(\rho^2 \lfloor t/2 \rfloor^{1-2r} + \sum_{k=1}^{\lfloor t/2 \rfloor} (c_{2k} - c_{2k-1})^2\right). \tag{A.2}$$

Using Inequality (A.1), the mean squared parameter estimation error can be bounded as

$$\mathbb{E}\left[\|\theta_t - \theta\|^2\right] \leq \mathbb{E}\left[\frac{1}{\lambda_{\min}(\mathscr{J}_t)} \|\mathscr{J}_t^{-1/2} \tilde{\varepsilon}_t\|^2\right] \leq \frac{1+\overline{p}^2}{L_t} \mathbb{E}\left[\tilde{\varepsilon}_t^\top \mathscr{J}_t^{-1} \tilde{\varepsilon}_t\right]. \tag{A.3}$$

We now establish an upper bound on $\mathbb{E}\left[\tilde{\varepsilon}_t^\top \mathscr{J}_t^{-1} \tilde{\varepsilon}_t\right]$ by adopting a similar approach as [58, Lemma 1]. More specifically, we establish a recursive inequality relating $\mathbb{E}\left[\tilde{\varepsilon}_t^\top \mathscr{J}_t^{-1} \tilde{\varepsilon}_t\right]$ to $\mathbb{E}\left[\tilde{\varepsilon}_{t-1}^\top \mathscr{J}_{t-1}^{-1} \tilde{\varepsilon}_{t-1}\right]$. It holds that

$$\begin{aligned}
&\mathbb{E}\left[\tilde{\varepsilon}_t^\top \mathscr{J}_t^{-1} \tilde{\varepsilon}_t\right] \\
&\quad = \mathbb{E}\left[\left(\tilde{\varepsilon}_{t-1} + \begin{bmatrix} p_t \\ 1 \end{bmatrix} \varepsilon_t\right)^\top \mathscr{J}_t^{-1} \left(\tilde{\varepsilon}_{t-1} + \begin{bmatrix} p_t \\ 1 \end{bmatrix} \varepsilon_t\right)\right] \\
&\quad = \mathbb{E}\left[\tilde{\varepsilon}_{t-1}^\top \mathscr{J}_t^{-1} \tilde{\varepsilon}_{t-1}\right] + 2\mathbb{E}\left[\tilde{\varepsilon}_{t-1}^\top \mathscr{J}_t^{-1} \begin{bmatrix} p_t \\ 1 \end{bmatrix} \varepsilon_t\right] + \mathbb{E}\left[\begin{bmatrix} p_t \\ 1 \end{bmatrix}^\top \mathscr{J}_t^{-1} \begin{bmatrix} p_t \\ 1 \end{bmatrix} \varepsilon_t^2\right].
\end{aligned} \tag{A.4}$$

Using the fact that $\tilde{\varepsilon}_{t-1}$, p_t, and $\mathscr{J}_t$ are all measurable according to the σ-algebra generated by $\varepsilon_1, \ldots, \varepsilon_{t-1}$, and the law of iterated expectations, we get

$$\begin{aligned}
\mathbb{E}\left[\tilde{\varepsilon}_{t-1}^\top \mathscr{J}_t^{-1} \begin{bmatrix} p_t \\ 1 \end{bmatrix} \varepsilon_t\right] &= \mathbb{E}\left[\mathbb{E}\left[\tilde{\varepsilon}_{t-1}^\top \mathscr{J}_t^{-1} \begin{bmatrix} p_t \\ 1 \end{bmatrix} \varepsilon_t \middle| \varepsilon_1, \ldots, \varepsilon_{t-1}\right]\right] \\
&= \mathbb{E}\left[\tilde{\varepsilon}_{t-1}^\top \mathscr{J}_t^{-1} \begin{bmatrix} p_t \\ 1 \end{bmatrix} \mathbb{E}\left[\varepsilon_t \middle| \varepsilon_1, \ldots, \varepsilon_{t-1}\right]\right] \\
&= 0,
\end{aligned} \tag{A.5}$$

where the last identity follows from the fact that ε_t is independent of $\varepsilon_1, \ldots, \varepsilon_{t-1}$ and is zero-mean. Using a similar argument, we get

$$\begin{aligned}\mathbb{E}\left[\begin{bmatrix} p_t \\ 1 \end{bmatrix}^\top \mathscr{J}_t^{-1} \begin{bmatrix} p_t \\ 1 \end{bmatrix} \varepsilon_t^2\right] &= \mathbb{E}\left[\begin{bmatrix} p_t \\ 1 \end{bmatrix}^\top \mathscr{J}_t^{-1} \begin{bmatrix} p_t \\ 1 \end{bmatrix}\right] \mathbb{E}\left[\varepsilon_t^2\right] \\ &\leq \mathbb{E}\left[\begin{bmatrix} p_t \\ 1 \end{bmatrix}^\top \mathscr{J}_t^{-1} \begin{bmatrix} p_t \\ 1 \end{bmatrix}\right] \frac{(\overline{\varepsilon} - \underline{\varepsilon})^2}{4}, \end{aligned} \tag{A.6}$$

where the last inequality follows from Popoviciu's inequality on variances. By combining Equations (A.4) and (A.5) with Inequality (A.6), we get

$$\mathbb{E}\left[\tilde{\varepsilon}_t^\top \mathscr{J}_t^{-1} \tilde{\varepsilon}_t\right] \leq \mathbb{E}\left[\tilde{\varepsilon}_{t-1}^\top \mathscr{J}_t^{-1} \tilde{\varepsilon}_{t-1}\right] + \mathbb{E}\left[\begin{bmatrix} p_t \\ 1 \end{bmatrix}^\top \mathscr{J}_t^{-1} \begin{bmatrix} p_t \\ 1 \end{bmatrix}\right] \frac{(\overline{\varepsilon} - \underline{\varepsilon})^2}{4}. \tag{A.7}$$

Here, we bound each term in the right hand side of Inequality (A.7) separately. For the first term, using the Sherman-Morrison formula, we get

$$\mathscr{J}_t^{-1} = \left(\mathscr{J}_{t-1} + \begin{bmatrix} p_t \\ 1 \end{bmatrix} \begin{bmatrix} p_t \\ 1 \end{bmatrix}^\top\right)^{-1} = \mathscr{J}_{t-1}^{-1} - \frac{\mathscr{J}_{t-1}^{-1} \begin{bmatrix} p_t \\ 1 \end{bmatrix} \begin{bmatrix} p_t \\ 1 \end{bmatrix}^\top \mathscr{J}_{t-1}^{-1}}{1 + \begin{bmatrix} p_t \\ 1 \end{bmatrix}^\top \mathscr{J}_{t-1}^{-1} \begin{bmatrix} p_t \\ 1 \end{bmatrix}}.$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\tilde{\varepsilon}_{t-1}^\top \mathscr{J}_t^{-1} \tilde{\varepsilon}_{t-1}\right] &= \mathbb{E}\left[\tilde{\varepsilon}_{t-1}^\top \mathscr{J}_{t-1}^{-1} \tilde{\varepsilon}_{t-1}\right] - \mathbb{E}\left[\frac{\left(\tilde{\varepsilon}_{t-1}^\top \mathscr{J}_{t-1}^{-1} \begin{bmatrix} p_t \\ 1 \end{bmatrix}\right)^2}{1 + \begin{bmatrix} p_t \\ 1 \end{bmatrix}^\top \mathscr{J}_{t-1}^{-1} \begin{bmatrix} p_t \\ 1 \end{bmatrix}}\right] \\ &\leq \mathbb{E}\left[\tilde{\varepsilon}_{t-1}^\top \mathscr{J}_{t-1}^{-1} \tilde{\varepsilon}_{t-1}\right], \end{aligned} \tag{A.8}$$

where the equality follows from the fact that the random variable in the second expectation is non-negative almost surely.

For the second term in Inequality (A.7), we have that

$$\begin{aligned}\begin{bmatrix} p_t \\ 1 \end{bmatrix}^\top \mathscr{J}_t^{-1} \begin{bmatrix} p_t \\ 1 \end{bmatrix} &= \frac{1}{J_t} \begin{bmatrix} p_t \\ 1 \end{bmatrix}^\top \begin{bmatrix} 1 & -\bar{p}_t \\ -\bar{p}_t & (1/t)\sum_{k=1}^t p_k^2 \end{bmatrix} \begin{bmatrix} p_t \\ 1 \end{bmatrix} \\ &= \frac{1}{J_t}\left((p_t - \bar{p}_t)^2 + \frac{1}{t} J_t\right) \\ &= \frac{J_t - J_{t-1}}{J_t} + \frac{1}{t}, \end{aligned} \tag{A.9}$$

where the last inequality follows from the fact that $J_t - J_{t-1} = (p_t - \bar{p}_t)^2$. Now using Inequalities (A.7), (A.8), and (A.9) we get

$$\mathbb{E}\left[\tilde{\varepsilon}_t^\top \mathscr{J}_t^{-1} \tilde{\varepsilon}_t\right] \leq \mathbb{E}\left[\tilde{\varepsilon}_{t-1}^\top \mathscr{J}_{t-1}^{-1} \tilde{\varepsilon}_{t-1}\right] + \left(\frac{J_t - J_{t-1}}{J_t} + \frac{1}{t}\right) \frac{(\overline{\varepsilon} - \underline{\varepsilon})^2}{4}.$$

By summing both sides of the above inequality from 3 to t we get

$$\mathbb{E}\left[\tilde{\varepsilon}_t^\top \mathscr{J}_t^{-1} \tilde{\varepsilon}_t\right] \leq \mathbb{E}\left[\tilde{\varepsilon}_2^\top \mathscr{J}_2^{-1} \tilde{\varepsilon}_2\right] + \frac{(\overline{\varepsilon} - \underline{\varepsilon})^2}{4} \mathbb{E}\left[\sum_{k=3}^t \left(\frac{J_k - J_{k-1}}{J_k} + \frac{1}{k}\right)\right].$$

It is straightforward to show that

$$\mathbb{E}\left[\tilde{\varepsilon}_2^\top \mathscr{J}_2^{-1} \tilde{\varepsilon}_2\right] = \mathbb{E}\left[\varepsilon_1^2 + \varepsilon_2^2\right] \leq \frac{1}{2}(\overline{\varepsilon} - \underline{\varepsilon})^2.$$

Note that $\sum_{k=3}^t (1/k) \leq \log(t)$. We also have that

$$\begin{aligned}\sum_{k=3}^t \frac{J_k - J_{k-1}}{J_k} &= \sum_{k=3}^t \int_{J_{k-1}}^{J_k} \frac{dx}{J_k} \\ &\leq \sum_{k=3}^t \int_{J_{k-1}}^{J_k} \frac{dx}{x} \\ &= \int_{J_2}^{J_t} \frac{dx}{x} \\ &\leq \log(J_t) \\ &\leq \log(t\overline{p}^2), \end{aligned}$$

where the last inequality follows from the fact that $(p_k - \bar{p}_t)^2 \leq \overline{p}^2$ almost surely.

Finally, we get

$$\begin{aligned}\mathbb{E}\left[\tilde{\varepsilon}_t^\top \mathscr{J}_t^{-1}\tilde{\varepsilon}_t\right] &\leq \frac{1}{2}(\overline{\varepsilon}-\underline{\varepsilon})^2\left(1+\log(\bar{p})+\log(t)\right)\\ &\leq \frac{1}{2}(\overline{\varepsilon}-\underline{\varepsilon})^2\left(2+\log(\bar{p})\right)\log(t),\end{aligned}$$

where the last inequality follows from the fact that $\log(t) \geq 1$ for $t \geq 3$. Finally, by applying the above inequality to the bound on the mean squared parameter estimation error (A.3), we get

$$\mathbb{E}\left[\|\theta_t-\theta\|^2\right] \leq \frac{1}{2}(1+\overline{p}^2)(\overline{\varepsilon}-\underline{\varepsilon})^2\left(2+\log(\bar{p})\right)\frac{\log(t)}{L_t}. \tag{A.10}$$

To complete the proof, we set $r = 1/4$. For this choice of r, we have that $L_t \geq \rho^2\sqrt{\lfloor t/2\rfloor}/8 \geq \rho^2\sqrt{t}/16$. Setting $\mu_2 := 8(1+\overline{p}^2)(\overline{\varepsilon}-\underline{\varepsilon})^2\left(2+\log(\bar{p})\right)$ concludes the proof.

A.2 Proof of Theorem 1

We introduce an additional assumption on the variation in the sequence of wholesale electricity prices.[1] Namely, let $\sigma \geq 0$ be nonnegative constant such that $|c_t - c_{t-1}| \geq \sigma$ for all $t \geq 1$. Ultimately, we will establish the desired result for $\sigma = 0$, the setting considered in the statement of the Theorem.

[1]Such assumption will prove useful in facilitating the proof of Theorem 2.

We begin with the following upper bound on the T-period regret.

$$\begin{aligned}
\Delta_T^{\pi} &= a \sum_{t=1}^{T} \mathbb{E}\left[(p_t - p_t^*)^2\right] \\
&\leq a \sum_{t=1}^{\lfloor \frac{T+1}{2} \rfloor} \mathbb{E}\left[(\widehat{p}_{2t-1} - p_{2t-1}^*)^2 + (\widehat{p}_{2t-1} - p_{2t-1}^* + \rho\delta_{2t})^2\right] \\
&\leq a \sum_{t=1}^{\lfloor \frac{T+1}{2} \rfloor} \left(3\mathbb{E}\left[(\widehat{p}_{2t-1} - p_{2t-1}^*)^2\right] + 2\rho^2\delta_{2t}^2\right) \\
&= K_0 + 2a\rho^2 \sum_{t=1}^{\lfloor \frac{T+1}{2} \rfloor} (2t)^{-2r} + 3a \sum_{t=1}^{\lfloor \frac{T+1}{2} \rfloor - 1} \mathbb{E}\left[(\widehat{p}_{2t+1} - p_{2t+1}^*)^2\right], \qquad \text{(A.11)}
\end{aligned}$$

where the second inequality follows from the fact that $x^2 + (x+y)^2 \leq 3x^2 + 2y^2$ for any pair of scalars $x, y \in \mathbb{R}$. Here, the constant K_0 is defined as

$$K_0 = 3a(p_1 - p_1^*)^2.$$

Recall that p_1 is assumed to be a deterministic constant. We now establish upper bounds on each term of the bound (A.11) separately.

Second term: For all $T \geq 3$, we have that

$$\begin{aligned}
\sum_{t=1}^{\lfloor \frac{T+1}{2} \rfloor} (2t)^{-2r} &\leq \int_0^{\lfloor \frac{T+1}{2} \rfloor} (2t)^{-2r} dt \\
&= \frac{1}{2(1-2r)} \left(2 \left\lfloor \frac{T+1}{2} \right\rfloor\right)^{1-2r} \\
&\leq \frac{2}{3(1-2r)} T^{1-2r}, \qquad \text{(A.12)}
\end{aligned}$$

where the last inequality follows from the fact that $(\frac{T+1}{T})^{1-2r} \leq 4/3$ for all $T \geq 3$ and all $r \in [0, 1/2)$.

Third term: Using the upper bound on the pricing error (2.14), we get

$$(\widehat{p}_{2t+1} - p_{2t+1}^*)^2 \leq 2\kappa_3^2 \|\widehat{\theta}_{2t} - \theta\|^2 + 2\kappa_2^2 (F_{2t}^{-1}(\alpha) - F^{-1}(\alpha))^2.$$

Then,

$$\sum_{t=1}^{\lfloor \frac{T+1}{2} \rfloor - 1} \mathbb{E}\left[(\widehat{p}_{2t+1} - p^*_{2t+1})^2\right] \leq \sum_{t=1}^{\lfloor \frac{T+1}{2} \rfloor - 1} \mathbb{E}\left[2\kappa_3^2 \|\widehat{\theta}_{2t} - \theta\|^2 + 2\kappa_2^2 (F_{2t}^{-1}(\alpha) - F^{-1}(\alpha))^2\right]$$
$$\leq \sum_{t=1}^{\lfloor \frac{T+1}{2} \rfloor - 1} \left(\kappa_4 \frac{\log(2t)}{L_{2t}} + 2\kappa_2^2 \mathbb{E}\left[(F_{2t}^{-1}(\alpha) - F^{-1}(\alpha))^2\right]\right), \tag{A.13}$$

where the last inequality follows from the upper bound (A.10) on the mean squared parameter estimation error and $\kappa_4 := (1 + \overline{p}^2)(\overline{\varepsilon} - \underline{\varepsilon})^2 (2 + \log(\bar{p}))\, \kappa_3^2$. Using the fact that for a continuous nonnegative random variable X, it holds that $\mathbb{E}[X] = \int_0^\infty \mathbf{P}\left(X \geq x\right) dx$, we get

$$\begin{aligned}
\mathbb{E}\left[(F_{2t}^{-1}(\alpha) - F^{-1}(\alpha))^2\right] &= \int_0^\infty \mathbf{P}\left((F_{2t}^{-1}(\alpha) - F^{-1}(\alpha))^2 \geq \gamma\right) d\gamma \\
&\leq \int_0^\infty 2\exp(-\mu_1 \gamma(2t)) d\gamma \\
&= \frac{1}{\mu_1 t},
\end{aligned} \tag{A.14}$$

where the inequality follows from the bound (2.9). By combining Inequalities (A.11), (A.12), (A.13), and (A.14), we get

$$\begin{aligned}
\Delta_T^\pi &\leq K_0 + \frac{4a}{3(1-2r)} \rho^2 T^{1-2r} + 3a \sum_{t=1}^{\lfloor \frac{T+1}{2} \rfloor - 1} \left(\kappa_4 \frac{\log(2t)}{L_{2t}} + \frac{2\kappa_2^2}{\mu_1 t}\right) \\
&\leq K_0 + \frac{4a}{3(1-2r)} \rho^2 T^{1-2r} + 24a\kappa_4 \sum_{t=1}^{\lfloor \frac{T+1}{2} \rfloor - 1} \frac{\log(2t)}{\rho^2 t^{1-2r} + \sigma^2 t} \\
&\quad + \frac{6a\kappa_2^2}{\mu_1}(1 + \log(T)),
\end{aligned} \tag{A.15}$$

where the last inequality follows from the definition of L_{2t} in Equation (A.2) and the assumption that $|c_t - c_{t-1}| \geq \sigma$ for all t. For $\sigma = 0$, it is straightforward to show that a choice of $r = 1/4$ minimizes the asymptotic order of the upper bound (A.15) with respect to the horizon T up to multiplicative logarithmic

factors. Setting $r = 1/4$ and $\sigma = 0$ yields

$$\Delta_T^\pi \leq K_1 + K_2 \log(T) + K_3 \rho^2 \sqrt{T} + \frac{K_4}{\rho^2} \sum_{t=1}^{\lfloor \frac{T+1}{2} \rfloor - 1} \frac{\log(2t)}{\sqrt{t}}, \tag{A.16}$$

where $K_1 := K_0 + K_2$, $K_2 := 6a\kappa_2^2/\mu_1$, $K_3 := 8a/3$, and $K_4 := 24a\kappa_4$. It holds that

$$\begin{aligned}
\sum_{t=1}^{\lfloor \frac{T+1}{2} \rfloor - 1} \frac{\log(2t)}{\sqrt{t}} &\leq \log(2) + \sum_{t=2}^{\lfloor \frac{T+1}{2} \rfloor - 1} \frac{\log(2t)}{\sqrt{t}} \\
&\leq \log(2) + \int_1^{T/2} \frac{\log(2t)}{\sqrt{t}} dt \\
&\leq \log(2) + 2\sqrt{\frac{T}{2}} \log(T).
\end{aligned}$$

Finally, we define the nonnegative constants C_0, C_1, and C_2 as follows to conclude the proof.

$$C_0 := K_1 + \frac{K_4 \log(2)}{\rho^2} \tag{A.17}$$

$$C_1 := \frac{\sqrt{2} K_4}{\rho^2} + K_3 \rho^2 \tag{A.18}$$

$$C_2 := K_2. \tag{A.19}$$

A.3 Proof of Theorem 2

Inequality (A.15) is a valid upper bound on the T-period regret incurred by perturbed myopic policy, under the assumption that $\sigma > 0$. By setting $\rho = 0$, the upper bound simplifies to

$$\Delta_T^\pi \leq K_0 + 24a\kappa_4 \sum_{t=1}^{\lfloor \frac{T+1}{2} \rfloor - 1} \frac{\log(2t)}{\sigma^2 t} + \frac{6a\kappa_2^2}{\mu_1}(1 + \log(T)),$$

It holds that

$$\sum_{t=1}^{T/2} \frac{\log(2t)}{t} \leq \log(2) + \int_1^{T/2} \frac{\log(2t)}{t} dt \leq \log(2) + \log^2(T).$$

We define the nonnegative constants M_0, M_1, and M_2 as follows to conclude the proof.

$$M_0 := K_0 + M_1 \tag{A.20}$$

$$M_1 := \frac{6a\kappa_2^2}{\mu_1} \tag{A.21}$$

$$M_2 := 24a\kappa_4. \tag{A.22}$$

A.4 Proof of Lemma 6

It is straightforward to show that the characteristic polynomial of $\mathscr{J}_t$ is given by

$$\lambda^2 - \lambda\left(t + \sum_{k=1}^{t} p_k^2\right) + tJ_t = 0.$$

Then,

$$\lambda_{\max}(\mathscr{J}_t) + \lambda_{\min}(\mathscr{J}_t) = t + \sum_{k=1}^{t} p_k^2,$$

$$\lambda_{\max}(\mathscr{J}_t)\lambda_{\min}(\mathscr{J}_t) = tJ_t.$$

From the first identity it follows that

$$\lambda_{\max}(\mathscr{J}_t) \leq t + \sum_{k=1}^{t} p_k^2 \leq t(1+\overline{p}^2).$$

Thus, we get

$$\lambda_{\min}(\mathscr{J}_t) = \frac{tJ_t}{\lambda_{\max}(\mathscr{J}_t)} \geq \frac{J_t}{1+\overline{p}^2}.$$

We now bound the random process $\{J_t\}$ from below by a deterministic sequence. Fix t. A direct substitution of the perturbed myopic policy yields

$$J_t \geq \sum_{k=1}^{\lfloor t/2 \rfloor} \left\{ \left(\widehat{p}_{2k-1} - \bar{p}_t\right)^2 + \left(\widehat{p}_{2k-1} - \bar{p}_t + \frac{1}{2}\left(c_{2k} - c_{2k-1}\right) + \rho\delta_{2k}\right)^2 \right\}.$$

The above inequality can be further relaxed to eliminate its explicit dependency on the (random) price process. Namely, it is straightforward to show that

$$J_t \geq \frac{1}{2} \sum_{k=1}^{\lfloor t/2 \rfloor} \frac{\rho^2}{(2k)^{2r}} + \frac{1}{8} \sum_{k=1}^{\lfloor t/2 \rfloor} \left(c_{2k} - c_{2k-1} \right)^2 . \tag{A.23}$$

One can further relax inequality (A.23) by using the facts that

$$\sum_{k=1}^{t} \frac{1}{k^{2r}} \geq \int_1^{t+1} \frac{1}{x^{2r}} dx = \frac{(t+1)^{1-2r} - 1}{1-2r},$$

and

$$(t+1)^{1-2r} - 1 \geq t^{1-2r} \left(1 - \frac{1}{2^{1-2r}} \right).$$

It follows that

$$J_t \geq \frac{\rho^2}{2^{1+2r}} \frac{\lfloor t/2 \rfloor^{1-2r}}{1-2r} \left(1 - \frac{1}{2^{1-2r}} \right) + \frac{1}{8} \sum_{k=1}^{\lfloor t/2 \rfloor} \left(c_{2k} - c_{2k-1} \right)^2 \geq L_t, \tag{A.24}$$

where L_t is defined as

$$L_t := \frac{1}{8} \left(\rho^2 \lfloor t/2 \rfloor^{1-2r} + \sum_{k=1}^{\lfloor t/2 \rfloor} \left(c_{2k} - c_{2k-1} \right)^2 \right).$$

APPENDIX B

PROOFS OF RESULTS IN CHAPTER 3

In this Appendix chapter, we provide a detailed proof of the theoretical results including Lemma 2, and Theorems 4 and 7 of Chapter 3. To facilitate the presentation of the mathematical proofs, we introduce the following notation. For all $t \geq 1$, define events $\mathcal{V}_t$, $\mathcal{S}_t$, and $\mathcal{Z}_t$ as

$$\begin{aligned}
\mathcal{V}_t &= \{\lambda_{\min}(V_t) \leq ct^{\nu}\}, \\
\mathcal{S}_t &= \left\{\text{LCB}_t(X_t^{\mathsf{CE}}) \leq b\right\}, \\
\mathcal{Z}_t &= \left\{N_t \geq \left\lceil \frac{ct^{\nu}}{\mu\rho^2\sigma^2} \right\rceil \right\}.
\end{aligned}$$

We denote by N_t the number of safe exploration stages until stage t for all $t \geq 1$.

B.1 Proof of Lemma 2

Recall from the definition of X_t^{S} that $\mathbf{P}\left(\langle X_t^{\mathsf{S}}, \theta^* \rangle \geq b_0\right) \geq 1 - \delta_t$. Thus, with probability $1 - \delta_t$, it holds that,

$$\begin{aligned}
\langle X_t^{\mathsf{SE}}, \theta^* \rangle &= \langle (1-\rho) X_t^{\mathsf{S}} + \rho U_t, \theta^* \rangle \\
&= \langle (1-\rho) X_t^{\mathsf{S}} + \rho \bar{x} + \rho H^{1/2} \zeta_t, \theta^* \rangle \\
&= \langle X_t^{\mathsf{S}}, \theta^* \rangle - \rho \langle X_t^{\mathsf{S}} - \bar{x}, \theta^* \rangle + \rho \langle H^{1/2} \zeta_t, \theta^* \rangle \\
&\geq b_0 - \rho \|X_t^{\mathsf{S}} - \bar{x}\| \|\theta^*\| - \rho \sqrt{\lambda_{\max}(H)} \|\theta^*\|, \qquad \text{(B.1)}
\end{aligned}$$

where the inequality follows from the Cauchy-Schwarz inequality and the fact that $\|\zeta_t\| = 1$. For any $x \in \mathcal{X}$, it holds that

$$\|x - \bar{x}\| \leq \|x - \bar{x}\|_{H^{-1}} \sqrt{\lambda_{\max}(H)} \leq \sqrt{\lambda_{\max}(H)}, \qquad \text{(B.2)}$$

where the inequality follows from the definition of $\mathcal{X}$ in Equation (3.1). By applying Inequality (B.2) to Inequality (B.1), with probability $1 - \delta_t$, it holds that

$$\langle X_t^{\mathsf{SE}}, \theta^* \rangle \geq b_0 - 2\rho\sqrt{\lambda_{\max}(H)}\|\theta^*\|.$$

Recall Assumption 2 that $\|\theta^*\| \leq S$. Thus, in order to guarantee that $\mathbf{P}\left(\langle X_t^{\mathsf{SE}}, \theta^* \rangle \geq b\right) \geq 1 - \delta_t$ it suffices to choose ρ such that

$$\rho \leq \frac{b_0 - b}{2S\sqrt{\lambda_{\max}(H)}}. \tag{B.3}$$

B.2 Proof of Theorem 4

From Lemma 2, it follows that the safe exploration arm X_t^{SE} is (δ_t, b)-safe by construction. Moreover, under the SEGE algorithm the greedy arm X_t^{CE} is only played if $\text{LCB}_t(X_t^{\mathsf{CE}}) \geq b$, which, in turn, implies

$$\mathbf{P}\left(\langle X_t^{\mathsf{CE}}, \theta^* \rangle \geq b\right) \geq 1 - \delta_t.$$

Thus, the greedy arm X_t^{CE} if played is (δ_t, b)-safe.

B.3 Proof of Theorem 5

We upper bound the expected regret during safe exploration stages and greedy exploitation stages separately. Recall the definition of expected regret R_T

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \langle X^* - X_t, \theta^* \rangle\right].$$

The expected regret can be decomposed into two parts,

$$R_T = \mathbb{E}\left[\sum_{t\in\{1,\dots,T\}\cap\{\mathcal{V}_t\cup\mathcal{S}_t\}} \langle X^* - X_t^{\mathsf{SE}}, \theta^*\rangle\right] + \mathbb{E}\left[\sum_{t\in\{1,\dots,T\}\cap\{\mathcal{V}_t^c\cap\mathcal{S}_t^c\}} \langle X^* - X_t^{\mathsf{CE}}, \theta^*\rangle\right]. \tag{B.4}$$

Expected regret during safe exploration stages: From the fact that $\|x\| \leq L$ for all $x \in \mathcal{X}$, Assumption 2 that $\|\theta^*\| \leq S$, and the Cauchy-Schwarz inequality it almost surely holds that $\langle X^* - X_t^{\mathsf{SE}}, \theta^*\rangle \leq 2LS$. Thus,

$$\mathbb{E}\left[\sum_{t\in\{1,\dots,T\}\cap\{\mathcal{V}_t\cup\mathcal{S}_t\}} \langle X^* - X_t^{\mathsf{SE}}, \theta^*\rangle\right] \leq 2LS\mathbb{E}\left[N_T\right].$$

Roughly speaking, the expected number of safe exploration stages grows linearly with the minimum eigenvalue of the information matrix. The following Lemma, establishes an upper bound on the expected number of safe exploration stages under the SEGE algorithm.

Lemma 7 (Safe Exploration Stages)**.** Let $\{\delta_t\}_{t=1}^{\infty}$ be any sequence of risk levels satisfying Inequality (3.18) for all $t \geq 1$. Let $c > 2dKL^2\sigma_\eta^2/(b_0 - b)^2$ and $\nu \in [\overline{\nu}, 1)$. Then, there exists a finite positive constant C_0 such that under the SEGE Algorithm 1, it holds that

$$\mathbb{E}\left[N_t\right] \leq C_0 t^\nu,$$

for all $t \geq 1$.

The proof of Lemma 7 is postponed to Appendix B.4. Using Lemma 7, the expected regret during the safe exploration stages is upper bound as

$$\mathbb{E}\left[\sum_{t\in\{1,\dots,T\}\cap\{\mathcal{V}_t\cup\mathcal{S}_t\}} \langle X^* - X_t^{\mathsf{SE}}, \theta^*\rangle\right] \leq C_1 T^\nu, \tag{B.5}$$

where $C_1 := 2LSC_0$.

Expected regret during greedy exploitation stages: Recall that the greedy arm is only played if $\text{LCB}_t(X_t^{\mathsf{CE}}) \geq b$ and $\lambda_{\min}(V_t) \geq ct^{\nu}$, i.e., the event $\mathcal{V}_t^c \cap \mathcal{S}_t^c$ occurs. Moreover, it almost surely holds that $\langle x, \theta^* \rangle \leq LS$ for all $x \in \mathcal{X}$. Then, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in\{1,\ldots,T\}\cap\{\mathcal{V}_t^c\cap\mathcal{S}_t^c\}} \langle X^* - X_t^{\mathsf{CE}}, \theta^*\rangle\right] \\
&\quad = \sum_{t=1}^{T} \int_0^{2LS} \mathbf{P}\left(\left\{\langle X^* - X_t^{\mathsf{CE}}, \theta^*\rangle \geq \gamma\right\} \cap \mathcal{V}_t^c \cap \mathcal{S}_t^c\right)\, d\gamma \\
&\quad \leq \sum_{t=1}^{T} \int_0^{2LS} \mathbf{P}\left(\left\{\langle X^* - X_t^{\mathsf{CE}}, \theta^*\rangle \geq \gamma\right\} \cap \mathcal{V}_t^c\right)\, d\gamma. \qquad \text{(B.6)}
\end{aligned}$$

To bound the integrand in Inequality (B.6), we first establish an upper bound on the stagewise regret under the greedy arm in terms of the parameter estimation error.

Lemma 8 (Stagewise Regret)**.** The conditional expected reward given the greedy (certainty equivalent) arm X_t^{CE} is almost surely lower bounded as

$$\langle X_t^{\mathsf{CE}}, \theta^*\rangle \geq \langle X^*, \theta^*\rangle - k_1 \left\|\theta^* - \widehat{\theta}_{t-1}\right\|^2 \qquad \text{(B.7)}$$

for all $t \geq 1$, where the constant k_1 is given by

$$k_1 = \frac{2\|X_0\|\lambda_{\max}(H)}{b_0\sqrt{\lambda_{\min}(H)}}.$$

The proof of Lemma 8 is postponed to Appendix B.5. By applying Inequality

(B.7) to (B.6), we get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\in\{1,\ldots,T\}\cap\{\mathcal{V}_t^c\cap\mathcal{S}_t^c\}}\langle X^* - X_t^{\mathsf{CE}}, \theta^*\rangle\right] \\
&\leq \sum_{t=1}^{T}\int_0^{2LS}\mathbf{P}\left(\left\{k_1\|\theta^* - \widehat{\theta}_{t-1}\|^2 \geq \gamma\right\}\cap\mathcal{V}_t^c\right)\,d\gamma \\
&\leq \sum_{t=1}^{T}\int_0^{2LS}\mathbf{P}\left(\|\theta^* - \widehat{\theta}_{t-1}\|_{V_{t-1}}^2 \geq \frac{\gamma c t^\nu}{k_1}\right)\,d\gamma, \qquad \text{(B.8)}
\end{aligned}$$

where the last inequality follows from the fact that $\|\theta^* - \widehat{\theta}_{t-1}\|_{V_{t-1}} \geq \|\theta^* - \widehat{\theta}_{t-1}\|\sqrt{\lambda_{\min}(V_{t-1})}$. We now utilize Theorem 3 to bound the integral in Inequality (B.8). More precisely, we define a parameter $\delta_t^\dagger(\gamma)$ for which it holds that $\gamma c t^\nu/k_1 = r_t^2(\delta_t^\dagger(\gamma))$, i.e.,

$$\delta_t^\dagger(\gamma) = \left(1 + \frac{tL^2}{\lambda}\right)\exp\left(-\frac{1}{d\sigma_\eta^2 k_1}\left(\sqrt{\gamma c t^\nu} - \sqrt{k\lambda}S\right)^2\right). \qquad \text{(B.9)}$$

Define γ_{t-1} as

$$\gamma_{t-1} = k_7\frac{\log(t)}{ct^\nu}, \qquad \text{(B.10)}$$

where k_7 is defined as

$$k_7 = 2k_1 d\sigma_\eta^2\left(\log(1 + L^2/\lambda) + \frac{2\lambda S^2}{d\sigma_\eta^2}\right). \qquad \text{(B.11)}$$

We then have that

$$\begin{aligned}
&\int_0^{2LS}\mathbf{P}\left(\|\theta^* - \widehat{\theta}_{t-1}\|_{V_{t-1}}^2 \geq \frac{\gamma c t^\nu}{k_1}\right)\,d\gamma \\
&\leq \gamma_{t-1} + \int_{\gamma_{t-1}}^{2LS}\mathbf{P}\left(\|\theta^* - \widehat{\theta}_{t-1}\|_{V_{t-1}}^2 \geq \frac{\gamma c t^\nu}{k_1}\right)\,d\gamma \\
&= \gamma_{t-1} + \int_{\gamma_{t-1}}^{2LS}\mathbf{P}\left(\|\theta^* - \widehat{\theta}_{t-1}\|_{V_{t-1}} \geq r_t(\delta_t^\dagger(\gamma))\right)\,d\gamma \\
&\leq \gamma_{t-1} + \int_{\gamma_{t-1}}^{2LS}\delta_t^\dagger(\gamma)\,d\gamma. \qquad \text{(B.12)}
\end{aligned}$$

We now establish an upper bound on the integral in Inequality (B.12). Using the fact that for any two real numbers $x, y > 0$, we have that $(\sqrt{x} - \sqrt{y})^2 \geq x/2 - y$,

we get

$$\delta_t^\dagger(\gamma) \leq (1 + tL^2/\lambda) \exp\left(\frac{2\lambda S^2}{d\sigma_\eta^2}\right) \exp\left(-\frac{ct^\nu}{2k_1 d\sigma_\eta^2}\gamma\right).$$

Then,

$$\begin{aligned}
\int_{\gamma_{t-1}}^{2LS} \delta_t^\dagger(\gamma)\, d\gamma &\leq (1 + tL^2/\lambda) \exp\left(\frac{2\lambda S^2}{d\sigma_\eta^2}\right) \int_{\gamma_{t-1}}^{2LS} \exp\left(-\frac{ct^\nu}{2k_1 d\sigma_\eta^2}\gamma\right)\, d\gamma \\
&\leq (1 + tL^2/\lambda) \exp\left(\frac{2\lambda S^2}{d\sigma_\eta^2}\right) \exp\left(-\frac{ct^\nu}{2k_1 d\sigma_\eta^2}\gamma_{t-1}\right) \frac{2k_1 d\sigma_\eta^2}{ct^\nu} \\
&\leq \frac{2k_1 d\sigma_\eta^2}{ct^\nu}, \qquad \text{(B.13)}
\end{aligned}$$

where the last inequality follows from the definition of γ_{t-1}. By applying Inequalities (B.12) and (B.13) to (B.8), we get

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\in\{1,\ldots,T\}\cap\{\mathcal{V}_t^c\cap\mathcal{S}_t^c\}} \langle X^* - X_t^{\mathsf{CE}}, \theta^*\rangle\right] &\leq \sum_{t=1}^{T}\left(k_7\frac{\log(t)}{ct^\nu} + \frac{2k_1 d\sigma_\eta^2}{ct^\nu}\right) \\
&\leq \frac{1}{c}(k_7\log(T) + 2k_1 d\sigma^2)\sum_{t=1}^{T} t^{-\nu} \\
&\leq C_1 \log(T) T^{1-\nu},
\end{aligned}$$

where C_1 is defined as

$$C_1 = \frac{1}{c(1-\overline{\nu})}(k_7 + 2k_1 d\sigma^2).$$

B.4 Proof of Lemma 7

From the definition of N_t, for $t \geq 0$ we have,

$$N_{t+1} = \begin{cases} N_t, & \lambda_{\min}(V_t) \geq ct^\nu \text{ and } \mathrm{LCB}_{t+1}(X_{t+1}^{\mathsf{CE}}) \geq b, \\ N_t + 1, & \text{otherwise}, \end{cases} \qquad \text{(B.14)}$$

where $N_0 = 0$.

Fix $\mu \in (0, 1)$. Define the random process $\{Z_t\}_{t=1}^{\infty}$ as follows. For any $t \geq 0$, Z_t is defined as

$$Z_t = 0 \vee \left(N_t - \left\lceil \frac{ct^{\nu}}{\mu \rho^2 \sigma^2} \right\rceil \right). \tag{B.15}$$

We show that $\mathbb{E}[Z_t]$ is finite for all $t \geq 0$ and thus establish an $O(t^{\nu})$ upper bound on expected number of safe exploration stages until stage t. Note that conditioned on the event $\mathcal{Z}_t^c$, we have $Z_{t+1} = 0$. Thus,

$$\begin{aligned}
\mathbb{E}[Z_{t+1}] &= \mathbb{E}[Z_t] + \mathbb{E}[\mathbb{1}\{\mathcal{Z}_t \cap (\mathcal{V}_t \cup \mathcal{S}_{t+1})\}] \\
&= \mathbb{E}[Z_t] + \mathbf{P}(\mathcal{Z}_t \cap (\mathcal{V}_t \cup \mathcal{S}_{t+1})) \\
&= \sum_{k=1}^{t} \mathbf{P}(\mathcal{Z}_k \cap (\mathcal{V}_k \cup \mathcal{S}_{k+1})).
\end{aligned} \tag{B.16}$$

Thus, in order to establish an upper bound on $\mathbb{E}[Z_{t+1}]$, it suffices to establish an upper bound on $\mathbf{P}(\mathcal{Z}_t \cap (\mathcal{V}_{t-1} \cup \mathcal{S}_t))$ for all $t \geq 0$. Using the union bound, it holds that

$$\begin{aligned}
\mathbf{P}(\mathcal{Z}_t \cap (\mathcal{V}_t \cup \mathcal{S}_{t+1})) &= \mathbf{P}\left(\mathcal{Z}_t \cap \left(\mathcal{V}_t \cup (\mathcal{S}_{t+1} \cap \mathcal{V}_{t-1}^c)\right)\right) \\
&\leq \mathbf{P}(\mathcal{Z}_t \cap \mathcal{V}_t) + \mathbf{P}(\mathcal{S}_{t+1} \cap \mathcal{V}_t^c).
\end{aligned} \tag{B.17}$$

We upper bound each term in inequality (B.17), separately.

Fist term: Qualitatively, we gain more information as we play the safe exploration arm more frequently. More specifically, we expect $\lambda_{\min}(V_t)$ not to be too small if N_t is large. The following Lemma quantifies the relationship between these two random variables.

Lemma 9 (Random Exploration)**.** Under the SEGE Algorithm, for any $\mu \in (0,1)$ it holds that

$$\mathbf{P}\left(\lambda_{\min}(V_t) \leq \mu\rho^2\sigma^2 N_t \mid N_t = n\right) \leq de^{-k_4(1-\mu)^2 n}, \tag{B.18}$$

where k_4 is defined as

$$k_4 = \frac{\rho^4\sigma^4}{2(2\rho((1-\rho)L + \rho\|\bar{x}\|)\sqrt{\lambda_{\max}(H)} + \rho^2\lambda_{\max}(H) - \rho^2\sigma^2 d)^2}.$$

We postpone the proof of Lemma 9, which relies on the Matrix Azuma-Hoeffding inequality (Theorem 8) to Appendix B.6. Using Lemma 9 and total probability theorem, we have

$$\begin{aligned}
\mathbf{P}\left(\mathcal{Z}_t \cap \mathcal{V}_t\right) &= \sum_{n=\left\lceil \frac{ct^\nu}{\mu\rho^2\sigma^2}\right\rceil}^{\infty} \mathbf{P}\left(\lambda_{\min}(V_t) \leq ct^\nu | N_t = n\right) \mathbf{P}\left(N_t = n\right) \\
&\leq \sum_{n=\left\lceil \frac{ct^\nu}{\mu\rho^2\sigma^2}\right\rceil}^{\infty} d\exp\left(-k_4\left(1 - \frac{ct^\nu}{\rho^2\sigma^2 n}\right)^2 n\right) \mathbf{P}\left(N_t = n\right) \\
&\leq d\exp\left(-\frac{k_4\left(1-\mu\right)^2}{\mu\rho^2\sigma^2} ct^\nu\right) \sum_{n=\left\lceil \frac{ct^\nu}{\mu\rho^2\sigma^2}\right\rceil}^{\infty} \mathbf{P}\left(N_t = n\right) \\
&\leq d\exp\left(-\frac{k_4\left(1-\mu\right)^2}{\mu\rho^2\sigma^2} ct^\nu\right).
\end{aligned} \tag{B.19}$$

Second term: We establish a lower bound on $\mathbf{P}\left(\mathcal{S}_{t+1} \cap \mathcal{V}_t^c\right)$. It holds that

$$\mathrm{LCB}_t(X_{t+1}^{\mathsf{CE}}) = \langle X_{t+1}^{\mathsf{CE}}, \widehat{\theta}_t\rangle - r_t(\delta_t)\|X_{t+1}^{\mathsf{CE}}\|_{V_t^{-1}} \geq \langle X_{t+1}^{\mathsf{CE}}, \widehat{\theta}_t\rangle - \frac{L}{\sqrt{\lambda_{\min}(V_t)}} r_t(\delta_t).$$

It holds that

$$\begin{aligned}\langle X_{t+1}^{\mathsf{CE}}, \widehat{\theta}_t\rangle &= \langle X_{t+1}^{\mathsf{CE}} - X^*, \widehat{\theta}_t\rangle + \langle X^*, \widehat{\theta}_t - \theta^*\rangle + \langle X^*, \theta^*\rangle \\ &\geq \langle X^*, \widehat{\theta}_t - \theta^*\rangle + \langle X^*, \theta^*\rangle \\ &\geq -\frac{L}{\sqrt{\lambda_{\min}(V_t)}}\|\widehat{\theta}_t - \theta^*\|_{V_t} + \langle X^*, \theta^*\rangle \\ &\geq b_0 - \frac{L}{\sqrt{\lambda_{\min}(V_t)}}\|\widehat{\theta}_t - \theta^*\|_{V_t},\end{aligned}$$

where the first inequality follows from the fact that X_{t+1}^{CE} is the optimal arm for reward parameter $\widehat{\theta}_t$ and the last inequality follows from the fact that $\langle X^*, \theta^*\rangle \geq \langle X_0, \theta^*\rangle \geq b_0$. Then,

$$\mathrm{LCB}_t(X_{t+1}^{\mathsf{CE}}) \geq b_0 - \frac{L}{\sqrt{\lambda_{\min}(V_t)}}\|\widehat{\theta}_t - \theta^*\|_{V_t} - \frac{L}{\sqrt{\lambda_{\min}(V_t)}} r_t(\delta_t).$$

Then,

$$\begin{aligned}\mathbf{P}\left(\mathcal{S}_{t+1} \mid \mathcal{V}_t^c\right) &= \mathbf{P}\left(\|\widehat{\theta}_t - \theta^*\|_{V_t} \geq \frac{b_0 - b}{L}\sqrt{\lambda_{\min}(V_t)} - r_t(\delta_t) \;\middle|\; \mathcal{V}_t^c\right) \\ &\leq \mathbf{P}\left(\|\widehat{\theta}_t - \theta^*\|_{V_t} \geq \frac{b_0 - b}{L}\sqrt{ct^\nu} - r_t(\delta_t) \;\middle|\; \mathcal{V}_t^c\right)\end{aligned}$$

Then, using the fact that $\delta_t \geq \overline{\delta}e^{-Kt^{\overline{\nu}}} \geq \overline{\delta}e^{-Kt^\nu}$, we get

$$\mathbf{P}\left(\mathcal{S}_{t+1} \mid \mathcal{V}_t^c\right) \leq \mathbf{P}\left(\|\widehat{\theta}_t - \theta^*\|_{V_t} \geq \frac{b_0 - b}{L}\sqrt{ct^\nu} - r_t\left(\overline{\delta}e^{-Kt^\nu}\right) \;\middle|\; \mathcal{V}_t^c\right).$$

In order to utilize Theorem 3, we define $\tilde{\delta}_t$ as

$$\tilde{\delta}_t = \exp\left(\frac{4\lambda S^2}{\sigma_\eta^2 d}\right)\frac{(1 + tL^2/\lambda)^2}{\overline{\delta}}\exp\left(-\left(\frac{(b_0 - b)^2 c}{2L^2\sigma_\eta^2 d} - K\right)t^\nu\right). \tag{B.20}$$

It is straightforward to verify that

$$r_t(\tilde{\delta}_t) \leq \frac{b_0 - b}{L}\sqrt{ct^\nu} - r_t\left(\overline{\delta}e^{-Kt}\right).$$

Thus, using Theorem 3, we get

$$\mathbf{P}\left(\mathcal{S}_{t+1} \cap \mathcal{V}_t^c\right) \leq \mathbf{P}\left(\|\widehat{\theta}_t - \theta^*\|_{V_t} \geq r_t(\tilde{\delta}_t)\right) \leq \tilde{\delta}_t. \tag{B.21}$$

By applying Inequalities (B.17), (B.19), and (B.21) to Inequality (B.16), we get

$$\mathbb{E}\left[Z_{t+1}\right] \leq b \sum_{k=1}^{t} \exp\left(-\frac{k_4 \left(1-\mu\right)^2}{\mu\rho^2\sigma^2} c k^{\nu}\right) + \exp\left(\frac{4\lambda S^2}{\sigma_\eta^2 d}\right) \sum_{k=1}^{t} \frac{(1+kL^2/\lambda)^2}{\overline{\delta}} \exp\left(-\left(\frac{(b_0-b)^2 c}{2L^2\sigma_\eta^2 d} - K\right) k^{\nu}\right).$$

Note that, using the integral test, one can verify that $\sum_{k=1}^{\infty} k^2 \exp(-sk^r)$ converges for all $s > 0$ and $r > 0$. Thus, using the fact that $K < (b_0 - b)^2 c/(2L^2\sigma_\eta^2 d)$, there exists a finite k_6 such that

$$\mathbb{E}\left[Z_t\right] \leq k_6$$

for all $t \geq 1$. Thus,

$$\mathbb{E}\left[N_t\right] \leq \mathbb{E}\left[Z_t\right] + \left\lceil \frac{ct^{\nu}}{\mu\rho^2\sigma^2} \right\rceil \leq k_6 + 1 + \frac{ct^{\nu}}{\mu\rho^2\sigma^2}$$

for all $t \geq 1$. Defining $C_0 := k_6 + 1 + c/(\mu\rho^2\sigma^2)$ concludes the proof.

B.5 Proof of Lemma 8

For each $t \geq 1$, the expected reward under the greedy arm is lower bounded as

$$\begin{aligned}
\langle X_t^{\mathsf{CE}}, \theta^* \rangle &= \langle X^*, \theta^* \rangle - \langle X^* - X_t^{\mathsf{CE}}, \theta^* \rangle \\
&= \langle X^*, \theta^* \rangle - \langle X^* - X_t^{\mathsf{CE}}, \theta^* - \widehat{\theta}_{t-1} \rangle - \langle X^* - X_t^{\mathsf{CE}}, \widehat{\theta}_{t-1} \rangle \\
&\geq \langle X^*, \theta^* \rangle - \langle X^* - X_t^{\mathsf{CE}}, \theta^* - \widehat{\theta}_{t-1} \rangle,
\end{aligned}$$

where the inequality follows from the fact that $\langle X_t^{\mathsf{CE}}, \widehat{\theta}_{t-1} \rangle \geq \langle x, \widehat{\theta}_{t-1} \rangle$ for all $x \in \mathcal{X}$ as the greedy arm is the optimal arm for the reward parameter $\widehat{\theta}_{t-1}$. Using the Cauchy-Schwarz inequality, we get

$$\begin{aligned}
\langle X_t^{\mathsf{CE}}, \theta^* \rangle &\geq \langle X^*, \theta^* \rangle - \|X^* - X_t^{\mathsf{CE}}\| \|\theta^* - \widehat{\theta}_{t-1}\| \\
&= \langle X^*, \theta^* \rangle - \left\| \frac{H\theta^*}{\|\theta^*\|_H} - \frac{H\widehat{\theta}_{t-1}}{\|\widehat{\theta}_{t-1}\|_H} \right\| \|\theta^* - \widehat{\theta}_{t-1}\|. \qquad \text{(B.22)}
\end{aligned}$$

We now show that

$$\left\| \frac{H\theta^*}{\|\theta^*\|_H} - \frac{H\widehat{\theta}_{t-1}}{\|\widehat{\theta}_{t-1}\|_H} \right\| \leq \frac{2\|X_0\|\lambda_{\max}(H)}{b_0\sqrt{\lambda_{\min}(H)}}\|\theta^* - \widehat{\theta}_{t-1}\|. \tag{B.23}$$

Using the triangle inequality, we get

$$\begin{aligned}
\left\| \frac{H\theta^*}{\|\theta^*\|_H} - \frac{H\widehat{\theta}_{t-1}}{\|\widehat{\theta}_{t-1}\|_H} \right\| &= \left\| \frac{H\theta^*}{\|\theta^*\|_H} - \frac{H\widehat{\theta}_{t-1}}{\|\theta^*\|_H} + \frac{H\widehat{\theta}_{t-1}}{\|\theta^*\|_H} - \frac{H\widehat{\theta}_{t-1}}{\|\widehat{\theta}_{t-1}\|_H} \right\| \\
&\leq \left\| \frac{H\theta^*}{\|\theta^*\|_H} - \frac{H\widehat{\theta}_{t-1}}{\|\theta^*\|_H} \right\| + \left\| \frac{H\widehat{\theta}_{t-1}}{\|\theta^*\|_H} - \frac{H\widehat{\theta}_{t-1}}{\|\widehat{\theta}_{t-1}\|_H} \right\| \\
&= \frac{1}{\|\theta^*\|_H}\|H(\theta^* - \widehat{\theta}_{t-1})\| + \frac{\|H\widehat{\theta}_{t-1}\|}{\|\theta^*\|_H\|\widehat{\theta}_{t-1}\|_H}\left|\|\theta^*\|_H - \|\widehat{\theta}_{t-1}\|_H\right| \\
&\leq \frac{1}{\|\theta^*\|_H}\|H(\theta^* - \widehat{\theta}_{t-1})\| + \frac{\|H\widehat{\theta}_{t-1}\|}{\|\theta^*\|_H\|\widehat{\theta}_{t-1}\|_H}\|\theta^* - \widehat{\theta}_{t-1}\|_H,
\end{aligned}$$

where the last inequality follows from the reverse triangle inequality. Using the Cauchy-Schwarz inequality and the fact that $\|H\| = \lambda_{\max}(H)$, we get

$$\begin{aligned}
&\frac{1}{\|\theta^*\|_H}\|H(\theta^* - \widehat{\theta}_{t-1})\| + \frac{\|H\widehat{\theta}_{t-1}\|}{\|\theta^*\|_H\|\widehat{\theta}_{t-1}\|_H}\|\theta^* - \widehat{\theta}_{t-1}\|_H \\
&\leq \frac{\lambda_{\max}(H)}{\|\theta^*\|_H}\|\theta^* - \widehat{\theta}_{t-1}\| + \frac{\sqrt{\lambda_{\max}(H)}\|\widehat{\theta}_{t-1}\|_H}{\|\theta^*\|_H\|\widehat{\theta}_{t-1}\|_H}\sqrt{\lambda_{\max}(H)}\|\theta^* - \widehat{\theta}_{t-1}\| \\
&= \frac{2\lambda_{\max}(H)}{\|\theta^*\|_H}\|\theta^* - \widehat{\theta}_{t-1}\|,
\end{aligned} \tag{B.24}$$

where the inequality follows from the fact that $\|H\widehat{\theta}_{t-1}\| \leq \|H^{1/2}\|\|H^{1/2}\widehat{\theta}_{t-1}\| = \sqrt{\lambda_{\max}(H)}\|\widehat{\theta}_{t-1}\|_H$. Recall from Assumption 4 that $\langle X_0, \theta^* \rangle \geq b_0$. So, $\|X_0\|\|\theta^*\| \geq b_0$

$$\|\theta^*\|_H \geq \frac{\sqrt{\lambda_{\min}(H)}b_0}{\|X_0\|}. \tag{B.25}$$

Thus, by applying Inequality (B.25) to (B.24), we get Inequality (B.23). Finally, combining Inequalities (B.22) and (B.23) yields the desired lower bound on the expected reward of the greedy arm.

B.6 Proof of Lemma 9

Let $\mathcal{N}_t^{\mathsf{SE}}$ be the set of stages in which a safe exploration arm is played up to and including stage t. Our objective is to establish a lower bound on the minimum eigenvalue of V_t in terms on N_t. As $yy^\top$ is a positive semidefinite matrix for any $y \in \mathbb{R}^d$, it holds that

$$\begin{aligned}
V_t &= \lambda I + \sum_{k=1}^{t} X_t X_t^\top \\
&\succeq \sum_{k \in \mathcal{N}_t^{\mathsf{SE}}} X_k^{\mathsf{SE}} X_k^{\mathsf{SE}\top} \\
&= \sum_{k \in \mathcal{N}_t^{\mathsf{SE}}} \left(((1-\rho)X_k^{\mathsf{S}} + \rho\bar{x} + \rho H^{1/2}\zeta_k)((1-\rho)X_k^{\mathsf{S}} + \rho\bar{x} + \rho H^{1/2}\zeta_k)^\top \right) \\
&\succeq \sum_{k \in \mathcal{N}_t^{\mathsf{SE}}} \Big(((1-\rho)X_k^{\mathsf{S}} + \rho\bar{x})(\rho H^{1/2}\zeta_k)^\top \\
&\qquad\qquad + \rho H^{1/2}\zeta_k((1-\rho)X_k^{\mathsf{S}} + \rho\bar{x})^\top + \rho^2 H^{1/2}\zeta_k\zeta_k^\top H^{1/2} \Big) \\
&= \sum_{k \in \mathcal{N}_t^{\mathsf{SE}}} \left(\rho^2 H^{1/2}\mathbb{E}\left[\zeta_k\zeta_k^\top\right] H^{1/2} + W_k \right) \\
&\succeq \sum_{k \in \mathcal{N}_t^{\mathsf{SE}}} \left(\rho^2 \lambda_{\min}\left(H^{1/2}\mathbb{E}\left[\zeta_k\zeta_k^\top\right] H^{1/2} \right) + W_k \right),
\end{aligned}$$

where W_k is defined as

$$\begin{aligned}
W_k :=& ((1-\rho)X_k^{\mathsf{S}} + \rho\bar{x})(\rho H^{1/2}\zeta_k)^\top \\
&+ \rho H^{1/2}\zeta_k((1-\rho)X_k^{\mathsf{S}} + \rho\bar{x})^\top + \rho^2 H^{1/2}(\zeta_k\zeta_k^\top - \mathbb{E}\left[\zeta_k\zeta_k^\top\right])H^{1/2}. \qquad \text{(B.26)}
\end{aligned}$$

Recall that σ^2 is defined as the minimum eigenvalue of the covariance matrix of U_k, i.e.,

$$\sigma^2 = \lambda_{\min}\left(\mathbb{E}\left[(U_k - \mathbb{E}\left[U_k\right])\left(U_k - \mathbb{E}\left[U_k\right]\right)^\top \right] \right) = \lambda_{\min}\left(H^{1/2}\mathbb{E}\left[\zeta_k\zeta_k^\top\right] H^{1/2} \right).$$

Thus, using the fact that $|\mathcal{N}_t^{\mathsf{SE}}| = N_t$, we get

$$V_t \succeq \rho^2 \sigma^2 N_t I + \sum_{k \in \mathcal{N}_t^{\mathsf{SE}}} W_k.$$

Using Weyl's inequality, it immediately follows that

$$\lambda_{\min}(V_t) \geq \rho^2 \sigma^2 N_t - \lambda_{\max} \left(\sum_{k \in \mathcal{N}_t^{\mathsf{SE}}} W_k \right). \tag{B.27}$$

We rely on the Matrix Azuma Inequality (B.28) to establish an upper bound on $\lambda_{\max}(\sum_{k \in \mathcal{N}_t} W_k)$, which holds with high probability.

Theorem 8 (Matrix Azuma Inequality)**.** [89, Theorem 7.1. and Remark 7.8.] Let $\{\mathcal{F}_k\}_{k=0}^{\infty}$ be a filtration. Consider the random process $\{Y_k\}_{k=1}^{\infty}$ adapted to the filtration $\{\mathcal{F}_k\}_{k=1}^{\infty}$. Each Y_k is a self-adjoint matrix with dimension d such that

$$\mathbb{E}\left[Y_k \mid \mathcal{F}_{k-1}\right] = 0 \text{ for } k = 1, 2, 3, \ldots,$$

and

$$Y_k^2 \preceq A_k^2 \text{ almost surely for } k = 1, 2, 3, \ldots,$$

where $\{A_k\}_{k=1}^{\infty}$ is a sequence of deterministic matrices. Moreover, the sequence $\{Y_k\}_{k=1}^{\infty}$ is conditionally symmetric, i.e., $Y_k \sim -Y_k$ conditional on $\mathcal{F}_{k-1}$. Then, for all $\delta \geq 0$ and $t \geq 1$, it holds that

$$\mathbf{P}\left(\lambda_{\max}\left(\sum_{k=1}^{t} Y_k \right) \geq \delta \right) \leq d \cdot \exp\left(- \frac{\delta^2}{2 \left\| \sum_{k=1}^{t} A_k^2 \right\|} \right). \tag{B.28}$$

In order to apply the Matrix Azuma Inequality (B.28), we first show that the sequence of random matrices $\{W_k\}_{k=1}^{\infty}$ satisfy the assumptions of Theorem 8. From the definition of W_k in Equation (B.26), it follows that $W_k = W_k^{\top}$ for all $k \geq 1$. Define the filtration $\mathcal{F}_k = \sigma(X_1^{\mathsf{S}}, \ldots, X_{k+1}^{\mathsf{S}}, \zeta_1, \ldots, \zeta_k)$ for all $k \geq 1$. It immediately follows that W_k is $\mathcal{F}_k$-measurable, conditionally symmetric, and

$\mathbb{E}[W_k \mid \mathcal{F}_{k-1}] = 0$. We now construct the sequence of deterministic matrices $\{A_k\}_{k=1}^{\infty}$ such that it almost surely holds that $W_k^2 \preceq A_k^2$. Using the fact that the trace of a matrix is equal to the sum of its eigenvalues, it almost surely holds that $\lambda_{\max}(W_k) \leq \text{trace}(W_k)$. Then,

$$\begin{aligned}\lambda_{\max}(W_k) &\leq 2((1-\rho)X_k^{\mathsf{S}} + \rho\bar{x})^\top(\rho H^{1/2}\zeta_k) \\ &\quad + \rho^2\zeta_k^\top H\zeta_k - \rho^2\text{trace}\left(H^{1/2}\mathbb{E}\left[\zeta_k\zeta_k^\top\right]H^{1/2}\right) \\ &\leq 2((1-\rho)X_k^{\mathsf{S}} + \rho\bar{x})^\top(\rho H^{1/2}\zeta_k) + \rho^2\lambda_{\max}(H) - \rho^2\sigma^2 d,\end{aligned}$$

where the inequality follows from the fact that $\|\zeta_k\| = 1$ for all $k \geq 1$ and the definition of σ^2. Using the fact that $\|X_k^{\mathsf{S}}\| \leq L$, and the Cauchy-Schwarz inequality, it almost surely holds that

$$\lambda_{\max}(W_k) \leq k_3,$$

where k_3 is defined as

$$k_3 = 2\rho((1-\rho)L + \rho\|\bar{x}\|)\sqrt{\lambda_{\max}(H)} + \rho^2\lambda_{\max}(H) - \rho^2\sigma^2 d.$$

Define $A_k = k_3 I$ for all $k \geq 1$. Then, it almost surely holds that $W_k^2 \preceq \lambda_{\max}(W_k)^2 I \preceq A_k^2$ for all $k \geq 1$. Thus, the sequence of random matrices $\{W_k\}_{k=1}^{\infty}$ satisfies all the assumptions of Theorem 8. Using the Cauchy-Schwarz inequality, we get

$$\left\|\sum_{k\in\mathcal{N}_t^{\mathsf{SE}}} A_k^2\right\| \leq \sum_{k\in\mathcal{N}_t^{\mathsf{SE}}} \left\|A_k^2\right\| \leq N_t k_3^2.$$

Using the Matrix Azuma Inequality (B.28), for any $\delta \geq 0$, it holds that

$$\mathbf{P}\left(\lambda_{\max}\left(\sum_{k\in\mathcal{N}_t^{\mathsf{SE}}} W_k\right) \geq \delta \;\middle|\; N_t = n\right) \leq d \cdot \exp\left(-\frac{\delta^2}{2nk_3^2}\right).$$

By setting $\delta = (1-\mu)\rho^2\sigma^2 N_t$, we get

$$\begin{aligned}\mathbf{P}\left(\lambda_{\max}\left(\sum_{k\in\mathcal{N}_t^{\mathsf{SE}}} W_k\right) \geq (1-\mu)\rho^2\sigma^2 N_t \;\middle|\; N_t = n\right) &\leq d\cdot\exp\left(-\frac{(1-\mu)^2\rho^4\sigma^4 n^2}{2nk_3^2}\right)\\ &\leq d\cdot\exp\left(-k_4(1-\mu)^2 n\right),\end{aligned}$$

where k_4 is defined as

$$k_4 := \frac{\rho^4\sigma^4}{2k_3^2}.$$

APPENDIX C

PROOFS OF RESULTS IN CHAPTER 4

We introduce the following quantities, which will be useful in the sequel. Define $M_0 = g(\mu_\lambda) = \max_{p\in[0,\mu_\lambda]} g(p)$, $m_1 = \min_{p\in[0,\mu_\lambda]} g'(p)$, $M_1 = \max_{p\in[0,\mu_\lambda]} g'(p)$, and $M_2 = \max_{p\in[0,\mu_\lambda]} |g''(p)|$. Note that $m_1 > 0$ as g is assumed to be strictly increasing.

C.1 Proof of Lemma 3

Given a fixed pair (Q, p), we have that

$$\pi_t(Q,p) = \mu_\lambda Q - pg(p) + \mu_\rho^+ \mathbb{E}\left[(g(p) - Q + \varepsilon_t)_+\right] - \mu_\rho^- \mathbb{E}\left[(Q - g(p) - \varepsilon_t)_+\right].$$

We show that π_t is concave in (Q, p). First, we prove that $pg(p)$ is convex. From Assumption 6 it follows that

$$(pg(p))'' = 2g'(p) + pg''(p) \geq 2g'(p) - 2p\frac{g'(p)^2}{g(p)} \geq 0,$$

where the last inequality follows from concavity of g and the fact that $g(0) \geq 0$. The concavity of $\mu_\rho^+ \mathbb{E}\left[(g(p) - Q + \varepsilon_t)_+\right] - \mu_\rho^- \mathbb{E}\left[(Q - g(p) - \varepsilon_t)_+\right]$ follows from Assumption 5 and the fact that g is concave .

Thus, one can characterize the unique maximizer of π_t as the solution to the first-order optimality conditions:

$$\frac{d\pi_t(Q,p)}{dp} = 0 \quad \text{and} \quad \frac{d\pi_t(Q,p)}{dQ}.$$

By Fubini's theorem, we have

$$\begin{aligned}\frac{d}{dp}\mathbb{E}\left[(g(p)-Q+\varepsilon_t)_+\right] &= \mathbb{E}\left[g'(p)\mathbb{1}\left\{g(p)-Q+\varepsilon_t \geq 0\right\}\right] \\ &= g'(p)\mathbf{P}\left(g(p)-Q+\varepsilon_t \geq 0\right) \\ &= g'(p)(1-F(Q-g(p)).\end{aligned}$$

We then get

$$\frac{d\pi_t(Q,p)}{dp} = -g(p) - pg'(p) + \mu_\rho^+ g'(p)(1-F(Q-g(p))) + \mu_\rho^- g'(p)F(Q-g(p)), \tag{C.1}$$

$$\frac{d\pi_t(Q,p)}{dQ} = \mu_\lambda - \mu_\rho^+(1-F(Q-g(p))) - \mu_\rho^- F(Q-g(p)). \tag{C.2}$$

By replacing Equations (C.1) and (C.2) in the first-order optimality conditions, we get Equations (4.4) and (4.5), respectively.

C.2 Proof of Theorem 6

Let $t \geq 1$, and fix (Q_t, p_t). To streamline the proof, we define $Y_t := Q_t - g(p_t)$ for each time period t. It follows that the expected profit of the aggregator can be expressed as

$$\pi_t(Q_t,p_t) = \mu_\lambda Y_t + (\mu_\lambda - p_t)g(p_t) + \mu_\rho^+\mathbb{E}\left[(\varepsilon_t - Y_t)_+\right] - \mu_\rho^-\mathbb{E}\left[(Y_t-\varepsilon_t)_+\right].$$

It will be helpful to decompose the expected profit as $r_t(Q_t,p_t) = r_{1t}(Q_t,p_t) + r_{2t}(Q_t,p_t)$, where

$$\begin{aligned}\pi_{1t}(Q_t,p_t) &:= (\mu_\lambda - p_t)g(p_t) \\ \pi_{2t}(Q_t,p_t) &:= \mu_\lambda Y_t + \mu_\rho^+\mathbb{E}\left[(\varepsilon_t - Y_t)_+\right] - \mu_\rho^-\mathbb{E}\left[(Y_t-\varepsilon_t)_+\right].\end{aligned}$$

We first show that there exists a positive finite C_0 such that for all $t \geq 1$, we have

$$\pi_{1t}(Q^*, p^*) - \pi_{1t}(Q_t, p_t) \leq C_0(p_t - p^*)^2. \tag{C.3}$$

It holds that

$$\begin{aligned}\pi_{1t}(Q^*, p^*) - \pi_{1t}(Q_t, p_t) &= (\mu_\lambda - p^*)g(p^*) - (\mu_\lambda - p_t)g(p_t) \\ &= (\mu_\lambda - p^*)(g(p^*) - g(p_t)) + (p_t - p^*)g(p_t).\end{aligned}$$

By Taylor's theorem, there exists q_{1t} and q_{2t} with $|q_{1t} - p^*| \leq |p_t - p^*|$ and $|q_{2t} - p^*| \leq |p_t - p^*|$ such that

$$\begin{aligned}g(p_t) &= g(p^*) + (p_t - p^*)g'(q_{1t}), \\ g(p_t) &= g(p^*) + (p_t - p^*)g'(p^*) + \frac{1}{2}(p_t - p^*)^2 g''(q_{2t}).\end{aligned}$$

Using Equation (4.4), we get

$$\begin{aligned}\pi_{1t}(Q^*, p^*) - \pi_{1t}(Q_t, p_t) &= -(\mu_\lambda - p^*)(p_t - p^*)\left(g'(p^*) + \frac{1}{2}(p_t - p^*)g''(q_{2t})\right) \\ &\quad + (p_t - p^*)g(p^*) + (p_t - p^*)^2 g'(q_{1t}) \\ &= (p_t - p^*)^2\left(-\frac{1}{2}(\mu_\lambda - p^*)g''(q_{2t}) + g'(q_{1t})\right) \\ &\leq C_0(p_t - p^*)^2,\end{aligned}$$

where C_0 is defined as $C_0 := \frac{1}{2}\mu_\lambda M_2 + M_1$.

We now show that there exists a positive finite C_1 such that for all $t \geq 1$, we have

$$\pi_{2t}(Q^*, p^*) - \pi_{2t}(Q_t, p_t) \leq C_1(Y_t - Y^*)^2, \tag{C.4}$$

where $Y^* := Q^* - g(p^*) = F^{-1}(\zeta)$. First, consider the case in which $Y_t \geq Y^*$. It

follows that

$$
\begin{aligned}
&\pi_{1t}(Q^*, p^*) - \pi_{1t}(Q_t, p_t) \\
&= \mu_\lambda(Y^* - Y_t) + \mu_\rho^+ \int_{Y^*}^{\infty} (\varepsilon_t - Y^*)\, dF - \mu_\rho^+ \int_{Y_t}^{\infty} (\varepsilon_t - Y_t)\, dF \\
&\quad - \mu_\rho^- \int_{-\infty}^{Y^*} (Y^* - \varepsilon_t)\, dF + \mu_\rho^- \int_{-\infty}^{Y_t} (Y_t - \varepsilon_t)\, dF \\
&= \mu_\lambda(Y^* - Y_t) + \mu_\rho^+ \int_{Y^*}^{\infty} (Y_t - Y^*)\, dF + \mu_\rho^- \int_{-\infty}^{Y^*} (Y_t - Y^*) dF \\
&\quad + (\mu_\rho^- - \mu_\rho^+) \int_{Y_t^*}^{Y_t} (Y_t - \varepsilon_t) dF \\
&= (Y^* - Y_t)\big(\mu_\lambda - \mu_\rho^+(1 - F(Y^*)) - \mu_\rho^- F(Y^*)\big) + (\mu_\rho^- - \mu_\rho^+) \int_{Y^*}^{Y_t} (Y_t - \varepsilon_t)\, dF \\
&= (\mu_\rho^- - \mu_\rho^+) \int_{Y^*}^{Y_t} (Y_t - \varepsilon_t)\, dF, \qquad \text{(C.5)}
\end{aligned}
$$

where the last equality follows from Equation (C.2) and the fact that $F(Y^*) = F(F^{-1}(\zeta)) = \zeta$. We have

$$
\begin{aligned}
\int_{Y^*}^{Y_t} (Y_t - \varepsilon_t)\, dF &\leq \int_{Y^*}^{Y_t} (Y_t - Y^*)\, dF \\
&= (Y_t - Y^*) F(Y_t - Y^*) \\
&\leq L(Y_t - Y^*)^2, \qquad \text{(C.6)}
\end{aligned}
$$

where $L := \max_{x,y \neq x \in \mathbb{R}} |F(x) - F(y)|/|x - y|$. Thus, by applying Inequality (C.6) to Equality (C.5), we get the desired inequality with $C_1 := L(\mu_\rho^- - \mu_\rho^+)$. For the case in which $Y_t < Y^*$, one can obtain an identical upper bound using an analogous approach as above.

Finally, combining Inequality (C.4) and Equation (C.3) yields the desired upper bound on regret.

C.3 Proof of Corollary 3

By applying Theorem 6 to the PCE policy (4.12) and (4.13), we get

$$R_T^{\gamma} \leq C_2 \sum_{i=1}^{I_T} L_i \delta_i^2 + 3C_0 \sum_{i=1}^{I_T} L_i \mathbb{E}\left[(\widehat{p}_i - p^*)^2\right] + 3C_1 \sum_{i=1}^{I_T} L_i \mathbb{E}\left[\left(\widehat{Q}_{i-1} - Q^* - g(\widehat{p}_{i-1}) + g(p^*)\right)^2\right], \tag{C.7}$$

where $C_2 := 2C_0 + 2C_1 M_1$. We now upper bound the last term in Inequality (C.7). From the definition of empirical quantile function in Equation (4.9), for all $\eta \in (0,1)$, we have

$$\widehat{F}_i^{-1}(\eta) = g(\widehat{p}_i) - \widehat{\alpha}_i \widehat{p}_i + \widehat{\beta}_i + F_i^{-1}(\eta). \tag{C.8}$$

where $F_i^{-1}(\eta)$ is the quantile function associated with the sequence of demand shocks defined in Equation (4.17). Using Equation (C.8), we get

$$\widehat{Q}_i - Q^* - g(\widehat{p}_i) + g(p^*) = F_{i-1}^{-1}(\eta) - F^{-1}(\eta). \tag{C.9}$$

C.4 Proof of Lemma 4

We first find the CE pricing error in terms of the linearization error. Using the definition of p^* in terms of the demand function linearization in Equation (4.7) and the definition of the CE price in Equation (4.10), we get

$$\begin{aligned} |\widehat{p}_{i+1} - p^*| &= \left| \mathscr{P}\left(\frac{1}{2}\left(\mu_\lambda - \frac{\widehat{\beta}_i}{\widehat{\alpha}_i}\right)\right) - \frac{1}{2}\left(\mu_\lambda - \frac{\beta(p^*)}{\alpha(p^*)}\right)\right| \\ &\leq \frac{1}{2}\left|\frac{\widehat{\beta}_i}{\widehat{\alpha}_i} - \frac{\beta(p^*)}{\alpha(p^*)}\right| \\ &\leq \frac{1}{2}\left|\frac{\beta(\widehat{p}_i)}{\alpha(\widehat{p}_i)} - \frac{\beta(p^*)}{\alpha(p^*)}\right| + \frac{1}{2}\left|\frac{\widehat{\beta}_i}{\widehat{\alpha}_i} - \frac{\beta(\widehat{p}_i)}{\alpha(\widehat{p}_i)}\right|, \end{aligned} \tag{C.10}$$

where the last inequality follows from the triangle inequality. We now bound each term in Inequality (C.10) separately.

First term: Using the definition of α and β, we get

$$\frac{\beta(\widehat{p}_i)}{\alpha(\widehat{p}_i)} - \frac{\beta(p^*)}{\alpha(p^*)} = \frac{g(\widehat{p}_i)}{g'(\widehat{p}_i)} - \widehat{p}_i - \frac{g(p^*)}{g'(p^*)} + p^*.$$

By Taylor's theorem, there exists $q \in \mathbb{R}$ with $|q - p^*| \leq |\widehat{p}_i - p^*|$ such that

$$\frac{g(\widehat{p}_i)}{g'(\widehat{p}_i)} = \frac{g(p^*)}{g'(p^*)} + \left(1 - \frac{g(q)g''(q)}{g'(q)^2}\right)(\widehat{p}_i - p^*).$$

Then,

$$\frac{1}{2}\left|\frac{\beta(\widehat{p}_i)}{\alpha(\widehat{p}_i)} - \frac{\beta(p^*)}{\alpha(p^*)}\right| = \frac{1}{2}\left(\frac{g(q)|g''(q)|}{g'(q)^2}\right)|\widehat{p}_i - p^*| \leq \kappa|\widehat{p}_i - p^*|, \tag{C.11}$$

where the last inequality follows from Assumption 6.

Second term: We first bound the estimation error of the parameters of the demand function linearization. The closed from solution to the LSE (4.8) is given by

$$\widehat{\alpha}_i = \frac{\sum_{t\in\mathcal{E}_i}\left(D_t - \bar{D}_i\right)(p_t - \bar{p}_i)}{\sum_{t\in\mathcal{E}_i}(p_t - \bar{p}_i)^2},$$
$$\widehat{\beta}_i = \bar{D}_i - \widehat{\alpha}_i\bar{p}_i.$$

where $\bar{D}_i := 1/(2L_i)\sum_{t\in\mathcal{E}_i} D_t$ and $\bar{p}_i = 1/(2L_i)\sum_{t\in\mathcal{E}_i} p_t$. After some elementary computations, we get

$$\widehat{\alpha}_i = \frac{g(\widehat{p}_i + \delta_i) - g(\widehat{p}_i)}{\delta_i} + \frac{1}{\delta_i}(W_{i2} - W_{i1}), \tag{C.12}$$

$$\widehat{\beta}_i = \beta(\widehat{p}_i) + W_{i1} - (\widehat{\alpha}_i - \alpha(\widehat{p}_i))\widehat{p}_i, \tag{C.13}$$

where

$$W_{i1} := \frac{1}{L_i} \sum_{t=T_{i-1}+1}^{T_{i-1}+L_i} \varepsilon_t,$$
$$W_{i2} := \frac{1}{L_i} \sum_{t=T_{i-1}+L_i+1}^{T_{i-1}+2L_i} \varepsilon_t.$$

By Taylor's theorem, there exists $q_i \in [\widehat{p}_i, \widehat{p}_i + \delta_i]$ such that

$$\frac{g(\widehat{p}_i + \delta_i) - g(\widehat{p}_i)}{\delta_i} = \alpha(\widehat{p}_i) + \frac{\delta_i}{2} g''(q_i).$$

Then, from Equations (C.12) and (C.13) it follows that

$$\begin{aligned}
\left| \frac{\widehat{\beta}_i}{\widehat{\alpha}_i} - \frac{\beta(\widehat{p}_i)}{\alpha(\widehat{p}_i)} \right| &= \frac{1}{|\widehat{\alpha}_i \alpha(\widehat{p}_i)|} \left| W_{i1} \alpha(\widehat{p}_i) + g(\widehat{p}_i)(\alpha(\widehat{p}_i) - \widehat{\alpha}_i) \right| \\
&\leq \frac{1}{|\widehat{\alpha}_i| \delta_i} \left(\delta_i + \frac{g(\widehat{p}_i)}{|\alpha(\widehat{p}_i)|} \right) (|W_{i1}| + |W_{i2}|) + \delta_i \frac{g(\widehat{p}_i)|g''(q_i)|}{2|\widehat{\alpha}_i \alpha(\widehat{p}_i)|} \\
&\leq Z_i, \qquad \text{(C.14)}
\end{aligned}$$

where

$$Z_i := \frac{1}{|\widehat{\alpha}_i|} \left(\left(\delta_0 + \frac{M_0}{m_1} \right) \frac{|W_{i1}| + |W_{i2}|}{\delta_i} + \kappa M_1 \delta_i \right) \qquad \text{(C.15)}$$

By combining Inequalities (C.10), (C.11), and (C.14), we get

$$|\widehat{p}_{i+1} - p^*| \leq \kappa |\widehat{p}_i - p^*| + \frac{1}{2} Z_i.$$

It immediately follows that

$$(\widehat{p}_{i+1} - p^*)^2 \leq \frac{1 + \kappa^2}{2} (\widehat{p}_i - p^*)^2 + \frac{1 + \kappa^2}{2(1 - \kappa^2)} Z_i^2. \qquad \text{(C.16)}$$

In order to establish an upper bound on the mean squared pricing error using Inequality (C.16), we define a high probability event under which Z_i is appropriately bounded. More precisely, define event $\mathcal{W}$ as

$$\mathcal{W} = \left\{ \max\{|W_{i1}|, |W_{i1}|\} \leq \frac{m_1 \sqrt{\log(L_i)}}{4\sqrt{L_i}} \right\}.$$

By the total expectation theorem, we have

$$\begin{aligned}\mathbb{E}\left[(\widehat{p}_{i+1}-p^*)^2\right] &= \mathbb{E}\left[(\widehat{p}_{i+1}-p^*)^2 \mid \mathcal{W}\right]\mathbf{P}\left(\mathcal{W}\right) \\ &\quad + \mathbb{E}\left[(\widehat{p}_{i+1}-p^*)^2 \mid \mathcal{W}^c\right]\mathbf{P}\left(\mathcal{W}^c\right) \\ &\leq \mathbb{E}\left[(\widehat{p}_{i+1}-p^*)^2 \mid \mathcal{W}\right] + \mu_\lambda^2 \mathbf{P}\left(\mathcal{W}^c\right), \end{aligned} \tag{C.17}$$

where the last inequality follows from the fact that $\widehat{p}_{i+1} \in [0, \mu_\lambda]$ by definition. We now bound the first term in Inequality (C.17). Using Inequality (C.16), we get

$$\mathbb{E}\left[(\widehat{p}_{i+1}-p^*)^2 \mid \mathcal{W}\right] \leq \frac{1+\kappa^2}{2}\mathbb{E}\left[(\widehat{p}_i-p^*)^2 \mid \mathcal{W}\right] + \frac{1+\kappa^2}{2(1-\kappa^2)}\mathbb{E}\left[Z_i^2 \mid \mathcal{W}\right]. \tag{C.18}$$

We now bound $\mathbb{E}\left[Z_i^2 \mid \mathcal{W}\right]$. Conditioned on $\mathcal{W}$, using the triangle inequality, we have

$$\begin{aligned}|\widehat{\alpha}_i| &\geq \frac{g(\widehat{p}_i+\delta_i)-g(\widehat{p}_i)}{\delta_i} - \frac{1}{\delta_i}(|W_{i1}+|W_{i2}|) \\ &\geq m_1 - \frac{1}{\delta_i}\frac{m_1\sqrt{\log(L_i)}}{2\sqrt{L_i}} \\ &\geq \frac{m_1}{2}.\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{E}\left[Z_i^2 \mid \mathcal{W}\right] &\leq \frac{4}{m_1^2}\left(\left(\delta_0 + \frac{M_0}{m_1}\right)\frac{|W_{i1}|+|W_{i2}|}{\delta_i} + \kappa M_1 \delta_i\right)^2 \\ &\leq \frac{4}{m_1^2}\left(\frac{M_0+\delta_0 m_1}{2} + \kappa M_1\right)^2 \frac{\log(L_i)}{\sqrt{L_i}}, \end{aligned} \tag{C.19}$$

where the last inequality follows from the definition of $\delta_i = \delta_0 L_i^{-1/4}$. By applying Inequality (C.19) to (C.18), we get

$$\mathbb{E}\left[(\widehat{p}_{i+1}-p^*)^2 \mid \mathcal{W}\right] \leq \left(\frac{1+\kappa^2}{2}\right)^i (\widehat{p}_1-p^*)^2 + c_0\log(L_i)\sum_{j=1}^{i}\left(\frac{1+\kappa^2}{2}\right)^{i-j}\frac{1}{\sqrt{L_j}},$$

where c_0 is defined as

$$c_0 := \frac{4}{m_1^2}\left(\frac{M_0+\delta_0 m_1}{2} + \kappa M_1\right)^2 \frac{1+\kappa^2}{2(1-\kappa^2)}.$$

Using the definition of $L_i = \lfloor L_0 \nu^i \rfloor$, we get

$$\mathbb{E}\left[(\widehat{p}_{i+1} - p^*)^2 \mid \mathcal{W}\right] \leq \left(\frac{1+\kappa^2}{2}\right)^i \mu_\lambda^2 + c_0 \log(L_0 \nu^i) \frac{\sqrt{\nu}}{\sqrt{L_0 \nu - 1}} \frac{1}{\sqrt{\nu^i}} \sum_{j=1}^{i} \left(\frac{(1+\kappa^2)\sqrt{\nu}}{2}\right)^{i-j}.$$

From the assumption that $\nu \leq 4/(1+\kappa^2)$, it follows that

$$\mathbb{E}\left[(\widehat{p}_{i+1} - p^*)^2 \mid \mathcal{W}\right] \leq c_1 \frac{i}{\sqrt{\nu^i}}, \tag{C.20}$$

where c_1 is defined as

$$c_1 := \mu_\lambda^2 + c_0 \log(L_0 \nu) \frac{2\sqrt{\nu}}{\sqrt{L_0 \nu - 1}(2 - (1+\kappa^2)\sqrt{\nu})}.$$

By applying Inequality (C.20) to Inequality (C.17), we get

$$\mathbb{E}\left[(\widehat{p}_{i+1} - p^*)^2\right] \leq c_1 \frac{i}{\sqrt{\nu^i}} + \mu_\lambda^2 \mathbf{P}\left(\mathcal{W}^c\right).$$

Thus, we are left to bound the probability of event $\mathcal{W}$. Using the fact that $\{\varepsilon_t\}$ is an i.i.d. sequence of random variables, we have

$$\mathbf{P}\left(\mathcal{W}^c\right) \leq 2\mathbf{P}\left(|W_{i1}| \geq \frac{m_1 \sqrt{\log(L_i)}}{4\sqrt{L_i}}\right).$$

As $\varepsilon_t \in [\underline{\varepsilon}, \overline{\varepsilon}]$ almost surely, we utilize Hoeffding's inequality to bound $\mathbf{P}\left(\mathcal{W}^c\right)$.

Lemma 10 (Hoeffding's Inequality)**.** Let $\{X_k\}$ be a sequence of independent random variables such that $X_k \in [a_k, b_k]$ almost surely for all $k \in \mathbb{N}$. Then, for all $s \geq 0$ and $n \in \mathbb{N}$

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}(X_k - \mathbb{E}\left[X_k\right])\right| \geq s\right) \leq 2\exp\left(-\frac{2n^2 s^2}{\sum_{k=1}^{n}(b_i - a_i)^2}\right). \tag{C.21}$$

Using Hoeffding's inequality (C.21), we get

$$\mathbf{P}\left(|W_{i1}| \geq \frac{m_1 \sqrt{\log(L_i)}}{4\sqrt{L_i}}\right) \leq 2\exp\left(-\frac{m_1^2}{8(\overline{\varepsilon} - \underline{\varepsilon})^2}\right)\frac{1}{L_i}.$$

Defining C_2 as follows yields the desired upper bound on the mean squared pricing error (4.19).

$$C_2 = c_1 + 2c_2\mu_\lambda^2 \frac{\sqrt{\nu}}{\sqrt{L_0\nu - 1}}.$$

C.5 Proof of Lemma 5

For all $\eta \in [0,1]$, define $F_i(\eta)$ as

$$F_i(\eta) := \frac{1}{L_i} \sum_{t=T_{i-1}+1}^{T_{i-1}+L_i} \mathbb{1}\{\varepsilon_{it} \leq x\}.$$

Note that $L_i F_i(\eta) \sim \mathrm{Bern}(L_i, F(\eta))$ as $L_i F_i(\eta)$ is the sum of L_i Bernoulli random variables of the form $\mathbb{1}\{\varepsilon_{it} \leq x\}$. Then, for all $s > 0$, we have

$$\begin{aligned}
&\mathbf{P}\left(|F_i^{-1}(\eta) - F^{-1}(\eta)| \geq s\right) \\
&\leq \mathbf{P}\left(F_i^{-1}(\eta) \geq F^{-1}(\eta) + s\right) + \mathbf{P}\left(F_i^{-1}(\eta) \leq F^{-1}(\eta) - s\right) \\
&= \mathbf{P}\left(\eta \geq F_i\left(F^{-1}(\eta) + s\right)\right) + \mathbf{P}\left(\eta \leq F_i\left(F^{-1}(\eta) - s\right)\right) \\
&\leq \exp\left(-\frac{2}{L_i}\left(L_i F_i\left(F^{-1}(\eta) + s\right) - L_i\eta\right)^2\right) \\
&\quad + \exp\left(-\frac{2}{L_i}\left(L_i\eta - L_i F_i\left(F^{-1}(\eta) - s\right)\right)^2\right),
\end{aligned}$$

where the last inequality follows from Hoeffding's inequality (C.21). increasing. Then,

$$\mathbf{P}\left(|F_i^{-1}(\eta) - F^{-1}(\eta)| \geq s\right) \leq 2\exp\left(-\frac{2}{C_3} L_i s^2\right).$$

where C_3 is defined as

$$C_3 := \frac{1}{\min\left\{\eta - F_i\left(F^{-1}(\eta) - s\right), F_i\left(F^{-1}(\eta) + s\right) - \eta\right\}}.$$

From Assumption 7, we have that F is strictly increasing, and thus, $C_3 < \infty$. Then,

$$\begin{aligned}\mathbb{E}\left[(F_i^{-1}(\eta) - F^{-1}(\eta))^2\right] &= \int_0^\infty \mathbf{P}\left(|F_i^{-1}(\eta) - F^{-1}(\eta)| \geq \sqrt{s}\right)\, ds \\ &\leq \int_0^\infty 2\exp\left(-\frac{2}{C_3}L_i s\right)\, ds \\ &= C_3\frac{1}{L_i}.\end{aligned}$$

C.6 Proof of Theorem 7

By applying Inequalities (4.20) and (4.19) to (C.7), we get

$$\begin{aligned}R_T^\gamma &\leq C_2\sum_{i=1}^{I_T} L_i\delta_i^2 + 3C_0C_3\sum_{i=1}^{I_T} L_i\frac{i-1}{\sqrt{\nu^{i-1}}} \\ &\quad + 3C_1C_4\left((\widehat{Q}_0 - Q^* - g(\widehat{p}_0) + p^*)^2 + \sum_{i=2}^{I_T} L_i\frac{1}{L_{i-1}}\right) \\ &\leq C_2\delta_0\sqrt{L_0}\sum_{i=1}^{I_T}\nu^{i/2} + 3C_0C_3L_0\nu\sum_{i=1}^{I_T} i\nu^{i/2} \\ &\quad + 3C_1C_4\left((\widehat{Q}_0 - Q^* - g(\widehat{p}_0) + p^*)^2 + \frac{L_0\nu}{L_0\nu - 1}I_T\right) \\ &\leq C_6\sum_{i=1}^{I_T} i\nu^{i/2},\end{aligned}$$

where C_6 is defined as

$$C_6 := C_2\delta_0\sqrt{L_0} + 3C_0C_3L_0\nu + 3C_1C_4\left((\widehat{Q}_0 - Q^* - g(\widehat{p}_0) + p^*)^2 + \frac{L_0\nu}{L_0\nu - 1}\right).$$

Then,

$$\begin{aligned} R_T^\gamma &\leq C_6 \sum_{i=1}^{I_T} i\nu^{i/2} \\ &\leq C_6 I_T \sum_{i=1}^{I_T} \nu^{i/2} \\ &\leq C_6 \frac{\sqrt{\nu}}{\sqrt{\nu}-1} I_T \sqrt{\nu^{I_T}}. \end{aligned}$$

We not show that $I_T \leq \log(T) + C_7$. From the definition of $I_T = \min\{i \in \mathbb{N} \mid T_i \geq T\}$, it follows that $2\sum_{i=1}^{I_T-1} L_i \leq T$. Using the fact that $L_i = \lfloor L_0 \nu^i \rfloor \geq L_0 \nu^i - 1$, we have $\sum_{i=1}^{I_T-1} \nu^i \leq T/2 + I_T$. Note that $I_T \leq T$ so

$$\frac{\nu^{I_T} - 1}{\nu - 1} \leq \frac{3}{2} T.$$

Then, $I_T \log(\nu) \leq \log(3(\nu-1)T/2) + 1$. Thus,

$$I_T \leq \log(T) + C_7,$$

where $C_7 := (1 + \log(3(\nu-1)/2))/\log(\nu)$.

Finally, defining C_5 as follows yields the desired upper bound on regret.

$$C_6 = C_6 \frac{\sqrt{\nu}}{\sqrt{\nu}-1}(1 + C_7)\sqrt{\nu^{C_7}\log(\nu)}.$$

BIBLIOGRAPHY

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[2] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1, 2012.

[3] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

[4] Theodore W Anderson and John B Taylor. Some experimental results on the statistical properties of least squares estimates in control problems. *Econometrica: Journal of the Econometric Society*, pages 1289–1302, 1976.

[5] Masanao Aoki. *Optimization of stochastic systems: topics in discrete-time systems*. Academic Press, 1967.

[6] Kenneth J Arrow, Theodore Harris, and Jacob Marschak. Optimal inventory policy. *Econometrica: Journal of the Econometric Society*, pages 250–272, 1951.

[7] Peter Auer. Using upper confidence bounds for online learning. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 270–279. IEEE, 2000.

[8] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272, 2017.

[9] Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *arXiv preprint arXiv:1704.09011*, 2017.

[10] Donald A Berry and Larry M Pearson. Optimal designs for clinical trials with dichotomous responses. *Statistics in Medicine*, 4(4):497–508, 1985.

[11] Dimitris Bertsimas and Phebe Vayanos. Data-driven learning in dynamic pricing using adaptive optimization. 2014.

[12] Omar Besbes and Assaf Zeevi. On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science*, 61(4):723–739, 2015.

[13] Eilyan Bitar and Yunjian Xu. On incentive compatibility of deadline differentiated pricing for deferrable demand. In *Decision and control (CDC), 2013 IEEE 52nd annual conference on*, pages 5620–5627. IEEE, 2013.

[14] Eilyan Bitar and Yunjian Xu. Deadline differentiated pricing of deferrable electric loads. *IEEE Transactions on Smart Grid*, 8(1):13–25, 2017.

[15] Severin Borenstein, Michael Jaske, and Arthurenfeld Ros. Dynamic pricing, advanced metering, and demand response in electricity markets. *Journal of the American Chemical Society*, 128(12):4136–45, 2002.

[16] Vivek Borkar and Pravin Varaiya. Identification and adaptive control of markov chains. *SIAM Journal on Control and Optimization*, 20(4):470–489, 1982.

[17] Clay Campaigne and Shmuel S Oren. Firming renewable power with demand response: an end-to-end aggregator business model. *Journal of Regulatory Economics*, pages 1–37, 2015.

[18] Asaf Cassel, Shie Mannor, and Assaf Zeevi. A general approach to multi-armed bandits under risk criteria. In *Conference On Learning Theory*, pages 1295–1306, 2018.

[19] Hung-po Chao. Demand response in wholesale electricity markets: the choice of customer baseline. *Journal of Regulatory Economics*, 39(1):68–88, 2011.

[20] Hung-po Chao. Competitive electricity markets with consumer subscription service in a smart grid. *Journal of Regulatory Economics*, 41(1):155–180, 2012.

[21] Hung-po Chao and Robert Wilson. Priority service: Pricing, investment, and market organization. *The American Economic Review*, pages 899–916, 1987.

[22] Charalampos Chelmis, Muhammad Rizwan Saeed, Marc Frincu, and Viktor Prasanna. Curtailment estimation methods for demand response: Lessons learned by comparing apples to oranges. In *Proceedings of the 2015*

ACM Sixth International Conference on Future Energy Systems, pages 217–218. ACM, 2015.

[23] Chen Chen, Jianhui Wang, and Shalinee Kishore. A distributed direct load control approach for large-scale residential demand response. *IEEE Transactions on Power Systems*, 29(5):2219–2228, 2014.

[24] ConEdison. Energy efficiency and demand management procedure - general calculating customer baseline load. 2013.

[25] Katie Coughlin, Mary Ann Piette, Charles Goldman, and Sila Kiliccote. Statistical analysis of baseline load models for non-residential buildings. *Energy and Buildings*, 41(4):374–381, 2009.

[26] Claude Crampes and Thomas-Olivier Léautier. Demand response in adjustment markets for electricity. *Journal of Regulatory Economics*, 48(2):169–193, 2015.

[27] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, pages 355—-366, 2008.

[28] Yahel David, Balázs Szörényi, Mohammad Ghavamzadeh, Shie Mannor, and Nahum Shimkin. PAC bandits with risk constraints. In *ISAIM*, 2018.

[29] Arnoud V den Boer. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science*, 20(1):1–18, 2015.

[30] Arnoud V den Boer and Bert Zwart. Simultaneously learning and optimizing using controlled variance pricing. *Management science*, 60(3):770–783, 2013.

[31] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

[32] Torgeir Ericson. Direct load control of residential water heaters. *Energy Policy*, 37(9):3502–3512, 2009.

[33] Ahmad Faruqui, Ryan Hledik, and John Tsoukalis. The power of dynamic pricing. *The Electricity Journal*, 22(3):42–56, 2009.

[34] Ahmad Faruqui and Sanem Sergici. Household response to dynamic pricing of electricity: a survey of 15 experiments. *Journal of Regulatory Economics*, 38(2):193–225, 2010.

[35] FERC. Order 719, Wholesale competition in regions with organized electric markets. *Federal Energy Regulatory Commission*, 2008.

[36] Nicolas Galichet, Michele Sebag, and Olivier Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pages 245–260, 2013.

[37] Lingwen Gan, Ufuk Topcu, and Steven H Low. Optimal decentralized protocol for electric vehicle charging. *IEEE Transactions on Power Systems*, 28(2):940–951, 2013.

[38] Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. Improved algorithms for conservative exploration in bandits. *arXiv preprint arXiv:2002.03221*, 2020.

[39] Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[40] Roberto Gomez, Michael Chertkov, Scott Backhaus, and Hilbert J Kappen. Learning price-elasticity of smart consumers in power distribution systems. In *Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference on*, pages 647–652. IEEE, 2012.

[41] Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In *Conference on Learning Theory*, pages 861–898, 2015.

[42] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[43] José Iria, Filipe Soares, and Manuel Matos. Optimal supply and demand bidding strategy for an aggregator of small prosumers. *Applied Energy*, 2017.

[44] Shweta Jain, Balakrishnan Narayanaswamy, and Y Narahari. A multi-armed bandit incentive mechanism for crowdsourcing demand response

in smart grids. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[45] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

[46] Liyan Jia and Lang Tong. Day ahead dynamic pricing for demand response in dynamic environments. In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pages 5608–5613. IEEE, 2013.

[47] Liyan Jia, Lang Tong, and Qing Zhao. An online learning approach to dynamic pricing for demand response. *arXiv preprint arXiv:1404.1325*, 2014.

[48] D. Kalathil and R. Rajagopal. Online learning for demand response. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 218–222, Sept 2015.

[49] Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi, and Benjamin Van Roy. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems*, pages 3910–3919, 2017.

[50] N Bora Keskin and Assaf Zeevi. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research*, 62(5):1142–1167, 2014.

[51] N Bora Keskin and Assaf Zeevi. On incomplete learning and certainty-equivalence control. *Operations Research*, 66(4):1136–1167, 2018.

[52] Kia Khezeli and Eilyan Bitar. Risk-sensitive learning and pricing for demand response. *IEEE Transactions on Smart Grid*, 9(6):6000–6007, Nov 2018.

[53] Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384, 2018.

[54] Panqanamala Ramana Kumar and Pravin Varaiya. *Stochastic systems: Estimation, identification, and adaptive control*. SIAM, 2015.

[55] Soumya Kundu, Nikolai Sinitsyn, Scott Backhaus, and Ian Hiskens. Modeling and control of thermostatically controlled loads. *arXiv preprint arXiv:1101.2157*, 2011.

[56] TL Lai and Herbert Robbins. Iterated least squares in multiperiod control. *Advances in Applied Mathematics*, 3(1):50–73, 1982.

[57] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[58] Tze Leung Lai and Ching Zong Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, pages 154–166, 1982.

[59] Jonathan Levin, Liran Einav, Chiara Farronato, and Neel Sundaresan. Auctions versus posted prices in online markets. *Journal of Political Economy*.

[60] Chang Li, Branislav Kveton, Tor Lattimore, Ilya Markov, Maarten de Rijke, Csaba Szepesvári, and Masrour Zoghi. Bubblerank: Safe online learning to re-rank via implicit click feedback. In *The Conference on Uncertainty in Artificial Intelligence*, 2019.

[61] Na Li, Lijun Chen, and Steven H Low. Optimal demand response based on utility maximization in power networks. In *Power and Energy Society General Meeting, 2011 IEEE*, pages 1–8. IEEE, 2011.

[62] Sen Li, Wei Zhang, Jianming Lian, and Karanjit Kalsi. Market-based coordination of thermostatically controlled loads – part I: A mechanism design formulation. *IEEE Transactions on Power Systems*, 31(2):1170–1178, 2016.

[63] Weixuan Lin and Eilyan Bitar. Forward electricity markets with uncertain supply: Cost sharing and efficiency loss. In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, pages 1707–1713. IEEE, 2014.

[64] Weiwu Ma, Song Fang, Gang Liu, and Ruoyu Zhou. Modeling of district load forecasting for distributed energy system. *Applied Energy*, 204:181–205, 2017.

[65] Zhongjing Ma, Duncan S Callaway, and Ian A Hiskens. Decentralized charging control of large populations of plug-in electric vehicles. *IEEE Transactions on Control Systems Technology*, 21(1):67–78, 2013.

[66] Johanna L Mathieu, Maryam Kamgarpour, John Lygeros, Göran Andersson, and Duncan S Callaway. Arbitraging intraday wholesale energy market prices with aggregations of thermostatic loads. *IEEE Transactions on Power Systems*, 30(2):763–772, 2015.

[67] Amir-Hamed Mohsenian-Rad, Vincent WS Wong, Juri Jatskevich, Robert Schober, and Alberto Leon-Garcia. Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid. *Smart Grid, IEEE Transactions on*, 1(3):320–331, 2010.

[68] Deepan Muthirayan, Dileep Kalathil, Kameshwar Poolla, and Pravin Varaiya. Mechanism design for self-reporting baselines in demand response. In *American Control Conference (ACC), 2016*, pages 1446–1451. American Automatic Control Council (AACC), 2016.

[69] Ashutosh Nayyar, Matias Negrete-Pincetic, Kameshwar Poolla, and Pravin Varaiya. Duration-differentiated energy services with a continuum of loads. *IEEE Transactions on Control of Network Systems*, 3(2):182–191, 2016.

[70] Daniel O Neill, Marco Levorato, Andrea Goldsmith, and Urbashi Mitra. Residential demand response using reinforcement learning. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 409–414. IEEE, 2010.

[71] NYISO. Markets and operational data, 2016.

[72] NYSEG. Smart home rate implementation plan. Technical report, 07 2019. Available on `http://documents.dps.ny.gov/public/Common/ViewDoc.aspx?DocRefId=%7B5D79D7C5-8002-4B00-8C32-D9CDE1374F8B%7D`.

[73] Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

[74] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

[75] Walid Saad, Zhu Han, H Vincent Poor, and Tamer Basar. Game-theoretic methods for the smart grid: An overview of microgrid systems, demand-side management, and smart grid communications. *IEEE Signal Processing Magazine*, 29(5):86–105, 2012.

[76] Pedram Samadi, Amir-Hamed Mohsenian-Rad, Robert Schober, Vincent WS Wong, and Juri Jatskevich. Optimal real-time pricing algorithm based on utility maximization for smart grid. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 415–420. IEEE, 2010.

[77] Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.

[78] Gaurav Sharma, Le Xie, and PR Kumar. Large population optimal demand response for thermostatically controlled inertial loads. In *Smart Grid Communications (SmartGridComm), 2013 IEEE International Conference on*, pages 259–264. IEEE, 2013.

[79] Nasim Yahya Soltani, Seung-Jun Kim, and Georgios B Giannakis. Real-time load elasticity tracking and pricing for electric vehicle charging. *Smart Grid, IEEE Transactions on*, 6(3):1303–1313, 2015.

[80] Yanan Sui, Joel Burdick, Yisong Yue, et al. Stagewise safe bayesian optimization with gaussian processes. In *International Conference on Machine Learning*, pages 4788–4796, 2018.

[81] Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *International Conference on Machine Learning*, pages 997–1005, 2015.

[82] Wen Sun, Debadeepta Dey, and Ashish Kapoor. Safety-aware algorithms for adversarial contextual bandit. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3280–3288. JMLR. org, 2017.

[83] Kalyan T Talluri and Garrett J Van Ryzin. *The theory and practice of revenue management*, volume 68. Springer Science & Business Media, 2006.

[84] Chin-Woo Tan and Pravin Varaiya. Interruptible electric power service contracts. *Journal of Economic Dynamics and Control*, 17(3):495–517, 1993.

[85] Hamidreza Tavafoghi and Demosthenis Teneketzis. Optimal contract design for energy procurement. In *Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on*, pages 62–69. IEEE, 2014.

[86] John B Taylor. Asymptotic properties of multiperiod control rules in the linear regression model. *International Economic Review*, pages 472–484, 1974.

[87] Joshua A Taylor and Johanna L Mathieu. Index policies for demand response. *Power Systems, IEEE Transactions on*, 29(3):1287–1295, 2014.

[88] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[89] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[90] U.S. Department of Energy. Benefits of demand response in electricity markets and recommendations for achieving them. *A Report to the United State Congress Pursuant of Section 1252 of the Energy Policy Act of 2005*, February 2006.

[91] U.S. Energy Information Administration. Electric power monthly. *U.S. Department of Energy*, February 2017.

[92] Ilnura Usmanova, Andreas Krause, and Maryam Kamgarpour. Safe convex learning under uncertain constraints. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2106–2114, 2019.

[93] Sattar Vakili and Qing Zhao. Mean-variance and value at risk in multi-armed bandit problems. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1330–1335. IEEE, 2015.

[94] Sattar Vakili and Qing Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1093–1111, 2016.

[95] Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

[96] Qingsi Wang, Mingyan Liu, and Johanna L Mathieu. Adaptive demand response: Online learning of restless and controlled bandits. In *Smart Grid Communications (SmartGridComm), 2014 IEEE International Conference on*, pages 752–757. IEEE, 2014.

[97] Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *International Conference on Machine Learning*, pages 1254–1262, 2016.

[98] Yunjian Xu, Na Li, and Steven H Low. Demand response with capacity

constrained supply function bidding. *IEEE Transactions on Power Systems*, 31(2):1377–1394, 2016.

[99] Peng Yang, Gongguo Tang, and Arye Nehorai. A game-theoretic approach for optimal time-of-use electricity pricing. *IEEE Transactions on Power Systems*, 28(2):884–892, 2013.