Data-driven Clustering and Feature-based Retail Electricity Pricing with Smart Meters

N. Bora Keskin[1] Yuexing Li[1] Nur Sunar[2]

[1]Duke University, Fuqua School of Business, email: {bora.keskin, yuexing.li}@duke.edu

[2]University of North Carolina at Chapel Hill, Kenan-Flagler Business School, email: nur_sunar@kenan-flagler.unc.edu

We consider an electric utility company that serves retail electricity customers over a discrete-time horizon. In each period, the company observes the customers' consumption as well as high-dimensional features on customer characteristics and exogenous factors. A distinctive element of our work is that these features exhibit three types of heterogeneity—over time, customers, or both. Based on the consumption and feature observations, the company can dynamically adjust the retail electricity price at the customer level. The consumption depends on the features: there is an underlying structure of clusters in the feature space, and the relationship between consumption and features is different in each cluster. Initially, the company knows neither the underlying cluster structure nor the corresponding consumption models. We design a data-driven policy of joint spectral clustering and feature-based pricing and show that our policy achieves near-optimal performance, i.e., its average regret converges to zero at the fastest achievable rate. This work is the first to theoretically analyze joint clustering and feature-based pricing with different types of feature heterogeneity. Our case study based on real-life smart meter data from Texas illustrates that our policy increases company profits by 146% over a three-month period relative to the company policy, and is robust to various forms of model misspecification.

Key words: spectral clustering, feature-based dynamic pricing, data-driven analysis, retail electricity, smart meter, lasso regularization, exploration-exploitation.

History: This version: December 11, 2021.

1. Introduction

1.1. Overview and Practical Motivations

Smart meters—devices that enable two-way real-time communication between electric utility companies (*utilities*) and customers—have been widely deployed across the world. The Edison Foundation (2021) estimates that 75% of U.S. households had smart meters by the end of 2020. Similarly, in the E.U., 77% of consumers are expected to have smart meters for electricity by the end of 2024 (European Commission 2021). The broad installation of smart meters enables dynamic pricing of retail electricity, under which utilities sell electricity to their customers—end-users—at different rates depending on the time of the day and the season. With smart meters, utilities can rapidly communicate price signals with their customers so that customers can adjust their energy consumption accordingly.

There is a growing interest in dynamic pricing of retail electricity across the globe. U.S. utilities have launched several pilot programs; some implemented dynamic pricing programs (Faruqui and Bourbonnais 2019). For example, Commonwealth Edison, the largest utility in Illinois, implemented *real-time pricing* (*hourly pricing*) under which residential electricity rates change every hour (ComEd 2020). Overall, the number of U.S. customers on time-varying rates increased by more than 90% from 2013 to 2019 (EIA 2020). U.S. regulators are increasingly supportive of advanced meter investments, demanding detailed electricity rate design and implementation (Federal Energy Regulatory Commission 2019). Thus, dynamic pricing is an important emerging business model in the U.S. and likely to gain more popularity in the near future. In

Europe, there has been a broad roll-out of dynamic pricing programs (e.g., 71% of the Norwegian households are subject to hourly-changing electricity rates, ECOFYS 2018). Policy makers in the E.U. strongly support dynamic pricing of retail electricity, and define it as hourly pricing (The E.U. 2019). In fact, the E.U.'s Directive 2019/944 requires that every member state have at least one energy supplier offering real-time pricing to customers (The E.U. 2019, ECOFYS 2018). As a result of all these developments, the number of customers on dynamic pricing rates are forecasted to increase by more than 15-fold from 2018 to 2025 across the globe (De Clercq 2018).

Successful dynamic pricing of retail electricity involves designing well-performing customer segmentation and pricing strategies (Dutta and Mitra 2017). However, this design is challenging in practice as it requires the knowledge of the customers' price-sensitivity and willingness to pay, which are typically unknown to utilities (Dutta and Mitra 2017). Utilities might try to learn these unknown factors from historical consumption data. The caveat is, because dynamic pricing of retail electricity is a new business model for utilities in various parts of the world, many utilities only have access to historical data that display very little price variation.[1] Thus, these utilities, including the majority of U.S. utilities, cannot accurately learn the aforementioned unknown factors from historical data; rather, they must learn them in an *online* fashion. Smart meters and other innovative technologies, such as the Internet of Things and smart home appliances, provide utilities with highly granular data on customer consumption patterns. These technologies together with time-varying rates enable utilities' online learning of the unknown factors, facilitating the design and implementation of successful data-driven customer segmentation and dynamic pricing strategies.

According to the U.S. government, "personalized pricing involves the use of data analytics to provide distinct prices to consumers based on personal characteristics and behaviors" (OECD 2018). Indeed, smart meters and other innovative technologies exactly enable this by providing very granular customer-level consumption data. With such data, utilities can study the individual consumption behavior in detail, and design personalized prices (equivalently discounts or rewards) to induce a certain consumption behavior for each customer. In the U.S., "the mere fact that different customers pay different prices [for the same service] does not raise consumer protection concerns" (OECD 2018). Thus, like the implementations in other sectors, personalized pricing is feasible in the U.S. energy sector. According to McKinsey & Company (2019), personalization can be appealing to utilities because of potential profit improvements and boost in customer loyalty, compared to the status quo. Besides, customers might also favor them because of personalized price signals about the electricity consumption and potential monetary savings (Allcott 2009, Navigant Research 2017). There are many utilities that offer customized pricing plans to their customers.

[1] Consider U.S. utilities as an example. Historically, the vast majority of U.S. electricity end-users were subject to a flat retail electricity price, which does not vary with the time-of-use or season. For instance, in 2017, around 95% of U.S. electricity customers were subject to a flat rate (U.S. EIA 2021a,b). Furthermore, most U.S. utilities do not request a rate adjustment for two to five years under the traditional (flat rate) strategy (Lazaar 2016). Hence, for the majority of U.S. utilities, historical data have very little price variation.

Examples include Eligo Energy (in Illinois), Reliant Energy (in Texas), Xoom Energy LLC (one of the largest energy retailers in North America), Green Mountain Energy (in Texas), and all utilities on Energy Ogre (in Texas) (Eligo Energy 2021, Reliant Energy 2021, Xoom Energy 2021, Green Mountain Energy 2021, Energy Ogre 2021).

Motivated by these practical facts, we consider a utility, called the *seller*, that serves N customers over T periods. The seller has access to detailed information on customer characteristics as well as exogenous factors such as weather conditions. For example, Austin Energy, a prominent U.S. utility, has access to rich and highly granular smart meter data that include numerous types of information about households, such as demographics and living habits, as well as the weather information for Austin, Texas (see §6.3.1). We encode this detailed information as high-dimensional features. Based on our observations on a real-life data set from Austin Energy, the seller observes three types of multidimensional features in each period: (1) *time-heterogeneous features* that vary over time but are invariant over customers at any given time, (2) *user-heterogeneous features* that vary over customers but are invariant over time, and (3) *fully heterogeneous features* that vary over both time and customers. The consumption behavior of the customers depends on the features. The space of features is partitioned into clusters based on similarity, and in each cluster, there is a distinct relationship between consumption and features. In this regard, each cluster can be interpreted as a different customer segment in the market. The seller knows neither the underlying cluster structure nor the consumption models in distinct clusters, and aims to learn these from accumulating observations on features and consumption. To tease out the effect of different feature types, we consider two settings: (i) all features exhibit variation over both customers and time, and (ii) some features may be invariant over customers or time. In these settings, we measure the seller's performance by *average regret*, i.e., the expected profit loss per period per customer, relative to a clairvoyant who knows the underlying cluster structure as well as the underlying consumption models. Using this measure, we design a well-performing joint clustering and feature-based dynamic pricing policy. To establish the robustness of our policy's good performance, we extend our analysis in various directions, such as temporal shifts in feature means.

1.2. Motivation for Customer Segmentation via Spectral Clustering

We use spectral clustering to uncover underlying clusters (customer segments). At a high level, spectral clustering extracts the most essential information in the features and projects it onto a low-dimensional space where it is easier to identify the underlying cluster structure. An important advantage of this method over centroid- and neighborhood-based clustering methods is that it effectively identifies non-convex cluster structures. Thus, it typically outperforms traditional methods such as k-means clustering (Ng et al. 2002). We first demonstrate the effectiveness of spectral clustering with a simple numerical example in Figure 1. Consider a 2-dimensional feature space, with underlying clusters in the form of 3 concentric circles. Figure 1a displays a sample data set from the underlying clusters. Figure 1b illustrates the result of spectral clustering, which identifies the underlying cluster structure (up to a permutation of cluster labels). In contrast, as shown in Figure 1c, k-means clustering fails to identify the clusters due to the non-convex cluster structure.

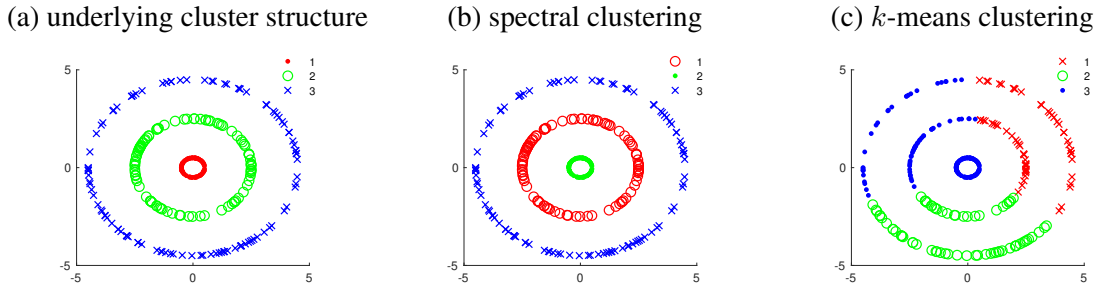


Figure 1 **Spectral clustering vs. k-means clustering.** Panel (a) displays 300 i.i.d. feature samples from a feature space $\mathcal{Z}$, where underlying clusters are in the form of 3 concentric circles with radii 0.5, 2.5, and 4.5, and the probability of observing a feature from each cluster is $1/3$. Panels (b) and (c) show the result of spectral and k-means clustering, respectively, based on the 300 feature samples. Similarity is measured via the Gaussian kernel, $\psi(\mathbf{z},\mathbf{z}') = \exp(-\|\mathbf{z}-\mathbf{z}'\|_2^2/\sigma^2)$ for $\mathbf{z},\mathbf{z}' \in \mathcal{Z}$ with $\sigma = 0.65$.

To demonstrate the value of spectral clustering in practice, we perform spectral clustering on our rich real-life smart meter data, which include 157 features; see §6.3.1 for a detailed description. Figure 2 illustrates 2-dimensional cross sections of weather-related features in the data set and the corresponding cluster structures. Note that cluster 1 in Figure 2a and cluster 2 in Figure 2b exhibit non-convexity, emphasizing the relevance of non-convex cluster structures in practice.

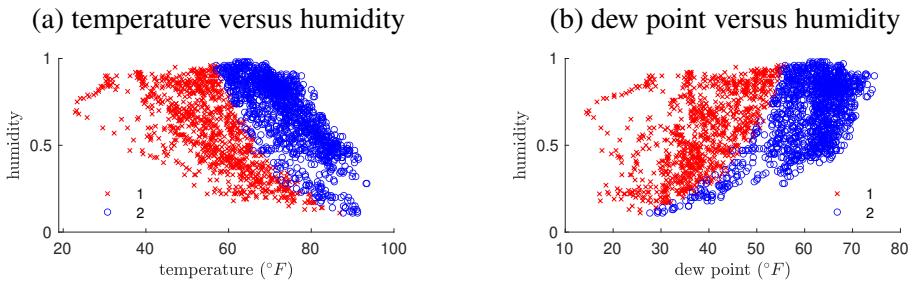


Figure 2 **Non-convex cluster structure from the real-life data set explained in §6.3.1.**

Using the same data set, we also compare spectral clustering with k-means clustering and hierarchical clustering based on Ward's linkage (Ward Jr 1963) in terms of the clustering quality. Each method is evaluated by three commonly used metrics: Davies-Bouldin index (Davies and Bouldin 1979), silhouette value (Rousseeuw 1987), and Dunn index (Dunn 1974). Table 1 summarizes the quality of clustering under each of the three methods.[2] Clustering with smaller Davies-Bouldin index, larger silhouette value, or larger Dunn index is more preferable. For all three metrics, spectral clustering performs the best. This further highlights the advantage of using spectral clustering to analyze real-life data sets.

Table 1 **Comparison of Clustering Methods**

	Davies-Bouldin index	Silhouette value	Dunn index
spectral	0.8612	0.6471	0.0202
k-means	0.8620	0.6460	0.0165
hierarchical	0.8959	0.5275	0.0189

1.3. Main Contributions

Our paper makes the following technical and practical contributions to the literature.

[2] The real-life data set is clustered into two clusters under each method as 2 turns out to be the ideal number of clusters for each method based on multiple standard metrics, including the silhouette value.

1.3.1. Dealing with different forms of feature heterogeneity. A distinctive element of our work is the consideration of multiple types of features: (1) time-heterogeneous, (2) user-heterogeneous and (3) fully heterogeneous features. In contrast to the vast majority of the feature-based pricing literature that only considers (3) without clustering, the presence of (1) and (2) poses a unique challenge in the analysis: the feature vectors containing (1)-(3) no longer form an i.i.d. sequence over time or over customers. Because of this, the methods and performance results developed in the case of fully heterogeneous features are not applicable in our general setting. To overcome this challenge, we design and analyze a new method called *parallel partial regression* (see §4). We show that our method efficiently learns the underlying consumption model in each cluster (see Propositions 2 and 3). We also show that our method and analysis can be extended when a particular feature type is non-i.i.d. (see §7.1). Overall, our approach can facilitate research in other contexts where multiple types of high-dimensional features are observed.

1.3.2. Design and analysis of joint spectral clustering and feature-based pricing. To our knowledge, our paper is the first that analyzes data-driven joint spectral clustering and contextual dynamic pricing. In fact, we are not aware of any prior work that theoretically analyzes spectral clustering in a dynamic pricing context. We design a well-performing joint spectral clustering and feature-based dynamic pricing policy. The policy learns the underlying cluster structure and the consumption behavior in each cluster in an efficient manner, and dynamically updates the estimates of cluster structure and maximizes profits on the fly as new feature information becomes available. To establish the good performance of our policy, we characterize (i) the misclassification errors of spectral clustering, and (ii) how these misclassification errors affect parameter estimation errors, and eventually the regret performance. None of these characterizations are available in the literature, even for the case of only fully heterogeneous features. We prove that the misclassification error is a key determinant of regret performance (see the discussions following Theorems 1 and 2). This emphasizes the importance of judicious clustering decisions in feature-based pricing.

1.3.3. Developing new theoretical performance guarantees and identifying the impact of different forms of feature heterogeneity on performance. We derive distinct performance results in the case where all features are fully heterogeneous and in the case where all three types of features are present. We show that in the former case, the average regret of our policy is of order $1/\sqrt{NT}$, while in the latter case, it is of order $1/\sqrt{N \wedge T}$, where $\wedge$ denotes the minimum of two numbers (see Theorems 1 and 2). These results establish that the different forms of feature heterogeneity observed in our real-life data set have fundamentally different impacts on the performance of a learning policy.

To our knowledge, these theoretical results are the first to identify how the different forms of feature heterogeneity influence the regret performance. To obtain these results, we first employ functional analysis and martingale theory to explicitly characterize the misclassification errors of spectral clustering and the model estimation errors within each cluster. Unlike standard M-estimation settings, spectral clustering introduces randomness in the number of observations in each cluster and could contaminate the observations

used in parameter estimation, creating additional challenges in characterizing estimation errors. We derive concentration inequalities to tackle these challenges (see the discussions after Proposition 2). Thus, our regret analysis in both of the aforementioned cases are novel.

1.3.4. Managerial insights for retail electricity pricing with smart meters. Using real-life smart meter data, we conduct case studies to generate valuable insights for practitioners. Relative to the actual pricing decisions of the U.S. electric utility company in the real-life data set, our approach yields more than a 146% increase in profits. Moreover, we demonstrate how different components of our approach offer substantial value in practice. First, we find that using clustering and feature-based pricing jointly is key in improving performance. Ignoring either modeling element while they are present results in significant losses (see Table 2 in §6.3.4), whereas taking both into account while they are absent still yields good performance (see Figures 9a and 9b in §6.3.5). Second, we show that underestimating the number of clusters leads to significant profit losses while overestimating the number of clusters results in slight losses. Third, we demonstrate that dedicated price experimentation is necessary in this context. Lastly, we illustrate that lasso regularization is effective in estimating sparse consumption models with different forms of high-dimensional feature heterogeneity.

1.4. Related Literature

Technological advances and new business models are changing the landscape of the energy sector (see, e.g., He et al. 2019, Sunar and Birge 2019, Sunar and Swaminathan 2018, Huang et al. 2020, for some of these developments). Our paper contributes to the smart energy operations literature by designing a data-driven segmentation and feature-based dynamic pricing strategy for utilities; interested readers can find perspectives on the broader context of smart city operations in Qi and Shen (2019). Within this literature, there is a stream that studies clustering without considering retail electricity pricing. This stream applies various clustering techniques (e.g., k-means) to load-profile retail electricity customers. We refer readers to Wang et al. (2015) and Rajabi et al. (2020) for comprehensive reviews. Our paper differs from this stream in multiple ways. First, this literature mainly performs data analysis or numerical studies while we also provide a theoretical analysis. Second, we study feature-based dynamic pricing, which is not considered in this stream. Third, these studies typically use load data to cluster customers into different segments, while our work forms clusters based on multi-dimensional feature observations of different types. Apart from this stream, our paper is also related to the literature on dynamic pricing of retail electricity. However, to our knowledge, in this literature, there is no prior work that theoretically analyzes feature-based dynamic pricing or the combination of feature-based dynamic pricing with data-driven clustering, which are the subjects of our study. A recent comprehensive review on dynamic pricing of electricity is by Dutta and Mitra (2017). We also note that in this literature, several empirical papers provide evidence for the superiority of dynamic pricing over flat rates, providing further motivation for our paper. Considering a simulation study with realistic parameters, Borenstein (2005) shows that real-time pricing can provide significant benefit

to utilities and customers, compared to time-invariant pricing. Faruqui and Sergici (2010) survey 15 pilot programs that test various dynamic pricing schemes in the retail electricity market and demonstrate that price signals change retail energy usage, supporting the effectiveness of the dynamic pricing.

Our paper contributes to the nascent literature on personalized dynamic pricing; see, e.g., Nambiar et al. (2019), Javanmard and Nazerzadeh (2019), and **?**. In this literature, it is typical to assume a single type of feature, which exhibits some variability in a given sample. In contrast, as mentioned earlier, we consider three different types of feature heterogeneity, and design a policy that exhibits the best achievable regret performance. Moreover, unlike these earlier studies, our paper analyzes spectral clustering in conjunction with personalized dynamic pricing. Overall, our analysis extends the applicability of personalized dynamic pricing to a much more general framework.

There is a literature that studies various clustering methods, such as k-means (MacQueen et al. 1967), k-means++ (Arthur and Vassilvitskii 2006), mixture model-based approaches (Bertsimas et al. 2003) and spectral clustering (Ng et al. 2002, von Luxburg et al. 2008). We refer readers to Saxena et al. (2017) for a comprehensive review. Among the methods studied in this literature, spectral clustering stands out with its effectiveness in identifying possibly non-convex cluster structures, and typically outperforms traditional methods such as k-means clustering (Ng et al. 2002, von Luxburg et al. 2008). Motivated by these findings, we use spectral clustering in our paper. However, unlike the aforementioned studies that focus on clustering in isolation, we consider how a profit-maximizing seller should implement data-driven personalized dynamic pricing together with spectral clustering.

Our paper is also related to the literature on bandit clustering methods; see, e.g., Cesa-Bianchi et al. (2013), Gentile et al. (2014), Nguyen and Lauw (2014), Gentile et al. (2017), and Miao et al. (2019). Nevertheless, our paper is differentiated from this research stream in several ways. First, unlike these studies, we use spectral clustering. Second, our method accounts for different types of feature heterogeneity, which are not considered in the aforementioned studies. Third, the methods in the above studies form clusters in the space of users or products. In contrast, we construct clusters directly in the space of features. This difference makes our method more scalable. To find the underlying cluster of a user, the aforementioned bandit clustering methods need to make a large number of observations on that user. For a new user whose prior consumption is not observed, those methods cannot infer the underlying cluster, whereas our approach can employ this user's feature information to find the relevant underlying cluster. Fourth, we provide practical guidelines for smart meter pricing in the retail electricity sector.

There are several recent developments in the application of clustering to personalized decision making—e.g., Bernstein et al. (2019) study a Bayesian hierarchical model with a Dirichlet process prior; Aouad et al. (2019) and Kallus (2017) analyze tree-based approaches for customer segmentation. Unlike these works, we conduct regret analysis to establish theoretical performance guarantees for our policy and demonstrate its asymptotic optimality.

In a broader context, our paper also contributes to the operations literature that studies learning problems. Some of these papers (e.g., the ones explained in the personalized pricing literature above) employ frequentist formulations while others consider Bayesian ones (e.g., Bertsimas and Mersereau 2007, Ding et al. 2014, Harrison and Sunar 2015, Zhang 2018, Keskin and Birge 2019, Kim 2020, Sunar et al. 2020).

The remainder of the paper is organized as follows. Section 2 describes the model, and §3 presents preliminaries on spectral graph theory. Section 4 describes our data-driven policy, and §5 provides our theoretical analysis. Section 6 presents case studies based on simulation experiments and a real-life data set. Section 7 considers several extensions, and §8 concludes the paper.

2. Problem Formulation

2.1. Basic Model Elements

Consider an electric utility company, hereafter called the *seller*, that provides retail electricity to its residential customers, namely households. The seller observes various types of features characterizing the weather (e.g., temperature, humidity, and pressure), household demographics (e.g., education level, annual income, and time spent at home), and house constructions (e.g., heating, ventilation, and air conditioning (HVAC) and foundation). For example, see Table 3 for a partial list of features observed by Austin Energy. Based on the feature observations, the seller charges a price for energy usage to each customer, and can adjust prices on an hourly basis. Through smart meter readings, the seller also observes the energy consumption of each customer over time.

The seller serves N customers, indexed by $i \in \mathcal{N} = \{1, \ldots, N\}$, over a discrete time horizon of T periods, indexed by $t \in \mathcal{T} = \{1, \ldots, T\}$. At the beginning of each period t, the seller observes a d-dimensional feature vector $\mathbf{X}_{i,t} = [X_{i,t,1} \ \cdots \ X_{i,t,d}]^\intercal$ for each customer i.[3] In the context of retail electricity, there are three types of features: (1) *time-heterogeneous features* that vary over time but are the same for all customers at any given time, e.g., weather-related information, such as temperature, humidity, pressure, etc., (2) *user-heterogeneous features* that vary over customers but are invariant over time, e.g., square footage of residence, foundation type, etc., and (3) *fully heterogeneous features* that vary over both time and customers, e.g, occurrence of maintenance issues, such as HVAC failure and insulation problems. We denote these three types of features by $\mathbf{X}_{i,t,1} \in \mathbb{R}^{d_1}$, $\mathbf{X}_{i,t,2} \in \mathbb{R}^{d_2}$, and $\mathbf{X}_{i,t,3} \in \mathbb{R}^{d_3}$, respectively, for all i and t, where d_1, d_2, and d_3 are positive integers satisfying $d = d_1 + d_2 + d_3$. Thus, the feature vector takes the following form: $\mathbf{X}_{i,t}^\intercal = [\mathbf{X}_{i,t,1}^\intercal \ \mathbf{X}_{i,t,2}^\intercal \ \mathbf{X}_{i,t,3}^\intercal]$. Note that $\mathbf{X}_{i,t,1} = \mathbf{X}_{j,t,1}$ for all i, j, and t, and that $\mathbf{X}_{i,s,2} = \mathbf{X}_{i,t,2}$ for all i, s, and t. With slight abuse of notation, we suppress the dependency on redundant indices by writing $\mathbf{X}_{i,t,1} = \mathbf{X}_{t,1}$ and $\mathbf{X}_{i,t,2} = \mathbf{X}_{i,2}$ for all i and t. For ease of exposition, we assume that each feature type is i.i.d. over the varying dimension in our base model, and extend our analysis to consider non-i.i.d. features (e.g., any seasonal patterns) in §7.1 for full generality.

[3] Throughout the paper, we denote scalars by regular italic letters, vectors by bold upright letters, and matrices by bold italic letters. All vectors are assumed to be column vectors unless otherwise stated.

The features $\{\mathbf{X}_{t,1} : t \in \mathcal{T}\}$, $\{\mathbf{X}_{i,2} : i \in \mathcal{N}\}$, and $\{\mathbf{X}_{i,t,3} : i \in \mathcal{N}, t \in \mathcal{T}\}$ have compact supports $\mathcal{X}_1 \subset \mathbb{R}^{d_1}$, $\mathcal{X}_2 \subset \mathbb{R}^{d_2}$ and $\mathcal{X}_3 \subset \mathbb{R}^{d_3}$, respectively. The feature space is then defined by the products of these supports as $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$. Let $\mathbb{P}_1(\cdot)$, $\mathbb{P}_2(\cdot)$ and $\mathbb{P}_3(\cdot)$ be the probability measures governing the sequences $\{\mathbf{X}_{t,1} : t \in \mathcal{T}\}$, $\{\mathbf{X}_{i,2} : i \in \mathcal{N}\}$, and $\{\mathbf{X}_{i,t,3} : i \in \mathcal{N}, t \in \mathcal{T}\}$, respectively. We define the probability measure $\mathbb{P}_{\mathcal{X}}(\cdot)$ on the feature space $\mathcal{X}$ such that

$$\mathbb{P}_{\mathcal{X}}(\mathbf{X}_{i,t} \in \mathrm{d}\tilde{\mathbf{X}}_{i,t}) = \mathbb{P}_1(\mathbf{X}_{t,1} \in \mathrm{d}\tilde{\mathbf{X}}_{t,1})\, \mathbb{P}_2(\mathbf{X}_{i,2} \in \mathrm{d}\tilde{\mathbf{X}}_{i,2})\, \mathbb{P}_3(\mathbf{X}_{i,t,3} \in \mathrm{d}\tilde{\mathbf{X}}_{i,t,3})$$

for all $\tilde{\mathbf{X}}_{i,t}$ satisfying $\tilde{\mathbf{X}}_{i,t}^{\mathsf{T}} = [\tilde{\mathbf{X}}_{t,1}^{\mathsf{T}}\ \tilde{\mathbf{X}}_{i,2}^{\mathsf{T}}\ \tilde{\mathbf{X}}_{i,t,3}^{\mathsf{T}}]$. The expectation operator associated with $\mathbb{P}_{\mathcal{X}}(\cdot)$ is $\mathbb{E}_{\mathcal{X}}(\cdot)$.

The feature space $\mathcal{X}$ is segmented into clusters based on similarity among features. For instance, clusters may correspond to whether temperature is above or below a threshold for time-heterogenous features, whether square footage of residence falls into a certain category for user-heterogeneous features, or whether an insulation problem occurs and affects certain customers for fully heterogeneous features. The three types of features induce $\mathcal{K}_1, \mathcal{K}_2$, and $\mathcal{K}_3$ clusters, respectively, and the underlying total number of clusters is $\mathcal{K} = \mathcal{K}_1\mathcal{K}_2\mathcal{K}_3$. More precisely, for each $b \in \{1,2,3\}$, the support $\mathcal{X}_b$ is partitioned into $\mathcal{K}_b$ disjoint sets, $R_1, \ldots, R_{\mathcal{K}_b}$, such that (i) $\mathbb{P}_b(R_k) \geq c_{\min} > 0$ for all $k \in \{1, \ldots, \mathcal{K}_b\}$, and (ii) $\mathbb{P}_b(\overline{R}_j \cap \overline{R}_k) = 0$, where $\overline{R}_k$ denotes the closure of R_k, for all $j, k \in \{1, \ldots, \mathcal{K}_b\}$ with $j \neq k$. Condition (i) is for avoiding redundant clusters from which there is small probability of observing a feature. Condition (ii) ensures that there is zero probability observing a feature on the boundaries of clusters. The seller does not know $\mathcal{K}_1, \mathcal{K}_2$, or $\mathcal{K}_3$, but knows that they are bounded above by $\bar{\mathcal{K}}_1, \bar{\mathcal{K}}_2$, and $\bar{\mathcal{K}}_3$, respectively. Thus, the total number of clusters is bounded above by $\bar{\mathcal{K}} = \bar{\mathcal{K}}_1\bar{\mathcal{K}}_2\bar{\mathcal{K}}_3$. (For a guideline on how to choose $\bar{\mathcal{K}}$ by analyzing data, see the discussion following Theorem 1.) The underlying cluster structure is represented by a function $\mathcal{C} : \mathcal{X} \mapsto \{1, \ldots, \mathcal{K}\}$ that maps any feature vector to its cluster label such that, for any $\mathbf{x} \in \mathcal{X}$, $\mathcal{C}(\mathbf{x}) = k$ for some $k \in \{1, \ldots, \mathcal{K}\}$. The seller does not know the underlying cluster structure and needs to learn it from accumulating data. We denote the means of three types of features in all clusters as $\{\boldsymbol{\mu}_{k_1,1} : k_1 = 1, \ldots, \mathcal{K}_1\}$, $\{\boldsymbol{\mu}_{k_2,2} : k_2 = 1, \ldots, \mathcal{K}_2\}$, and $\{\boldsymbol{\mu}_{k_3,3} : k_3 = 1, \ldots, \mathcal{K}_3\}$, respectively. The corresponding covariance matrices, all of which are symmetric and positive definite, are denoted by $\{\boldsymbol{\Sigma}_{k_1,1} : k_1 = 1, \ldots, \mathcal{K}_1\}$, $\{\boldsymbol{\Sigma}_{k_2,2} : k_2 = 1, \ldots, \mathcal{K}_2\}$, and $\{\boldsymbol{\Sigma}_{k_3,3} : k_3 = 1, \ldots, \mathcal{K}_3\}$, respectively. They are all unknown to the seller.

Upon observing a feature vector $\mathbf{X}_{i,t} = \mathbf{x}_{i,t}$ for every customer i at the beginning of period t, the seller sets real-time prices $\{p_{i,t} \in \mathcal{P} = [p_{\min}, p_{\max}] : i \in \mathcal{N}\}$ for all customers, where $0 < p_{\min} < p_{\max} < \infty$. In response, the seller observes the energy consumption $D_{i,t}$ of every customer i in period t, given by

$$D_{i,t} = g\big(\gamma_{\mathcal{C}(\mathbf{x}_{i,t})} + \beta_{\mathcal{C}(\mathbf{x}_{i,t})} p_{i,t} + \boldsymbol{\alpha}_{\mathcal{C}(\mathbf{x}_{i,t})}^{\mathsf{T}} \mathbf{x}_{i,t}\big) + \varepsilon_{i,t} \tag{1}$$

for $i \in \mathcal{N}$ and $t \in \mathcal{T}$. In (1), $\mathcal{C}(\mathbf{x}_{i,t})$ is the label of the cluster where $\mathbf{x}_{i,t}$ resides, which is unknown to the seller; $\gamma_{\mathcal{C}(\mathbf{x}_{i,t})} \in \mathbb{R}$, $\beta_{\mathcal{C}(\mathbf{x}_{i,t})} \in \mathbb{R}$, and $\boldsymbol{\alpha}_{\mathcal{C}(\mathbf{x}_{i,t})} \in \mathbb{R}^d$ are cluster-specific consumption parameters, which are also unknown to the seller; $g : \mathbb{R} \to \mathbb{R}$ is a function known to the seller; and $\varepsilon_{i,t}$ is the unobservable and idiosyncratic consumption shock of customer i in period t. This model endows each cluster with its

own consumption parameters. For cluster $k \in \{1,\dots,\mathcal{K}\}$, the consumption parameters are expressed as a vector $\boldsymbol{\theta}_k$ satisfying $\boldsymbol{\theta}_k^\intercal = [\gamma_k \ \beta_k \ \boldsymbol{\alpha}_k^\intercal]$. Here, γ_k is a parameter associated with the baseline consumption, β_k represents the price-sensitivity of the customers, and $\boldsymbol{\alpha}_k$ captures how the features affect consumption in cluster k. In particular, $\boldsymbol{\alpha}_k^\intercal = [\boldsymbol{\alpha}_{k,1}^\intercal \ \boldsymbol{\alpha}_{k,2}^\intercal \ \boldsymbol{\alpha}_{k,3}^\intercal]$, where $\boldsymbol{\alpha}_{k,1}$, $\boldsymbol{\alpha}_{k,2}$, and $\boldsymbol{\alpha}_{k,3}$ are the coefficients of $\mathbf{x}_{t,1}$, $\mathbf{x}_{i,2}$, and $\mathbf{x}_{i,t,3}$, respectively. For $k \in \{1,\dots,\mathcal{K}\}$, the consumption parameter vector $\boldsymbol{\theta}_k$ lies in a compact rectangle $\Theta \subset \mathbb{R}^{d+2}$. We denote the collection of all consumption parameter vectors by $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_\mathcal{K}\}$.

We make minimal assumptions on $g(\cdot)$ in (1). Specifically, we let $g(\cdot)$ be any function that is twice continuously differentiable and increasing. These properties are satisfied by numerous commonly used functions, such as the exponential, inverse-logit, inverse-probit, and identity functions. All in all, (1) provides substantial model flexibility by allowing for general non-linear relationships between consumption and price.

We assume that the sequence of idiosyncratic consumption shocks, namely $\{\varepsilon_{i,t} : i \in \mathcal{N}, t \in \mathcal{T}\}$, follows a sub-Gaussian martingale difference sequence. More precisely, letting $m = (t-1)N + i$ for any $(i,t) \in \mathcal{N} \times \mathcal{T}$, we have $\mathbb{E}[\varepsilon_m | \mathcal{F}_{m-1}] = 0$, and there exist positive constants σ_0 and η_0 such that $\mathbb{E}[\varepsilon_m^2 | \mathcal{F}_{m-1}] < \sigma_0^2$ and $\mathbb{E}[\exp(\eta\varepsilon_m)|\mathcal{F}_{m-1}] < \infty$ for all η with $|\eta| < \eta_0$, where $\mathcal{F}_m = \sigma(p_1,\dots,p_m,\varepsilon_1,\dots,\varepsilon_m,\mathbf{X}_1,\dots,\mathbf{X}_{m+1})$ for all m. Note that we explicitly allow the consumption shocks to depend on the history of prices and features to capture potential serial correlations.

In practice, d is usually large, and consumption may be determined by a small subset of features. Thus, it is useful to account for model sparsity for estimation accuracy and scalability. Accordingly, we consider a sparse consumption model as follows: let $\mathcal{S}_k$ be the support of $\boldsymbol{\theta}_k$, i.e., the set of indices corresponding to the non-zero elements of $\boldsymbol{\theta}_k$. Define $\mathcal{S} = \bigcup_{k=1}^{\mathcal{K}} \mathcal{S}_k$, and assume that the cardinality of $\mathcal{S}$ is at most $\bar{s}$.

In the context of our real-life data set, about half of the energy provided by the seller is generated by inflexible resources in nuclear and coal power plants, meaning that the marginal cost of production is fixed (Austin Energy 2018). Moreover, the vast majority of the remaining energy is from renewable energy sources, whose marginal cost of production is negligible (Austin Energy 2018). Thus, the seller's marginal cost of production is essentially constant. Since prices can be treated as unit profit margins on top of the marginal cost of production, we normalize the marginal cost of production to zero without loss of generality, and use "profit" and "revenue" interchangeably. Given the underlying cluster structure $\mathcal{C}(\cdot)$, the consumption parameter vectors $\{\boldsymbol{\theta}_k^\intercal = [\gamma_k \ \beta_k \ \boldsymbol{\alpha}_k^\intercal] : k = 1,\dots,\mathcal{K}\}$, and the feature vector realizations $\{\mathbf{x}_{1,t},\dots,\mathbf{x}_{N,t}\}$, the seller's expected single period profit function for customer i in period t is given by

$$\mathcal{R}(p_{i,t};\boldsymbol{\theta}_{\mathcal{C}(\mathbf{x}_{i,t})},\mathbf{x}_{i,t}) = p_{i,t}\, g\big(\gamma_{\mathcal{C}(\mathbf{x}_{i,t})} + \beta_{\mathcal{C}(\mathbf{x}_{i,t})} p_{i,t} + \boldsymbol{\alpha}_{\mathcal{C}(\mathbf{x}_{i,t})}^\intercal \mathbf{x}_{i,t}\big). \tag{2}$$

Let $\varphi(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{x}_{i,t})},\mathbf{x}_{i,t}) = \operatorname{argmax}_p\{\mathcal{R}(p;\boldsymbol{\theta}_{\mathcal{C}(\mathbf{x}_{i,t})},\mathbf{x}_{i,t})\}$ be the profit-maximizing price with the corresponding maximum expected profit $\mathcal{R}^*(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{x}_{i,t})},\mathbf{x}_{i,t}) = \mathcal{R}(\varphi(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{x}_{i,t})},\mathbf{x}_{i,t});\boldsymbol{\theta}_{\mathcal{C}(\mathbf{x}_{i,t})},\mathbf{x}_{i,t})$. We assume that $\varphi(\boldsymbol{\theta},\mathbf{x})$ is in the interior of $\mathcal{P}$ and that $\mathcal{R}(p;\boldsymbol{\theta},\mathbf{x})$ has non-vanishing second derivatives with respect to price p at $\varphi(\boldsymbol{\theta},\mathbf{x})$ for the underlying consumption parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_\mathcal{K}\}$ and all $\mathbf{x} \in \mathcal{X}$.

As mentioned above, (2) captures the case with constant marginal production (or procurement) cost, which is in line with the operations of many U.S. utilities, including Austin Energy. For example, any utility with inflexible generation facilities incurs a constant marginal cost of production. Moreover, typically, any utility that does not own generation facilities but procures energy from independent power producers also incurs a constant marginal cost of procurement. In practice, a common way to procure energy for many years is to sign a long-term contract with an independent power producer (FERC 2007, CEC 2018). Under these contracts, procurement price is generally fixed (Hausman et al. 2008), implying a constant marginal procurement cost for utilities. There are many U.S. utilities whose primary source of energy is constant-price contracts. For instance, MCE (2018) and PCE (2017) meet the vast majority of their customer demands through these contracts.

We also note that even if the utility's marginal production (procurement) cost changes over time, the cost is observable to the utility, and thus, our analysis holds when the marginal cost of production (procurement) is time-varying. Therefore, our results are also valid for utilities that procure energy from wholesale electricity markets or are subject to volatile procurement costs. Furthermore, §7.2 extends our analysis to accommodate more general cost functions and establishes the robustness of our results.

2.2. Admissible Policies and Performance Metric

Because the seller does not know the underlying cluster structure $\mathcal{C}(\cdot)$ or the consumption parameter vectors $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_\mathcal{K}\}$, it has to make pricing decisions based on observed information. Let $\mathbf{I}_t$ be the collection of all past observations until right before the pricing decision in period $t+1$, i.e., $\mathbf{I}_t = \{\mathbf{p}_1, \ldots, \mathbf{p}_t, \mathbf{D}_1, \ldots, \mathbf{D}_t, \mathbf{X}_1, \ldots, \mathbf{X}_{t+1}\}$ for $t = 1, 2, \ldots,$ where $\mathbf{p}_s = [p_{1,s} \ \cdots \ p_{N,s}]^\intercal$, $\mathbf{D}_s = [D_{1,s} \ \cdots \ D_{N,s}]^\intercal$, and $\mathbf{X}_s = \{\mathbf{X}_{1,s}, \ldots, \mathbf{X}_{N,s}\}$ for all s, and $\mathbf{I}_0 = \mathbf{X}_1$. An *admissible policy* is a sequence of functions $\pi = \{\pi_t : t \in \mathcal{T}\}$, where $\pi_t : \mathbb{R}^{(2(t-1)+td)N} \mapsto \mathcal{P}^N$ is a measurable function such that $\pi_t(\mathbf{I}_{t-1}) = \mathbf{p}_t$ for all $t \in \mathcal{T}$. The set of all admissible policies is denoted by Π.

Given an admissible policy $\pi \in \Pi$, the underlying cluster structure $\mathcal{C}(\cdot)$, and the consumption parameter vectors $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k^\intercal = [\gamma_k \ \beta_k \ \boldsymbol{\alpha}_k^\intercal] : k = 1, \ldots, \mathcal{K}\} \in \Theta^\mathcal{K}$, we define a probability measure $\mathbb{P}_{\boldsymbol{\theta}}^\pi(\cdot)$ on the space of consumption sequences such that

$$\mathbb{P}_{\boldsymbol{\theta}}^\pi(\mathbf{D}_1 \in \mathrm{d}\tilde{\mathbf{D}}_1, \ldots, \mathbf{D}_T \in \mathrm{d}\tilde{\mathbf{D}}_T) = \prod_{t=1}^{T} \prod_{i=1}^{N} \mathbb{P}_\varepsilon \left\{ g\left(\gamma_{\mathcal{C}(\mathbf{X}_{i,t})} + \beta_{\mathcal{C}(\mathbf{X}_{i,t})} p_{i,t} + \boldsymbol{\alpha}_{\mathcal{C}(\mathbf{X}_{i,t})}^\intercal \mathbf{X}_{i,t}\right) + \varepsilon_{i,t} \in \tilde{D}_{i,t} \middle| \mathbf{I}_{t-1} \right\}$$

for $\tilde{\mathbf{D}}_1, \ldots, \tilde{\mathbf{D}}_T \in \mathbb{R}^N$, where $\mathbb{P}_\varepsilon(\cdot)$ is the probability measure governing the consumption shocks $\{\varepsilon_{i,t} : i \in \mathcal{N}, t \in \mathcal{T}\}$. Let $\mathbb{E}_{\boldsymbol{\theta}}^\pi(\cdot)$ be the expectation operator associated with $\mathbb{P}_{\boldsymbol{\theta}}^\pi(\cdot)$. Define the seller's *average conditional regret* under policy $\pi \in \Pi$ as the average profit loss over N customers and T periods relative to a clairvoyant who knows $\mathcal{C}(\cdot)$ and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k : k = 1, \ldots, \mathcal{K}\}$, which is given by

$$\Delta_{\boldsymbol{\theta}}^\pi(N, T; \mathbf{X}) = \frac{1}{NT} \mathbb{E}_{\boldsymbol{\theta}}^\pi \left\{ \sum_{t=1}^{T} \sum_{i=1}^{N} [\mathcal{R}^*(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t}) - \mathcal{R}(p_{i,t}^\pi; \boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t})] \middle| \mathbf{X} \right\},$$

where $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$ and $p_{i,t}^{\pi}$ is the price for customer i in period t under policy π. Define $\mathbb{P}_{\mathbf{X}}(\cdot)$ as the product probability measure on the space of $\mathbf{X}$, and $\mathbb{E}_{\mathbf{X}}(\cdot)$ as the corresponding expectation operator. The seller aims to minimize the *average expected regret*

$$\Delta_{\boldsymbol{\theta}}^{\pi}(N,T) = \mathbb{E}_{\mathbf{X}}[\Delta_{\boldsymbol{\theta}}^{\pi}(N,T;\mathbf{X})]$$

by choosing an admissible policy $\pi \in \Pi$. Regret is a widely adopted performance metric in the related literature (see, e.g., Keskin and Zeevi 2014, Qi et al. 2017, Feng et al. 2020). For notational brevity, we also let $\mathbb{E}_{\mathbf{X},\boldsymbol{\theta}}^{\pi}(\cdot) = \mathbb{E}_{\mathbf{X}}\{\mathbb{E}_{\boldsymbol{\theta}}^{\pi}(\cdot|\mathbf{X})\}$, and $\mathbb{P}_{\mathbf{X},\boldsymbol{\theta}}^{\pi}(\cdot)$ be the probability measure associated with $\mathbb{E}_{\mathbf{X},\boldsymbol{\theta}}^{\pi}(\cdot)$.

3. Preliminaries on Spectral Graph Theory

This section presents preliminary information on spectral graph theory, which is useful for describing our data-driven policy based on spectral clustering and feature-based pricing. Consider a generic feature space $\mathcal{Z}$ equipped with a probability measure $\mathbb{P}_{\mathcal{Z}}(\cdot)$. This feature space is segmented into K clusters, whose underlying structure is unknown. Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be n i.i.d. observations drawn from $\mathbb{P}_{\mathcal{Z}}(\cdot)$, and $\psi : \mathcal{Z} \times \mathcal{Z} \mapsto [0, \infty)$ be a nonnegative, real-valued, symmetric, and continuous function that gives the similarity between a pair of observations. Based on these, we can form an undirected graph by treating the n observations as nodes and their pairwise similarities as weighted edges. Intuitively, any pair of feature observations with larger similarity is more likely to be drawn from the same cluster in the underlying cluster structure. Given the observations $\mathbf{z}_1, \ldots, \mathbf{z}_n$, construct an $n \times n$ similarity matrix $\mathcal{Q}_n$ whose $(j,\ell)^{\text{th}}$ entry is $\psi(\mathbf{z}_j, \mathbf{z}_\ell)$ for $j, \ell \in \{1, \ldots, n\}$. Let $\mathcal{D}_n$ be a diagonal matrix whose j^{th} diagonal entry is $\sum_{\ell=1}^{n} \psi(\mathbf{z}_j, \mathbf{z}_\ell)$, namely the j^{th} row sum of $\mathcal{Q}_n$, for $j \in \{1, \ldots, n\}$. The j^{th} diagonal entry of $\mathcal{D}_n$ represents the cumulative similarity of $\mathbf{z}_j$ (i.e., node j) to all observations (i.e., all nodes) and is called the *degree* of $\mathbf{z}_j$. The following definition provides a normalized measure of similarity.

DEFINITION 1. The normalized Laplacian based on n observations is $\mathcal{L}_n = \boldsymbol{I}_n - \mathcal{D}_n^{-1/2} \mathcal{Q}_n \mathcal{D}_n^{-1/2}$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

The normalized Laplacian $\mathcal{L}_n$ encodes pairwise similarities among feature observations, using degrees for normalization. For $j, \ell \in \{1, \ldots, n\}$, the $(j,\ell)^{\text{th}}$ entry of $\mathcal{L}_n$ is $\mathbb{I}\{j = \ell\} - \psi(\mathbf{z}_j, \mathbf{z}_\ell)/\sqrt{d_j d_\ell}$, where $d_j = \sum_{m=1}^{n} \psi(\mathbf{z}_j, \mathbf{z}_m)$ and $d_\ell = \sum_{m=1}^{n} \psi(\mathbf{z}_\ell, \mathbf{z}_m)$ are the degrees of nodes j and ℓ, respectively. Thus, $\mathcal{L}_n$ depends on the choice of ψ through the similarity matrix $\mathcal{Q}_n$. A key property of $\mathcal{L}_n$ is the following.

FACT 1. (von Luxburg 2007, Proposition 3) *The smallest eigenvalue of $\mathcal{L}_n$ is 0. The eigenvector corresponding to this eigenvalue is $\mathcal{D}_n^{1/2}\mathbf{1}_n$, where $\mathbf{1}_n$ denotes the $n \times 1$ vector of all ones.*

To express further properties of $\mathcal{L}_n$, let a connected component of an undirected graph be a set A such that (i) there are no edges with positive weight according to ψ between any node in A and any node in the complement of A, and (ii) all nodes in A are either directly joined by an edge or through some intermediate nodes in A. That is, nodes within a connected component are similar to each other while nodes belonging to different connected components are distinct from each other. Based on this, we express another key property of $\mathcal{L}_n$ as follows.

FACT 2. (von Luxburg 2007, Proposition 4) *Let $A_1, \ldots, A_m$ be connected components of the undirected graph characterized by ψ. The algebraic multiplicity of the eigenvalue 0 of $\mathcal{L}_n$ is equal to the number of these connected components. The corresponding eigenspace is spanned by the vectors $\{\mathcal{D}_n^{1/2}\mathbb{I}_{A_i} : i = 1, \ldots, m\}$, where $\mathbb{I}_{A_i}$ is an indicator vector whose j^{th} component is $\mathbb{I}\{\mathbf{z}_j \in A_i\}$.*

Thus, if we view the connected components of the undirected graph as the estimated clusters based on n observations, then certain eigenvalues and eigenvectors of $\mathcal{L}_n$ can be used to estimate the underlying cluster structure. In the case of a perfect separation of clusters (i.e., when the corresponding undirected graph has multiple separate connected components), the eigenvectors of $\mathcal{L}_n$ corresponding to the eigenvalue 0 are given by the indicators of clusters to which the features belong, scaled by the square root of the degrees. In the case of an almost perfect separation where similarities between connected components are close to but not exactly 0, $\mathcal{L}_n$ can be treated as a perturbed version of a Laplacian with perfect separation, and the eigenvectors of $\mathcal{L}_n$ serve as an approximation to the eigenvectors in the perfect separation case. Based on perturbation theory, the smaller the perturbation from the Laplacian in the perfectly separated case and the farther away an eigenvalue λ is from the rest of the spectrum, the closer the eigenspace corresponding to λ in the case of imperfect separation is to that in the perfect separation case.

4. Data-Driven Clustering and Feature-Based Pricing Policy

In this section, we design our policy based on spectral clustering with feature-based pricing. For the similarity function in spectral clustering, we use the Gaussian kernel. Because there are three different types of features in our setting, we consider three different similarity functions; that is, given $b \in \{1, 2, 3\}$, we let $\psi_b : \mathcal{X}_b \times \mathcal{X}_b \mapsto [0, \infty)$ such that $\psi_b(\mathbf{x}, \mathbf{x}') = \exp(-||\mathbf{x} - \mathbf{x}'||_2^2/\sigma_b^2)$ for $\mathbf{x}, \mathbf{x}' \in \mathcal{X}_b$, where σ_b is a tuning parameter. Let $M \in \{2, \ldots, T\}$, $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \sigma_3]^\intercal$ be a vector of real numbers, and $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_k : k = 1, \ldots, \bar{\mathcal{K}}\}$ be a set of vectors of positive real numbers, where $\boldsymbol{\lambda}_k = [\lambda_{k,1}\ \lambda_{k,2}\ \lambda_{k,3}]^\intercal$. Our data-driven clustering and feature-based pricing policy with parameters $M, \boldsymbol{\sigma}, \boldsymbol{\lambda}$, denoted by $\hat{\pi}(M, \boldsymbol{\sigma}, \boldsymbol{\lambda})$, is as follows:[4]

Data-driven clustering and feature-based pricing policy $\hat{\pi}(M, \boldsymbol{\sigma}, \boldsymbol{\lambda})$

Step 1: Price experimentation. Generate a set of test prices $\{p_{i,t} : i = 1, \ldots, N, t = 1, \ldots, M\}$, drawn uniformly at random from $\mathcal{P}$. Charge the prices $\{p_{i,t} : i = 1, \ldots, N\}$ in period $t = 1, \ldots, M$. Observe the feature vectors $\mathbf{x}_{i,t}$ and the consumptions $D_{i,t}$ for all $i = 1, \ldots, N$, $t = 1, \ldots, M$.

Step 2: Estimation of the underlying cluster structure.

(a) Construct three normalized Laplacians $\mathcal{L}_M$, $\mathcal{L}_N$, and $\mathcal{L}_{NM}$ for the three types of features. Specifically, $\mathcal{L}_M$ is based on M time-heterogeneous feature observations, $\{\mathbf{x}_{t,1} : t = 1, \ldots, M\}$; $\mathcal{L}_N$ is based on N user-heterogeneous feature observations, $\{\mathbf{x}_{i,2} : i = 1, \ldots, N\}$; and $\mathcal{L}_{NM}$ is based on NM fully heterogeneous feature observations, $\{\mathbf{x}_{i,t,3} : i = 1, \ldots, N, t = 1, \ldots, M\}$.

[4] See also Appendix A for a numerical example illustrating how our policy works.

(b) Compute the first $\bar{\mathcal{K}}_1$ eigenvectors $\mathbf{v}_{1,1},\ldots,\mathbf{v}_{\bar{\mathcal{K}}_1,1}$ of $\mathcal{L}_M$, the first $\bar{\mathcal{K}}_2$ eigenvectors $\mathbf{v}_{1,2},\ldots,\mathbf{v}_{\bar{\mathcal{K}}_2,2}$ of $\mathcal{L}_N$, and the first $\bar{\mathcal{K}}_3$ eigenvectors $\mathbf{v}_{1,3},\ldots,\mathbf{v}_{\bar{\mathcal{K}}_3,3}$ of $\mathcal{L}_{NM}$, corresponding to the eigenvalues in increasing order. For $b\in\{1,2,3\}$, let $\boldsymbol{V}_b=[\mathbf{v}_{1,b}\ \cdots\ \mathbf{v}_{\bar{\mathcal{K}}_b,b}]$. The rows of $\boldsymbol{V}_b$ correspond to the observed feature vectors in the eigenspace of the Laplacians. Normalize each row of $\boldsymbol{V}_b$ by its Euclidean norm. The resulting matrix is denoted by $\tilde{\boldsymbol{V}}_b$ for $b\in\{1,2,3\}$.

(c) For $b\in\{1,2,3\}$, classify the rows of $\tilde{\boldsymbol{V}}_b$ into $\bar{\mathcal{K}}_b$ clusters using k-means clustering, up to some permutation of cluster labels. For all $i\in\mathcal{N}$ and $t\in\{1,\ldots,M\}$, compute the estimated cluster label of the feature observation $\mathbf{x}_{i,t}$ as follows: if the three types of features in $\mathbf{x}_{i,t}$ have cluster labels $\hat{k}_1$, $\hat{k}_2$, and $\hat{k}_3$, respectively, then the estimated cluster label of $\mathbf{x}_{i,t}$ is $\hat{\mathcal{C}}(\mathbf{x}_{i,t})=(\hat{k}_1-1)\bar{\mathcal{K}}_2\bar{\mathcal{K}}_3+(\hat{k}_2-1)\bar{\mathcal{K}}_3+\hat{k}_3$, that is, the linear index of the label vector $(\hat{k}_1,\hat{k}_2,\hat{k}_3)$.

Step 3: Estimation of the consumption parameters.

(a) For each estimated cluster $k\in\{1,\ldots,\bar{\mathcal{K}}\}$, define the *quasi-likelihood* (Wedderburn 1974) of the consumption parameter $\tilde{\boldsymbol{\theta}}$ given a lasso-regularization penalty $\tilde{\lambda}$ and observations $\mathbf{u}=\{\mathbf{u}_{i,t}:i=1,\ldots,N,\,t=1,\ldots,M\}$ as

$$Q_k(\tilde{\boldsymbol{\theta}};\tilde{\lambda},\mathbf{u})=\sum_{i=1}^{N}\sum_{t=1}^{M}\hat{\chi}_{i,t}(k)\int_{D_{i,t}}^{g(\tilde{\boldsymbol{\theta}}^{\mathsf{T}}\mathbf{u}_{i,t})}\frac{D_{i,t}-y}{\nu(y)}\mathrm{d}y-\tilde{\lambda}||\tilde{\boldsymbol{\theta}}||_1, \tag{3}$$

where $\hat{\chi}_{i,t}(k)=\mathbb{I}\{\hat{\mathcal{C}}(\mathbf{x}_{i,t})=k\}$, $\nu(y)=g'(g^{-1}(y))$ for $y\in\mathbb{R}$, and $||\tilde{\boldsymbol{\theta}}||_1$ denotes the ℓ_1-norm of $\tilde{\boldsymbol{\theta}}$. Perform a *parallel partial regression with lasso regularization* for each feature type separately to obtain the maximum quasi-likelihood estimators for the estimated cluster k as follows:

$$\hat{\boldsymbol{\vartheta}}_{k,b}(\lambda_{k,b})=\mathrm{argmax}_{\tilde{\boldsymbol{\theta}}}\left\{Q_k(\tilde{\boldsymbol{\theta}};\lambda_{k,b},\mathbf{u}_b)\right\}\ \text{ for } b\in\{1,2,3\},$$

where $\mathbf{u}_b=\{\mathbf{u}_{i,t,b}:i=1,\ldots,N,\,t=1,\ldots,M\}$ for $b\in\{1,2,3\}$, with $\mathbf{u}_{i,t,1}^{\mathsf{T}}=[1\ \mathbf{x}_{i,t,1}^{\mathsf{T}}]$, $\mathbf{u}_{i,t,2}^{\mathsf{T}}=[1\ \mathbf{x}_{i,t,2}^{\mathsf{T}}]$, and $\mathbf{u}_{i,t,3}^{\mathsf{T}}=[1\ p_{i,t}\ \mathbf{x}_{i,t,3}^{\mathsf{T}}]$.

(b) For all $k\in\{1,\ldots,\bar{\mathcal{K}}\}$, extract the estimated coefficients $\hat{\boldsymbol{\alpha}}_{k,1}(\lambda_{k,1})$ of $\{\mathbf{x}_{i,t,1}\}$ from $\hat{\boldsymbol{\vartheta}}_{k,1}(\lambda_{k,1})$, the estimated coefficients $\hat{\boldsymbol{\alpha}}_{k,2}(\lambda_{k,2})$ of $\{\mathbf{x}_{i,t,2}\}$ from $\hat{\boldsymbol{\vartheta}}_{k,2}(\lambda_{k,2})$, and the estimated coefficients $\hat{\boldsymbol{\alpha}}_{k,3}(\lambda_{k,3})$ of $\{\mathbf{x}_{i,t,3}\}$, $\hat{\beta}_k(\lambda_{k,3})$ of $\{p_{i,t}\}$, and $\hat{\gamma}_{k,3}(\lambda_{k,3})$ for the baseline consumption from $\hat{\boldsymbol{\vartheta}}_{k,3}(\lambda_{k,3})$. Compute an estimate $\hat{\gamma}_k(\boldsymbol{\lambda}_k)$ of the baseline consumption term in (1) as

$$\hat{\gamma}_k(\boldsymbol{\lambda}_k)=\frac{1}{N_k}\sum_{i=1}^{N}\sum_{t=1}^{M}\hat{\chi}_{i,t}(k)\left[\hat{\gamma}_{k,3}(\lambda_{k,3})-\hat{\boldsymbol{\alpha}}_{k,1}(\lambda_{k,1})^{\mathsf{T}}\mathbf{x}_{i,t,1}-\hat{\boldsymbol{\alpha}}_{k,2}(\lambda_{k,2})^{\mathsf{T}}\mathbf{x}_{i,t,2}\right],$$

where $N_k=\sum_{i=1}^{N}\sum_{t=1}^{M}\hat{\chi}_{i,t}(k)$ is the number of observations that are classified to cluster k.

(c) For all $k\in\{1,\ldots,\bar{\mathcal{K}}\}$, let $\hat{\boldsymbol{\vartheta}}_k(\boldsymbol{\lambda}_k)^{\mathsf{T}}=[\hat{\gamma}_k(\boldsymbol{\lambda}_k)\ \ \hat{\boldsymbol{\alpha}}_{k,1}(\lambda_{k,1})^{\mathsf{T}}\ \ \hat{\boldsymbol{\alpha}}_{k,2}(\lambda_{k,2})^{\mathsf{T}}\ \ \hat{\boldsymbol{\alpha}}_{k,3}(\lambda_{k,3})^{\mathsf{T}}\ \ \hat{\beta}_k(\lambda_{k,3})]$, and $\hat{\boldsymbol{\theta}}_k(\boldsymbol{\lambda}_k)=\mathrm{Proj}_{\Theta}(\hat{\boldsymbol{\vartheta}}_k(\boldsymbol{\lambda}_k))$, where $\mathrm{Proj}_{\Theta}(\cdot)$ denotes the projection onto Θ.

Step 4: Prediction, cluster updating, and pricing for profit maximization in the remaining periods.

(a) At the beginning of each period $t=M+1,\ldots,T$, observe the new features $\mathbf{x}_{i,t}^{\mathsf{T}}=[\mathbf{x}_{t,1}^{\mathsf{T}}\ \mathbf{x}_{i,2}^{\mathsf{T}}\ \mathbf{x}_{i,t,3}^{\mathsf{T}}]$ for $i\in\mathcal{N}$. Construct the normalized Laplacians $\mathcal{L}_t$ based on $\{\mathbf{x}_{s,1}:s=1,\ldots,t\}$ and $\mathcal{L}_{Nt}$ based on $\{\mathbf{x}_{i,s,3}:i=1,\ldots,N,s=1,\ldots,t\}$.

(b) Recompute the first $\bar{\mathcal{K}}_1$ eigenvectors $\mathbf{v}_{1,1},\ldots,\mathbf{v}_{\bar{\mathcal{K}}_1,1}$ of $\mathcal{L}_t$ and the first $\bar{\mathcal{K}}_3$ eigenvectors $\mathbf{v}_{1,3},\ldots,\mathbf{v}_{\bar{\mathcal{K}}_3,3}$ of $\mathcal{L}_{Nt}$, corresponding to the eigenvalues in increasing order. For $b\in\{1,3\}$, update $\boldsymbol{V}_b = [\mathbf{v}_{1,b}\ \cdots\ \mathbf{v}_{\bar{\mathcal{K}}_b,b}]$, and normalize each row by its Euclidean norm to get $\tilde{\boldsymbol{V}}_b$.

(c) For $b\in\{1,3\}$, classify the rows of $\tilde{\boldsymbol{V}}_b$ into $\bar{\mathcal{K}}_b$ clusters using k-means clustering. For all $i\in\mathcal{N}$ in period $t\in\{M+1,\ldots,T\}$, if $\mathbf{x}_{t,1}$ and $\mathbf{x}_{i,t,3}$ have cluster labels $\tilde{k}_1$ and $\tilde{k}_3$, respectively, then classify $\mathbf{x}_{i,t}$ to cluster $\tilde{k}=(\tilde{k}_1-1)\bar{\mathcal{K}}_2\bar{\mathcal{K}}_3+(\tilde{k}_2-1)\bar{\mathcal{K}}_3+\tilde{k}_3$ with $\tilde{k}_2=\hat{k}_2$.

(d) Charge the profit-maximizing price $p_{i,t}=\text{Proj}_{\mathcal{P}}(\varphi(\hat{\boldsymbol{\theta}}_{\tilde{k}}(\boldsymbol{\lambda}_{\tilde{k}}),\mathbf{x}_{i,t}))$ for customer i in period t, where $\text{Proj}_{\mathcal{P}}(\cdot)$ denotes the projection onto $\mathcal{P}$. Proceed to the next period unless $t=T$.

In Step 4, we combine new features with all historical observations to recompute the normalized Laplacians. In this way, we dynamically refine the estimated cluster structure in each period, and use it to obtain the predicted cluster labels for the new feature observations in the current period.[5]

5. Theoretical Analysis

In this section, we derive theoretical performance guarantees for our policy. First, we analyze the misclassification error of spectral clustering on a generic compact feature space $\mathcal{Z}$. Recall that $\hat{\mathcal{C}}(\cdot)$ is the estimated cluster function that maps any feature to its estimated cluster label. The following proposition characterizes this error for a random sample of size n.

PROPOSITION 1. (MISCLASSIFICATION ERROR) *There exist positive constants K_1 and K_2 such that for any random feature sample $\mathbf{Z}=\{\mathbf{Z}_1,\ldots,\mathbf{Z}_n\}$,*

$$\mathbb{P}_{\mathbf{Z}}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{\hat{\mathcal{C}}(\mathbf{Z}_i)\neq\mathcal{C}(\mathbf{Z}_i)\}\geq K_1\bar{\mathcal{K}}\sqrt{\frac{\log n}{n}}\right\}\leq\frac{K_2\bar{\mathcal{K}}}{n},$$

where $\mathbb{P}_{\mathbf{Z}}(\cdot)$ is the product probability measure over the space of $\mathbf{Z}$.

REMARK 1. Without loss of generality, the estimated cluster function, $\hat{\mathcal{C}}(\cdot)$, is such that the number of mismatches between the true and estimated cluster labels is the smallest over all permutations of labels.

Proposition 1 states that the average misclassification error based on n feature observations is in the order of $1/\sqrt{n}$ and at most linear in $\bar{\mathcal{K}}$, i.e., the upper bound on the number of clusters. To our knowledge, this is the first result that identifies the rate of misclassification error of spectral clustering when the number of clusters is known only up to an upper bound. The derivation of Proposition 1 is based on functional analysis and the eigendecomposition of the Laplacian $\mathcal{L}_n$ described in §3. In the proof of this result, we construct a linear operator that corresponds to $\mathcal{L}_n$, and analyze the rate at which the eigenfunctions of this operator converge to the eigenfunctions of a limit operator that represents the underlying cluster structure.

[5] In Step 4, there is no need to perform cluster updating for user-heterogeneous features because they do not change over time. However, our policy can handle the case of dynamic observations on user-heterogeneous features. Suppose that a new household with the user-heterogeneous feature $\tilde{\mathbf{x}}_{i,2}$ moves to the area served by the electric utility company. In this case, we can perform Steps 4(a)-(c) accounting for $\tilde{\mathbf{x}}_{i,2}$ to obtain the predicted class label $\tilde{k}_2$.

The eigenvectors of $\mathcal{L}_n$ characterize the estimated cluster labels, and are given by the eigenfunctions of the operator corresponding to $\mathcal{L}_n$ evaluated at the realized features. Based on this, we show how the estimated cluster labels converge to the true cluster labels, and that the average misclassification error among at most $\bar{\mathcal{K}}$ clusters converges to zero at a rate of order $1/\sqrt{n}$.

Next, we characterize parameter estimation errors. The details of this analysis depend on the different types of features explained in §2. For this reason, we consider two settings: (i) all features are fully heterogeneous over both time and users; (ii) the features are of mixed type, including time- and user-heterogeneous features, as well as fully heterogeneous ones.

5.1. Analysis for Fully Heterogeneous Features

We first consider a setting where all features are fully heterogeneous. Note that if $\bar{\mathcal{K}}$ is strictly greater than the number of underlying clusters, $\mathcal{K}$, then without loss of generality, we generate some dummy clusters to make the number of clusters match $\bar{\mathcal{K}}$. One way to achieve this is to set the underlying parameter vectors of these dummy clusters all equal to the parameter vector of one of the original $\mathcal{K}$ clusters. In this way, $\boldsymbol{\theta}_k$ is defined for all $k \in \{1,\dots,\bar{\mathcal{K}}\}$.

Based on the classification of features into clusters, the following proposition characterizes the estimation error for the consumption parameter vector of each estimated cluster.

PROPOSITION 2. (PARAMETER ESTIMATION ERROR WITH FULLY HETEROGENEOUS FEATURES) *Let* $\pi = \hat{\pi}(M,\sigma,\boldsymbol{\lambda})$, *where* $M \in \{2,\dots,T\}$, $\sigma \in \mathbb{R}$, *and* $\boldsymbol{\lambda} = \{\lambda_k : k = 1,\dots,\bar{\mathcal{K}}\}$ *with* $\lambda_k = \tilde{c}_k\sqrt{NM\log(dNM)}$ *and* $\tilde{c}_k > 0$. *Then, there exist positive constants* K_3 *and* K_4 *such that*

$$\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\hat{\boldsymbol{\theta}}_k(\lambda_k)-\boldsymbol{\theta}_k\|_2^2 \geq \frac{K_3\bar{s}\log(dNM)}{NM}\right\} \leq \frac{K_4\bar{s}\log(dNM)}{NM} \text{ for } k\in\{1,\dots,\bar{\mathcal{K}}\} \text{ and } N\in\{2,3,\dots\}.$$

REMARK 2. Because there is only one type of feature in the context of Proposition 2, it is sufficient to use a single tuning parameter σ and a single regularization parameter λ_k for each cluster k under $\hat{\pi}(M,\sigma,\boldsymbol{\lambda})$. The constants K_3 and K_4 in Proposition 2 are characterized in the proof of this proposition. Similarly, for each subsequent result, the constants in the result are provided in the associated proof.

Proposition 2 shows that when the features are all i.i.d. over both time and users, the squared estimation error of the quasi-likelihood estimator for each cluster converges to zero roughly at a rate of order $1/(NM)$. As explained earlier, this estimation uses lasso regularization and is performed without knowing the underlying sparsity in the true consumption parameter vectors $\{\boldsymbol{\theta}_k\}$.

The proof of Proposition 2 is based on deriving a lower bound on the minimum eigenvalues of the empirical Fisher information matrix and an upper bound on the interaction terms between the consumption shocks and the observed features; see Lemmas 1 and 2 in Appendix C. Building on these, we construct a probabilistic upper bound on the parameter estimation error by expressing this error in terms of the quasi-likelihood function and applying the aforementioned lemmas. In contrast to the parametric estimation results

in the literature on dynamic pricing and learning, deriving Proposition 2 involves two additional challenges. First, parameter estimation in our setting is directly affected by the misclassification error. An estimated cluster does not necessarily include all observations that should have been classified in that cluster, and may be contaminated by observations that should have been classified in other clusters. To address this challenge, we quantify the estimation error due to misclassification, and derive a probabilistic upper bound on this error. Second, the number of observations classified in each cluster is a random quantity, whereas in earlier work, the number of observations at any particular period is typically deterministic due to the absence of clustering. For this challenge, we derive concentration inequalities to obtain an upper bound on the event that the number of observations in an estimated cluster is small, and thereby prove the convergence of parameter estimation with high probability. By addressing these two additional challenges, Proposition 2 extends the parametric estimation results in the related literature to account for misclassification errors.

Combining the above results, we derive the following upper bound on the average regret when all features are fully heterogeneous.

THEOREM 1. (UPPER BOUND ON REGRET WITH FULLY HETEROGENEOUS FEATURES) *Let* $\pi = \hat{\pi}(M, \sigma, \boldsymbol{\lambda})$, *where* $M = \lceil \kappa\sqrt{T/N} \rceil \vee 2$, $\kappa \geq 1$, $\sigma \in \mathbb{R}$, *and* $\boldsymbol{\lambda} = \{\lambda_k : k = 1, \ldots, \bar{\mathcal{K}}\}$ *with* $\lambda_k = \tilde{c}_k \sqrt{NM \log(dNM)}$ *and* $\tilde{c}_k > 0$. *Then, there exists a positive constant* K_5 *such that*

$$\Delta^{\pi}_{\boldsymbol{\theta}}(N, T) \leq \frac{K_5 \bar{\mathcal{K}} \bar{s} \log(dNT)}{\sqrt{NT}} \text{ for } N, T \in \{2, 3, \ldots\}.$$

Theorem 1 states that when the features are fully heterogeneous, the average regret of our policy vanishes at a rate of order $1/\sqrt{NT}$. This result follows from Propositions 1 and 2. For a time horizon of T periods, Proposition 1 implies that the average misclassification error is of order $1/\sqrt{NT}$, and Proposition 2 implies that the average parameter estimation error is also of order $1/\sqrt{NT}$ (up to logarithmic terms). Thus, the average regret due to both errors converges at a rate essentially in the order of $1/\sqrt{NT}$. This is the fastest achievable convergence rate of regret in our setting of fully heterogeneous features. The reason is as follows. Keskin and Zeevi (2014) present a lower bound on regret in a special case where there is a single underlying cluster and the seller makes a single observation with no features in each period. That lower bound implies that the average regret after n observations is at least of order $1/\sqrt{n}$ (Keskin and Zeevi 2014, Theorem 1). In our setting, where the seller observes N fully heterogeneous features in each period and there are multiple clusters in the underlying cluster structure, Theorem 1 shows that our policy $\hat{\pi}(M, \sigma, \boldsymbol{\lambda})$ is rate-optimal in the sense that it achieves the best convergence rate of regret over NT observations, up to logarithmic terms.

Theorem 1 indicates that the average regret of our policy is at most linear in $\bar{\mathcal{K}}$, the upper bound on the number of clusters. We now provide a guideline for choosing $\bar{\mathcal{K}}$. According to spectral graph theory, when there are $\mathcal{K}$ well-separated underlying clusters, the first $\mathcal{K}$ eigenvalues of the normalized Laplacian, sorted in increasing order, would be close to 0. In contrast, the $(\mathcal{K}+1)^{\text{st}}$ eigenvalue would be substantially greater. In this case, one can choose $\bar{\mathcal{K}} = \mathcal{K}$ by looking for this "eigen-gap" in the spectrum of the normalized

Laplacian. This is known as the *eigen-gap heuristic* (von Luxburg 2007, p. 410). However, if the underlying clusters are not well-separated, the aforementioned eigen-gap can become less prominent, making it difficult to determine the number of clusters. In this case, one can choose $\bar{K}$ such that the first $\bar{K}$ eigenvalues include all the ones that are close to 0, and the $(\bar{K}+1)^{\text{st}}$ eigenvalue is substantially greater than 0. Based on this, our approach uses $\bar{K}$ as an input (instead of requiring the exact number of clusters), thereby providing more flexibility in the policy design and implementation. As noted in §1.2, it is also possible to use other criteria such as Davies-Bouldin index, silhouette value, and Dunn index to find a suitable number of clusters.

5.2. Analysis for Features of Mixed Types

We now consider our general setting with all three types of features. Since the combination of all three feature types does not satisfy the i.i.d. feature-sampling assumption, the method and analysis developed for fully heterogeneous features cannot be applied to this setting. Specifically, characterizing the parameter estimation error becomes a key challenge in this general setting, and the literature does not offer any methods that can directly address this challenge. Our policy overcomes this challenge by devising a new method, i.e., parallel partial regression. Accordingly, as described in §4, our policy performs three separate regressions over $\{\mathbf{X}_{t,1} : t = 1,\dots,M\}$, $\{\mathbf{X}_{i,2} : i = 1,\dots,N\}$, and $\{[p_{i,t}\ \mathbf{X}_{i,t,3}^\intercal]^\intercal : i = 1,\dots,N,\ t = 1,\dots,M\}$. Recall that the consumption parameter vector of cluster k is $\boldsymbol{\theta}_k^\intercal = [\gamma_k\ \beta_k\ \boldsymbol{\alpha}_{k,1}^\intercal\ \boldsymbol{\alpha}_{k,2}^\intercal\ \boldsymbol{\alpha}_{k,3}^\intercal]$, where γ_k is a parameter associated with baseline consumption, and β_k, $\boldsymbol{\alpha}_{k,1}$, $\boldsymbol{\alpha}_{k,2}$, $\boldsymbol{\alpha}_{k,3}$ are the coefficients of $\{p_{i,t}\}$, $\{\mathbf{X}_{t,1}\}, \{\mathbf{X}_{i,2}\}, \{\mathbf{X}_{i,t,3}\}$, respectively. The corresponding lasso-regularized quasi-likelihood estimator of $\boldsymbol{\theta}_k$ is $\hat{\boldsymbol{\theta}}_k(\boldsymbol{\lambda}_k)^\intercal = [\hat{\gamma}_k(\boldsymbol{\lambda}_k)\ \hat{\beta}_k(\lambda_{k,3})\ \hat{\boldsymbol{\alpha}}_{k,1}(\lambda_{k,1})^\intercal\ \hat{\boldsymbol{\alpha}}_{k,2}(\lambda_{k,2})^\intercal\ \hat{\boldsymbol{\alpha}}_{k,3}(\lambda_{k,3})^\intercal]$, where $\boldsymbol{\lambda}_k = [\lambda_{k,1}\ \lambda_{k,2}\ \lambda_{k,3}]^\intercal$ is the vector of regularization parameters. Given these, we have the following result.

PROPOSITION 3. (PARAMETER ESTIMATION ERROR WITH MIXED FEATURES) *Let* $\pi = \hat{\pi}(M, \boldsymbol{\sigma}, \boldsymbol{\lambda})$, *where* $M \in \{2,\dots,T\}$, $\boldsymbol{\sigma} \in \mathbb{R}^3$, *and* $\boldsymbol{\lambda} = \{(\lambda_{k,1}\ \lambda_{k,2}\ \lambda_{k,3}) : k = 1,\dots,\bar{K}\}$ *with* $\lambda_{k,1} = \tilde{c}_{k,1}\sqrt{M\log(d_1 M)}$, $\lambda_{k,2} = \tilde{c}_{k,2}\sqrt{N\log(d_2 N)}$, $\lambda_{k,3} = \tilde{c}_{k,3}\sqrt{NM\log(d_3 NM)}$, *and* $\tilde{c}_{k,1}, \tilde{c}_{k,2}, \tilde{c}_{k,3} > 0$. *Then, there exist positive constants* K_6 *and* K_7 *such that*

$$\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\hat{\boldsymbol{\theta}}_k(\boldsymbol{\lambda}_k) - \boldsymbol{\theta}_k\|_2^2 \geq \frac{K_6 \bar{s}\log(dNM)}{N \wedge M}\right\} \leq \frac{K_7 \bar{s}\log(dNM)}{N \wedge M} \quad \textit{for } k \in \{1,\dots,\bar{K}\} \textit{ and } N \in \{2,3,\dots\}.$$

Proposition 3 shows that in the case of mixed feature types, the squared estimation errors of the quasi-likelihood estimators are at most in the order of $1/(N \wedge M)$. Specifically, the squared estimation errors are of order $1/M$, $1/N$ and $1/(NM)$ for the regressions over time-, user- and fully heterogeneous features, respectively. The overall convergence rate is determined by the slowest convergence rate among the three separate regressions.

Performing the partial regressions in parallel segregates different types of features so that the features used in each regression are i.i.d. This allows us to theoretically analyze the convergence rates of the parameter estimation errors for different feature types. It also provides flexibility in choosing different lasso-regularization parameters that depend on the effective sample size of each feature type, thereby improving the estimation accuracy.

Based on the previous results, we now prove a general upper bound on the average regret of our policy in the presence of all three feature types.

THEOREM 2. (UPPER BOUND ON REGRET WITH MIXED FEATURES) *Let* $\pi = \hat{\pi}(M, \boldsymbol{\sigma}, \boldsymbol{\lambda})$, *where* $M = \lceil \kappa\sqrt{T} \rceil \vee 2$, $\kappa \geq 1$, $\boldsymbol{\sigma} \in \mathbb{R}^3$, *and* $\boldsymbol{\lambda} = \{(\lambda_{k,1}\ \lambda_{k,2}\ \lambda_{k,3}) : k = 1, \ldots, \bar{\mathcal{K}}\}$ *with* $\lambda_{k,1} = \tilde{c}_{k,1}\sqrt{M \log(d_1 M)}$, $\lambda_{k,2} = \tilde{c}_{k,2}\sqrt{N \log(d_2 N)}$, $\lambda_{k,3} = \tilde{c}_{k,3}\sqrt{NM \log(d_3 NM)}$, *and* $\tilde{c}_{k,1}, \tilde{c}_{k,2}, \tilde{c}_{k,3} > 0$. *Then, there exists a positive constant* K_8 *such that*

$$\Delta^{\pi}_{\boldsymbol{\theta}}(N,T) \leq \frac{K_8 \bar{\mathcal{K}} \bar{s} \log(dNT)}{\sqrt{N \wedge T}} \ \textit{for}\ N, T \in \{2, 3, \ldots\}. \tag{4}$$

Theorem 2 states that in general, the average regret of our policy vanishes at a rate of order $1/\sqrt{N \wedge T}$. This result is a direct consequence of Propositions 1 and 3. By Proposition 1, the misclassification error is in the order of either $1/\sqrt{T}$ or $1/\sqrt{N}$, whichever is larger. Moreover, by Proposition 3, the parameter estimation error is in the order of either $1/\sqrt{T}$ or $1/N$, whichever converges to zero more slowly. The reason for this is that the number of observations used for parameter estimation is of order $\sqrt{T}$ for the time-heterogeneous features and of order N for the user-heterogeneous features. Consequently, the convergence rate of the average regret is determined by the misclassification error.

Comparing Theorems 1 and 2, we observe that the presence of time- and user-heterogeneous features slows down the convergence rate from order $1/\sqrt{NT}$ to $1/\sqrt{N \wedge T}$. This is because these two types of features bear repetitive information over time and customers, respectively, imposing additional challenges in the seller's learning of the underlying cluster structure and consumption parameters. This contrast between Theorems 1 and 2 reveals how different forms of feature heterogeneity influence regret performance.

The convergence rate of regret in Theorem 2 is the best one in our setting of mixed features. As mentioned earlier, in a special case of our setting, Keskin and Zeevi (2014) provide a lower bound implying that the average regret after n observations is at least of order $1/\sqrt{n}$ (Keskin and Zeevi 2014, Theorem 1). This means that the average regret would vanish at a rate of order $1/\sqrt{T}$ if all features are time-heterogeneous, and order $1/\sqrt{N}$ if all features are user-heterogeneous. In the worst case, the average regret converges to zero at rate of order $(1/\sqrt{N}) \vee (1/\sqrt{T}) = 1/\sqrt{N \wedge T}$. Thus, in general, our policy $\hat{\pi}(M, \boldsymbol{\sigma}, \boldsymbol{\lambda})$ is rate-optimal in the sense that it achieves the fastest convergence rate of regret, up to logarithmic terms.

5.3. Connection between Inter-cluster Similarity and Regret

In this subsection, we discuss how inter-cluster similarity affects regret. For a pair of clusters $\mathscr{C}_1$ and $\mathscr{C}_2$, define the similarity between $\mathscr{C}_1$ and $\mathscr{C}_2$ as

$$\mathscr{S}(\mathscr{C}_1, \mathscr{C}_2) := \sup\{\psi(\mathbf{x}, \mathbf{x}') : \mathbf{x} \in \mathscr{C}_1, \mathbf{x}' \in \mathscr{C}_2\}.$$

This definition is in the same spirit as the *single linkage* metric (Everitt et al. 2011), which gives the smallest distance (i.e., largest similarity) between clusters. As the goal of clustering is to identify cluster boundaries, we use this metric to measure inter-cluster similarity. If the inter-cluster similarity is 0 for all cluster pairs,

then the number of clusters (or the number of connected components from the perspective of graph theory) equals the multiplicity of the eigenvalue 0 of the normalized Laplacian $\mathcal{L}_n$. As the inter-cluster similarity increases, the multiplicity of the eigenvalue 0 becomes 1 (because there would be only one connected component), and for $k \in \{2, \ldots, \mathcal{K}\}$, the k^{th} smallest eigenvalue (counting for multiplicity) increases and deviates from 0.

To use functional analysis theory, we consider an operator J_n that corresponds to $\mathcal{L}_n$, and let J be the limit of $\{J_n : n \in \mathbb{N}\}$ in the sense of compact convergence, representing the underlying cluster structure.[6] Hence, the eigenvalues and eigenfunctions of $\{J_n : n \in \mathbb{N}\}$ converge to those of J. According to spectral theory, the eigenvalues of $\mathcal{L}_n$ are those of the operator $\mathbf{1} - J_n$, where $\mathbf{1}$ denotes the identity operator (von Luxburg et al. 2008, Proposition 9). Thus, the eigenvalues of $\mathcal{L}_n$ are good approximations to those of the limit operator $\mathbf{1} - J$ for sufficiently large n. Consequently, as the inter-cluster similarity increases, the first $\mathcal{K}$ eigenvalues of $\mathbf{1} - J$ will also increase. Because the largest possible eigenvalue of $\mathbf{1} - J$ is 1, the remaining eigenvalues of $\mathbf{1} - J$ cannot increase as much as the first $\mathcal{K}$ eigenvalues, and thus they become less separated from the rest of the spectrum.

We now state a result formalizing on the above discussion. For a non-zero simple eigenvalue ρ of J with eigenfunction u, let $d_J(\rho)$ denote the distance from ρ to the rest of the spectrum of J. That is, $d_J(\rho) = \min\{|\rho - \rho'| : \rho' \in \text{Spec}(J) \setminus \{\rho\}\}$, where $|\cdot|$ denotes the complex norm and $\text{Spec}(J)$ is the spectrum of J. For $n \in \mathbb{N}$, consider the eigenvalue ρ_n of J_n such that $\{\rho_n : n \in \mathbb{N}\}$ converges to ρ, and let u_n be the eigenfunction corresponding to ρ_n. Define $\mathbb{P}_n$ as the empirical probability measure on a generic compact feature space $\mathcal{Z}$ based on n feature observations, and for brevity, let $\mathbb{P}_n f = \int f \mathrm{d}\,\mathbb{P}_n$ and $\mathbb{P}_{\mathcal{Z}} f = \int f \mathrm{d}\,\mathbb{P}_{\mathcal{Z}}$. Let $\{s_n : n \in \mathbb{N}\}$ be a sequence of signs with $s_n \in \{-1, 1\}$, and $\mathscr{F}$ be as in Definition 10 of von Luxburg et al. (2008). Then, we have the following proposition.

PROPOSITION 4. *For any given non-zero simple eigenvalue ρ of J, the smallest constant $\tilde{C}$ such that $\|s_n u_n - u\|_\infty \leq \tilde{C} \sup_{f \in \mathscr{F}} |\mathbb{P}_n f - \mathbb{P}_{\mathcal{Z}} f|$ is given by $\tilde{C} = 4R_\rho^2(3r_{\mathcal{J}} + 1)/(R_\rho \rho - 1)$, where $R_\rho = 2/d_J(\rho)$, $r_{\mathcal{J}} = \psi_{\max}/\psi_{\min}$, $\psi_{\max} = \max\{\psi(\mathbf{x}, \mathbf{y}) : (\mathbf{x}, \mathbf{y}) \in \mathcal{Z} \times \mathcal{Z}\}$, and $\psi_{\min} = \min\{\psi(\mathbf{x}, \mathbf{y}) : (\mathbf{x}, \mathbf{y}) \in \mathcal{Z} \times \mathcal{Z}\}$.*

Proposition 4 characterizes the value of $\tilde{C}$ that leads to the tightest upper bound on $\|s_n u_n - u\|_\infty$, which controls the misclassification error of spectral clustering and hence the average regret. This consequently indicates an inherent relationship between inter-cluster similarity and regret. As the inter-cluster similarity increases, $d_J(\rho)$ becomes smaller and R_ρ becomes larger. By Proposition 4, this means that the upper bound on $\|s_n u_n - u\|_\infty$ becomes looser. This in turn translates into an increase in the upper bound on the misclassification error of spectral clustering, and hence an increase in the upper bound on average regret. Our analysis shows that this increase is inversely proportional to $d_J(\rho)$.

[6] For formal definitions of J_n and J, see (B.3) and (B.4), respectively, in Appendix B.

6. Computational Results

In this section, we present three case studies that shed light on our analysis. We first run two sets of simulation experiments, and then conduct a case study based on real-life data from the retail electricity sector. The first set of simulation experiments examines how regret performance depends on $\bar{\mathcal{K}}$ (the upper bound on the number of clusters), d (the dimensionality of the features), and $\bar{s}$ (the model sparsity parameter). The second one studies the convergence rate of regret in terms of N and T, and also demonstrates the benefits of active price experimentation and lasso regularization. Based on Theorems 1 and 2, it suffices to consider fully heterogeneous features to investigate the effect of $\bar{\mathcal{K}}$, d, and $\bar{s}$ on regret. For this reason, we use fully heterogeneous features in the first set of experiments, and time- and user-heterogeneous features in the second set. Finally, in our case study based on real-life data, we demonstrate the practical value of our approach.

6.1. Simulation Experiments with Fully Heterogeneous Features

In this subsection, we conduct simulation experiments to examine how $\bar{\mathcal{K}}$, d, and $\bar{s}$ influence regret.

6.1.1. Effect of $\bar{\mathcal{K}}$ on regret. We consider a setting with 10-dimensional fully heterogeneous features that are randomly drawn from a mixture of $\mathcal{K} = 3$ multivariate Gaussian distributions with equal weight on each component. All other model parameters are summarized in Appendix F.1. We implement our policy over $T = 100$ periods, and there are $N = 20$ users in each period. The experimentation phase lasts $M = \lceil \kappa \sqrt{T/N} \rceil$ periods, where $\kappa = 10$. The set of feasible prices is $[5.5, 8]$, and the test prices are drawn uniformly at random from this set. The similarity parameter for the Gaussian kernel is $\sigma = 1$ and the lasso-regularization parameter is $\lambda = 0.3\sqrt{NM \log(NM)}$.

In this setting, we examine the performance of our policy with $\bar{\mathcal{K}}$ varying from 1 to 6. Figure 3 displays the average regret over 100 sample paths for these values of $\bar{\mathcal{K}}$. Because the number of underlying clusters is 3, our policy is correctly specified when $\bar{\mathcal{K}} = 3$, and thus incurs the smallest regret in this case. More importantly, while underestimating the number of clusters leads to a significant increase in regret, overestimation results only in a slight increase. Based on this, an important practical insight is that overestimating the number of clusters is much less harmful than underestimating it.

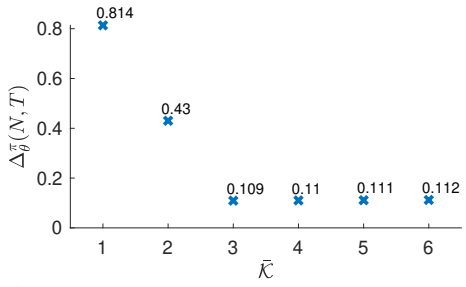


Figure 3 **The effect of $\bar{\mathcal{K}}$ on regret.**

6.1.2. Effect of d on regret. We first consider a base setting with 4-dimensional fully heterogeneous features randomly drawn from a mixture of $\mathcal{K} = 2$ multivariate Gaussian distributions with equal weights. To study how d affects regret, we generate variants of the base setting by gradually increasing d from 4 to 20. The model parameters for these variant settings are in Appendix F.2. To focus on the effect of d on regret, our policy uses $\bar{\mathcal{K}} = 2$ (i.e., there is no misspecification in the upper bound on the number of clusters). The set of feasible prices is $[5.5, 7]$. All other parameters are the same as in §6.1.1. Figure 4a shows how the regret of our policy depends on d. A semi-log regression on the data points in Figure 4a reveals that the average regret of our policy grows logarithmically in d (with $R^2 = 0.985$). This illustrates the theoretical result in Theorem 1.

6.1.3. Effect of $\bar{s}$ on regret. We start with the setting with 20-dimensional features in §6.1.2, and then create variants of that setting with $\bar{s}$ ranging from 10 to 16. The model parameters for these variant settings are in Appendix F.3, and all other parameters are as in §6.1.2. Figure 4b displays how the regret of our policy grows with $\bar{s}$. A linear regression on the displayed data indicates that the average regret of our policy increases linearly in $\bar{s}$ (with $R^2 = 0.996$). This observation is consistent with Theorem 1.

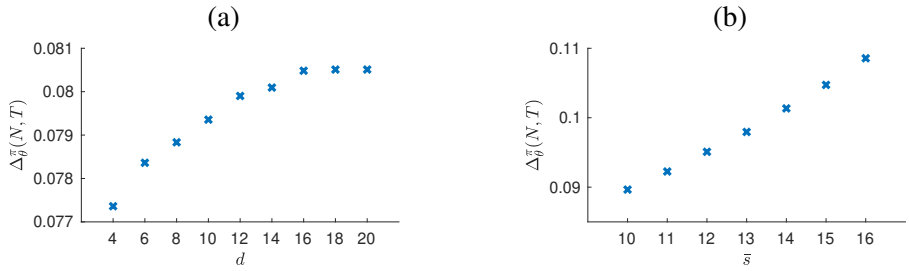


Figure 4 **The effect of d and $\bar{s}$ on regret.**

6.2. Simulation Experiments with Time- and User-heterogeneous Features

We now run a set of experiments to investigate how the average regret of our policy depends on N and T. To examine the convergence rate of regret in T, we fix $N = 2000$ and vary T in the set $\{300, 400, \ldots, 900\}$. After that, to study the convergence rate of regret in N, we fix $T = 2000$ and vary N in the set $\{300, 400, \ldots, 900\}$. All other parameters for these experiments are in Appendix F.4. Figure 5 shows the average regret of our policy for various values of N and T based on 100 sample paths. Performing a log-log regression on the data displayed on each panel of Figure 5, we deduce that the average regret of our policy converges to zero in the order of $T^{-0.50}$ (left panel) and $N^{-0.51}$ (right panel). This verifies the convergence rates in Theorem 2.

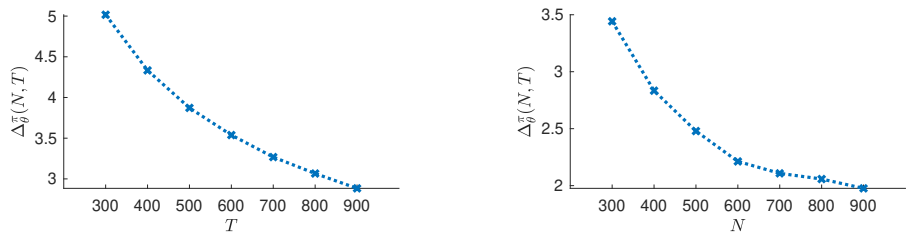


Figure 5 **Average regrets over T and N.**

6.2.1. Value of price experimentation. To see the value of using prices as a learning tool, consider a policy that focuses on short-term profits. Using the first 10 periods for initializing estimation, this policy updates the estimates of the underlying cluster structure and consumption parameters in each period $t \in \{11, \ldots, M\}$, choosing profit-maximizing prices based on the most recent estimates and feature observations. The policy is the same as our policy after M periods. We call this policy the *greedy policy* because it targets immediate profits. To assess the value of price experiments, we compare our policy with the greedy policy. First, we fix $T = 500$ and compare the regret performance of the two policies for $N \in \{200, 300, \ldots, 800\}$. Second, we fix $N = 500$ and make a similar regret comparison for $T \in \{200, 300, \ldots, 800\}$. The lasso-regularization parameters for both policies are $\lambda_{\text{time}} = 0.1\sqrt{M \log(d_1 M)}$ and $\lambda_{\text{user}} = 0.1\sqrt{N \log(d_2 N)}$. All other parameters are as described at the beginning of §6.2.

Figure 6 displays the percentage increase in average regret due to using the greedy policy instead of our policy for different values of T and N. The greedy policy is inferior to our policy in terms of regret: it increases the average regret by about 46.7% on average. This is because, with no price experimentation, the greedy policy could fail to identify the underlying cluster structure and consumption parameters, and thus its prices could get stuck at sub-optimal values. The poor regret performance of the greedy policy emphasizes the practical need for active price experimentation.

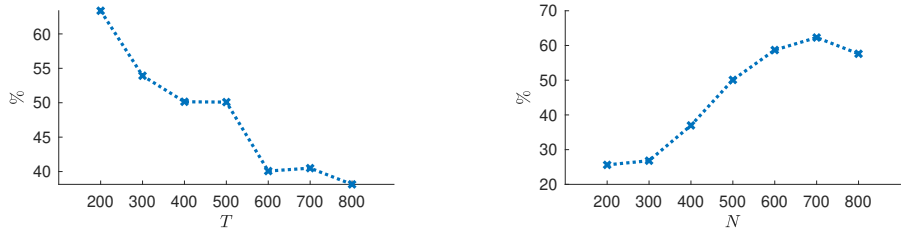


Figure 6 **Percentage increase in average regret due to lack of price experimentation.**

6.2.2. Value of regularization. To examine the impact of lasso regularization on regret performance, we consider a special case of our policy that uses no regularization. This is equivalent to setting $\lambda_{\text{time}} = \lambda_{\text{user}} = 0$ in our policy. All other parameters are as described at the beginning of §6.2 and in §6.2.1. Figure 7 shows the percentage increase in average regret due to using the unregularized policy instead of our policy. For each pair of N and T, our policy outperforms the unregularized policy, and using no regularization results in an increase in average regret by about 0.1% on average. The corresponding increase in cumulative regret is equal to the increase in average regret multiplied by N and T. As a result, the cumulative regret due to not using lasso regularization is significant.

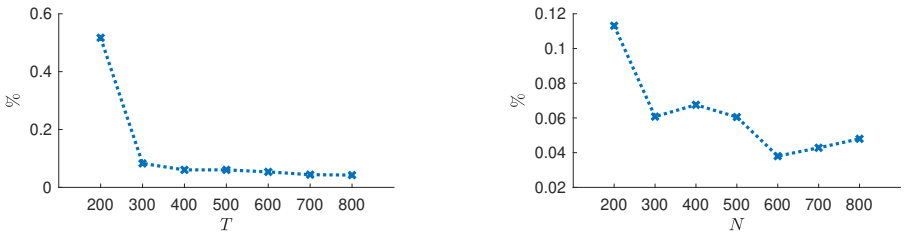


Figure 7 **Percentage increase in average regret due to lack of lasso regularization.**

6.3. Case Study on Real-life Data

We now illustrate the applicability and validity of our policy in the context of a real-life data set.

6.3.1. Description of data. We use the data set of household electricity consumption from Pecan Street Dataport (Pecan Street 2019). This data set contains: (i) smart meter data of electricity consumption in a neighborhood of Austin, Texas, at 15-minute intervals; (ii) audits and surveys gathered from all participating residents, including demographic information, construction-related information, some living habits, such as thermostat temperature in different seasons and times of day, and comments in text on installed equipment and experience; (iii) weather information in Austin, Texas, at 1-hour intervals, including temperature, pressure, humidity, etc. The survey data and the weather data constitute the user- and time-heterogeneous features, respectively.

The local electric utility serving the neighborhood in the data set is Austin Energy. We obtain the retail rate of Austin Energy from the U.S. Utility Rate Database (OpenEI 2019). On top of the base rate, Austin Energy ran price experiments from March 2013 through October 2014 to estimate the electricity consumption behavior of their customers. This program involved "pricing trials," in which some customers received financial incentives at the individual level to shift their consumption to off-peak hours. To take advantage of the price variations in these pricing trials, we focus on the aforementioned time period.

We also account for the utility's marginal cost of production in this case study. The vast majority of the energy provided by Austin Energy is generated by either inflexible resources whose marginal production cost is fixed, or renewable energy sources whose marginal production cost is negligible. Thus, as explained in §2, the marginal cost of production is essentially constant in the context of this data set. According to the 2014 annual performance report of Austin Energy (Austin Energy 2014), the average system fuel cost of Austin Energy was 0.03815 dollars per kWh in 2014.

6.3.2. Data preprocessing. To have the same level of granularity in the consumption and weather data, we define a period to be an hour. For the text information in the survey data, we use natural language processing tools to convert texts into predictor candidates for regression analysis. Specifically, we first tokenize the texts, lemmatize the words, erase punctuation marks, and remove a list of stop words to preprocess the texts. We then create a bag of words based on the preprocessed texts and apply sparse singular value decomposition to the matrix of word counts. The leading eigenvectors whose corresponding eigenvalues are greater than 0.001 are designated as predictor candidates. Via this procedure and other standard statistical methods, we obtain 157 linearly independent predictors with 8 time-heterogeneous and 149 user-heterogeneous features. A full list of features is in Appendix G.

6.3.3. Model calibration. We use the preprocessed data from March 1$^{\text{st}}$ through May 31$^{\text{st}}$ to calibrate our model. Because each period in the model corresponds to one hour in the data set, there are $T = 2207$ periods. There are 129 households (i.e., customers) in the data set. Because the sample size of the time-heterogeneous features is much larger than that of the user-heterogeneous features, prioritizing the time-heterogeneous

features in clustering avoids very small clusters that would lead to excessive noise in estimation. Thus, we calibrate the underlying cluster structure in the space of the time-heterogeneous features, treating the space of the user-heterogeneous features as a single cluster. Specifically, we use spectral clustering with 2, 3, and 4 clusters, and perform a silhouette analysis to calibrate the number of clusters and the underlying cluster structure. The resulting number of clusters is 2, which yields the largest average silhouette value. In each cluster, we use linear regression with backward elimination to calibrate the model parameters. The parameters that are left in the model are all statistically significant with $p \leq 0.05$. The parameters that are dropped have zero values. Out of the 157 parameters in the two clusters, there are 36 zero-valued parameters in the first cluster and 32 in the second one (see Appendix F.5 for the calibrated parameter values). This means that the sparsity in the underlying model parameters is non-trivial.

6.3.4. Results and discussion. We implement our policy using the Gaussian similarity parameter $\sigma = 10$ in spectral clustering and the lasso-regularization penalties $\lambda_1 = 0.05\sqrt{M \log(d_1 M)}$ and $\lambda_2 = 0.05\sqrt{N \log(d_2 N)}$ in the regressions for the time- and user-heterogeneous features, respectively, where $M = \lceil \kappa\sqrt{T} \rceil$, $\kappa = 8$, $N = 111$ (as there are 111 users in the first M periods), $d_1 = 8$, and $d_2 = 149$. To demonstrate the performance of our policy, we compare it with the performance of (i) a policy that ignores the underlying cluster structure, (ii) a policy that avoids feature-based pricing, (iii) the greedy policy that does not use price experimentation (see §6.2.1), (iv) a variant of our policy that does not use lasso regularization (see §6.2.2), and (v) the historical decisions of the electric utility company.[7]

Table 2 compares the percentage increase in average regret due to using the policies (i)-(v) instead of our joint clustering and feature-based pricing policy. The comparisons are all based on the sample path of features observed by the utility company in real life.

Table 2 Percentage Increase in Average Regret Relative to Joint Clustering and Feature-based Pricing

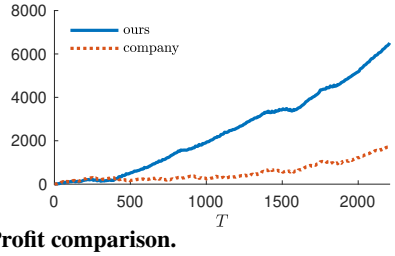
	$T = 1200$	$T = 1700$	$T = 2200$
No clustering	0.39%	7.65%	17.23%
No feature-based pricing	76.86%	65.53%	62.82%
Greedy (no experimentation)	262.00%	195.59%	195.10%
No regularization	3.19%	1.33%	23.84%
Company	125.44%	78.82%	72.97%

Our policy outperforms all of the policies (i)-(v). First, the policy that treats the feature space as a single cluster makes a costly model misspecification error, resulting in larger average regret. Second, avoiding feature-based pricing leads to substantially increased regret, suggesting that using feature information in pricing is key in the context of our real-life data. Third, as in §6.2.1, the greedy policy performs poorly, indicating the necessity of price experimentation in this context. Fourth, as in §6.2.2, using no regularization worsens performance because lasso regularization helps decrease parameter estimation errors, especially when the features are high-dimensional. Fifth, our approach significantly improves upon the utility company's historical decisions.

[7] We also conduct a robustness check with a model calibrated by forward addition; see Appendix H.

We also compare the cumulative profit of our policy with that of the company's historical decisions. Figure 12a shows how these cumulative profits of these policies evolve over time, and Figure 12b shows the cumulative profits of the same policies at the end of the three-month horizon. The results displayed Figure 12 are based on the sample path of features observed by the utility company in real life. Our approach significantly outperforms the company's historical pricing decisions in terms of profits. Compared to the company's historical decisions, our policy increases the cumulative profit by more than 146% over three months (this is based on the smaller of the profit improvements observed in the case-study setting described above and the setting of the robustness check in Appendix H).

(a) cumulative profit over T periods ($)



(b) three-month total profit ($)

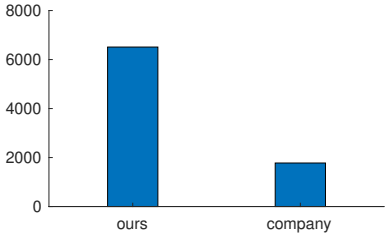


Figure 8 **Profit comparison.**

6.3.5. Impact of model misspecification. To study how model misspecification affects profit performance, we also investigate several additional settings. In the first setting, there are two clusters, and in each cluster, consumption depends only on price. Specifically, as in §6.3.3, we consider a linear model with two clusters over the time-heterogeneous features ($\mathcal{K}_1 = 2$) and a single cluster over the user-heterogeneous features ($\mathcal{K}_2 = 1$). But unlike §6.3.3, all feature coefficients are zero in this setting, as feature variations within a cluster do not influence consumption. As a benchmark, we consider a fully-specified policy that knows the number of clusters and the sparsity structure, and thus uses only clustering without feature-based pricing. Our policy is misspecified in this setting since it assumes a feature-based model and attempts to estimate feature coefficients to be used in pricing. The parameters of our policy are $\sigma_1 = 5$, $\tilde{c}_1 = 0.1$, and $\tilde{c}_2 = 0.5$. Figure 9a shows the cumulative profit of our policy, the benchmark policy, and the company's historical decisions in this setting. Despite the misspecification, our policy performs almost the same as the fully-specified benchmark policy and outperforms the company's historical decisions. This is largely because our policy uses lasso regularization, which controls the variance in parameter estimates.

In the second setting, we consider a feature-based linear model with a single cluster ($\mathcal{K}_1 = \mathcal{K}_2 = 1$) such that consumption depends on both price and features. As a benchmark, we consider a fully-specified policy that knows that there is a single cluster and implements feature-based pricing without clustering. Our policy is misspecified as it assumes that there are two clusters over the time-heterogeneous features. The parameters of our policy are $\sigma_1 = 10$ and $\tilde{c}_1 = \tilde{c}_2 = 0.001$. As shown in Figure 9b, our policy performs essentially the same as the fully-specified benchmark policy and significantly outperforms the company's historical decisions. This suggests that, as in §6.1.1, our policy is robust to the model misspecification caused by overestimating the number of clusters in the context of our real-life data set.

In the third setting, we consider a feature-based linear model with $\mathcal{K}_1 = 2$ and $\mathcal{K}_2 = 1$ as in §6.3.3. As a benchmark, we consider a policy that constructs two clusters over the time-heterogenous features but does not use feature-based pricing. We consider three versions of our policy with $\bar{\mathcal{K}}_1 \in \{3, 4, 5\}$. In each version, the policy parameters are $\sigma_1 = 5$ and $\tilde{c}_1 = \tilde{c}_2 = 0.5$. All policies are misspecified in this setting: the benchmark policy is misspecified because it does not use feature-based pricing with clusters, and the three versions of our policy are misspecified in terms of the number of clusters. As illustrated in Figure 9c, all three versions of our policy perform significantly better than the company's historical decisions, and overestimating the number of clusters does not affect profits substantially. However, failure to use feature-based pricing leads to a large profit loss. This again emphasizes the value of feature-based pricing.

In the fourth setting, we consider a feature-based logit model with $\mathcal{K}_1 = 2$ and $\mathcal{K}_2 = 1$. The link function is the logistic function multiplied by a positive constant representing the market size. As a benchmark, we consider a fully-specified policy that knows the number of underlying clusters as well as the link function and uses clustering with feature-based pricing (the constant multiplying the logistic function is estimated together with other model parameters). Our policy has $\bar{\mathcal{K}}_1 = 2$ and also uses clustering with feature-based pricing, but assumes a linear model. Other policy parameters are $\sigma_1 = 10$ and $\tilde{c}_1 = \tilde{c}_2 = 0.5$. As displayed in Figure 9d, our policy performs better than the benchmark policy most of the time. Our policy also dominates the company's historical decisions significantly, as in the previous settings.

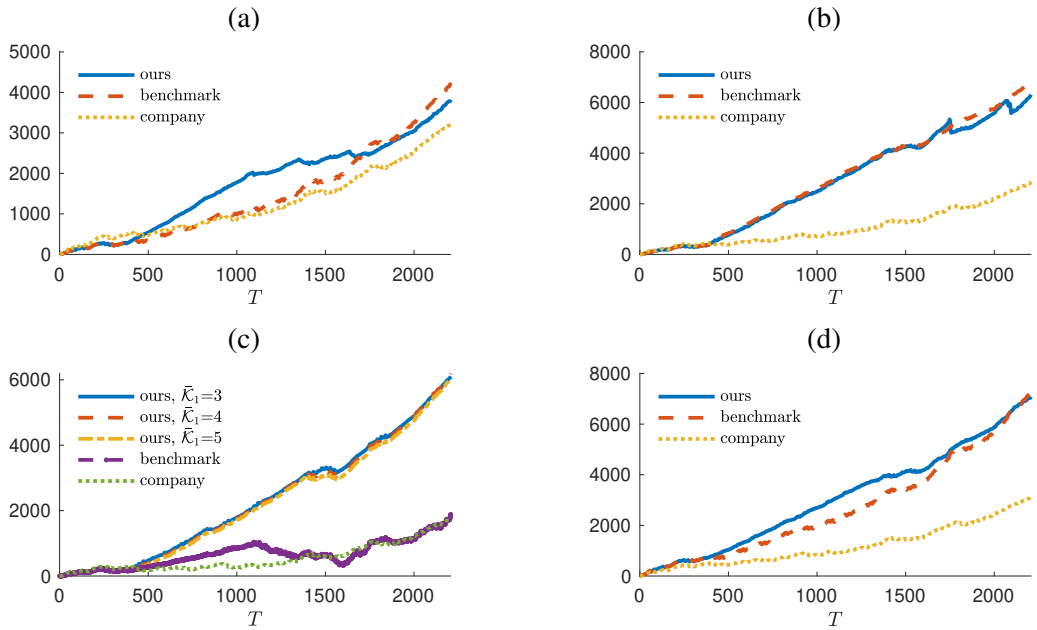


Figure 9 **Impact of model misspecification.** The four panels display the cumulative profits of various policies under four different settings over a three-month period. In the setting of panel (a), there is a linear model with two clusters and consumption depends only on prices and not on features, but our policy uses feature-based pricing. In the setting of panel (b), there is a feature-based linear model with a single cluster, while our policy assumes two clusters. In the setting of panel (c), there is a feature-based linear model with two clusters, but our policy overestimates the number of clusters. In the setting of panel (d), there is a feature-based logit model with two clusters, while our policy assumes a linear model. Our policy is robust to all these forms of misspecification and outperforms the company's historical decisions in all cases.

Combining the results in Table 2 and Figure 9, we see that ignoring either clustering or feature-based pricing when they should be implemented leads to significant profit loss, whereas taking both of them into account while they are redundant does essentially no harm. Based on these findings, we thus conclude that our data-driven joint clustering and feature-based pricing policy achieves good performance even if the underlying model is misspecified in the context of our real-life data set.

7. Extensions

7.1. Non-i.i.d. Features

In this subsection, we extend our base model and analysis to accommodate non-i.i.d. features. Specifically, we consider time-heterogeneous features whose mean can change over time. Other types of features with mean shifts can be analyzed in a similar way. Suppose that the sequence of time-heterogeneous features has a corresponding mean sequence $\{\boldsymbol{\mu}_t : t = 1, \ldots, T\}$ with at most $\mathscr{C} > 1$ change-points over T periods. Formally, there is a sequence of change-points $\{t_j^* : 0, 1, \ldots, \mathscr{C}+1\}$ with $1 = t_0^* < t_1^* < \cdots < t_{\mathscr{C}}^* < t_{\mathscr{C}+1} = T$ and $t_j^* = \inf\{t > t_{j-1}^* : \boldsymbol{\mu}_t \neq \boldsymbol{\mu}_{t_{j-1}^*}\}$. Suppose that any two distinct feature means are at least $\delta_{\min}$ apart for some $\delta_{\min} > 0$, i.e., $\inf\{||\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t'}||_2 : \boldsymbol{\mu}_t \neq \boldsymbol{\mu}_{t'}, 1 \leq t' < t \leq T\} \geq \delta_{\min}$. The seller knows neither the feature means nor the change-points.

We now modify our policy to add change-point detection. We divide the time horizon into cycles of length $\ell < M$, and detect change-points based on the means of feature observations at the end of each cycle. Let $\hat{\eta}_\tau = \sup\{\|\bar{\mathbf{X}}_{\tau,1} - \bar{\mathbf{X}}_{\tau',1}\|_2 \colon \tau' = L(\tau), \ldots, \tau - 1\}$, where $\bar{\mathbf{X}}_{\tau,1}$ is the mean of observed time-heterogeneous features in cycle τ, and $L(\tau) = \sup\{\tau' \leq \tau : \chi_{\tau'} = 1\}$ is the latest detection cycle preceding τ. Then, for any given threshold $\eta > 0$, define the sequence of indicators $\{\chi_\tau : \tau = 0, 1, \ldots, \lfloor T/\ell \rfloor\}$ such that $\chi_0 = 1$ and

$$\chi_{\tau+1} = \begin{cases} 1 & \text{if } \hat{\eta}_\tau > \eta, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Based on (5), we claim a detection in cycle τ if the mean of observations in this cycle differs by η (in Euclidean norm) from those in at least one previous cycle since the latest detection. At the beginning of the time horizon, we run the base policy as described in §4 and perform change-point detection on top of it. Once our policy claims a detection, it drops all past data, restarts the base policy in the next cycle, and keeps detecting change-points. We denote this modified policy as $\hat{\pi}(M, \boldsymbol{\sigma}, \boldsymbol{\lambda}, \eta)$.

Let τ_j^* be the cycle containing the j^{th} change-point. To define the cycle of the first detection claim after the j^{th} change-point, let $\hat{\tau}_j^+ = \inf\{\tau \geq \tau_j^* : \chi_\tau = 1\} \wedge \tau_{j+1}^*$ for $j = 1, \ldots, \mathscr{C}$, with $\hat{\tau}_0^+ = 0$. Similarly, to define the cycle of second detection claim after the j^{th} change-point, let $\hat{\tau}_j^- = \inf\{\tau > \tau_j^+ : \chi_\tau = 1\} \wedge \tau_{j+1}^*$ for $j = 0, 1, \ldots, \mathscr{C}$. If $\hat{\tau}_j^+ > \tau_j^*$, there is a detection delay following the j^{th} change-point, and if $\tau_j^- < \tau_{j+1}^*$, there is an early false alarm before the $(j+1)^{\text{st}}$ change-point. The next proposition establishes how these two types of detection errors contribute to average regret, when all three types of feature heterogeneity are present.

PROPOSITION 5. *Let* $\pi = \hat{\pi}(M, \boldsymbol{\sigma}, \boldsymbol{\lambda}, \eta)$, *where* $M = \lceil \kappa\sqrt{T} \rceil \vee 2$, $\kappa \geq 1$, $\boldsymbol{\sigma} \in \mathbb{R}^3$, *and* $\boldsymbol{\lambda} = \{(\lambda_{k,1} \ \lambda_{k,2} \ \lambda_{k,3}) : k = 1, \ldots, \bar{\mathcal{K}}\}$ *with* $\lambda_{k,1} = \tilde{c}_{k,1}\sqrt{M \log(d_1 M)}$, $\lambda_{k,2} = \tilde{c}_{k,2}\sqrt{N \log(d_2 N)}$, $\lambda_{k,3} = \tilde{c}_{k,3}\sqrt{NM \log(d_3 NM)}$, *and* $\tilde{c}_{k,1}, \tilde{c}_{k,2}, \tilde{c}_{k,3} > 0$. *Denote by* $\mathscr{E}$ *the set of time periods in the price experimentation step of* π.

(a) (REGRET DUE TO DETECTION DELAYS) *There exists a constant* $K_9 > 0$ *such that*

$$\frac{1}{NT}\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\sum_{t=\ell\tau_j^*+1}^{\ell\hat{\tau}_j^+}\sum_{i=1}^{N}[\mathcal{R}^*(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t}) - \mathcal{R}(p^{\pi}_{i,t}; \boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t})]\mathbb{I}\{t \notin \mathscr{E}\}\right\} \leq \frac{K_9 \bar{\mathcal{K}} \bar{s} \log(dNT)}{\sqrt{N \wedge T}}. \tag{6}$$

(b) (REGRET DUE TO EARLY FALSE ALARMS) *There exists a constant* $K_{10} > 0$ *such that*

$$\frac{1}{NT}\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\sum_{t=\ell\hat{\tau}_j^-+1}^{\ell\tau_{j+1}^*}\sum_{i=1}^{N}[\mathcal{R}^*(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t}) - \mathcal{R}(p^{\pi}_{i,t}; \boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t})]\mathbb{I}\{t \notin \mathscr{E}\}\right\} \leq \frac{K_{10}}{\sqrt{T}}. \tag{7}$$

For any $j \in \{1, \ldots, \mathscr{C}\}$, there could be multiple change-points between cycles $L(\tau_j^*)$ and τ_j^*. These changes contaminate the consumption observations between cycles $L(\tau_j^*)$ and τ_j^+. This translates into a more involved analysis of the regret caused by detection delays. On the other hand, there is no change-point between cycles τ_j^- and τ_{j+1}^*. As a result, the regret caused by early false alarms depends only on the tail distribution of the demeaned features. Based on Proposition 5, Theorem 3 derives an upper bound on average regret in the presence of mean-shifting time-heterogeneous features.

THEOREM 3. (UPPER BOUND ON REGRET WITH MEAN-SHIFTING FEATURES) *Let* $\pi = \hat{\pi}(M, \boldsymbol{\sigma}, \boldsymbol{\lambda}, \eta)$, *where* $M = \lceil \kappa\sqrt{T} \rceil \vee 2$, $\kappa \geq 1$, $\boldsymbol{\sigma} \in \mathbb{R}^3$, *and* $\boldsymbol{\lambda} = \{(\lambda_{k,1} \ \lambda_{k,2} \ \lambda_{k,3}) : k = 1, \ldots, \bar{\mathcal{K}}\}$ *with* $\lambda_{k,1} = \tilde{c}_{k,1}\sqrt{M \log(d_1 M)}$, $\lambda_{k,2} = \tilde{c}_{k,2}\sqrt{N \log(d_2 N)}$, $\lambda_{k,3} = \tilde{c}_{k,3}\sqrt{NM \log(d_3 NM)}$, *and* $\tilde{c}_{k,1}, \tilde{c}_{k,2}, \tilde{c}_{k,3} > 0$. *Then, there exists a positive constant* K_{11} *such that*

$$\Delta^{\pi}_{\boldsymbol{\theta}}(N, T) \leq \frac{K_{11} \bar{\mathcal{K}} \bar{s} \log(dNT)}{\sqrt{N \wedge T}} \ \textit{for} \ N, T \in \{2, 3, \ldots\}. \tag{8}$$

The preceding theorem shows that the average regret of our modified policy $\hat{\pi}(M, \boldsymbol{\sigma}, \boldsymbol{\lambda}, \eta)$ converges to zero in the order of $\sqrt{N \wedge T}$, which is the same as the convergence rate in Theorem 2.

7.2. General Cost Functions

This subsection extends our base model and analysis to allow for more general cost functions for the seller. Recall that our base model assumes that the seller's marginal cost of production/procurement is constant. Below, we present an extended analysis in which the seller has a piecewise linear cost function with a single kink. Furthermore, in Appendix K, we study the extensions to piecewise linear cost functions with multiple kinks and general differentiable cost functions. At the end of this subsection, we also explain how our analysis extends when the seller's cost function is time-varying.

Suppose that the seller incurs cost $C(q)$ when it produces or procures a total of q units of energy, where

$$C(q) = \begin{cases} a_1 q + b_1 & \text{if } q < \tilde{q}, \\ a_2 q + b_2 & \text{if } q \geq \tilde{q}. \end{cases} \tag{9}$$

Here, $\tilde{q}$ is a kink in the piecewise linear cost function, and a_1, b_1, a_2 and b_2 are cost parameters such that $a_1, a_2, b_1 \geq 0$ (this condition ensures that the cost function is non-negative and increasing in q). Let $\mathscr{K} = \{\varkappa = (a_1, a_2, b_1, b_2, \tilde{q}) : a_1, a_2, b_1 \geq 0,\ \tilde{q} \in [\underline{q}, \overline{q}],\ \overline{q} > \underline{q} > 0\}$ be the feasible set of cost parameters. If $a_1 > a_2$ and $b_1 < b_2$, then $C(\cdot)$ is concave; if $a_1 < a_2$ and $b_1 > b_2$, then $C(\cdot)$ is convex.

For analytical convenience, we change the decision variable from price to expected consumption. For $i \in \mathcal{N}$ and $t \in \mathcal{T}$, given the underlying cluster structure $\mathcal{C}(\cdot)$, the consumption parameter vectors $\{\boldsymbol{\theta}_k^{\intercal} = [\gamma_k\ \beta_k\ \boldsymbol{\alpha}_k^{\intercal}] : k = 1, \ldots, \mathcal{K}\}$, the price $p_{i,t}$, and the feature vector realization $\mathbf{x}_{i,t}$, the expected consumption of customer i in period t is $\tilde{d}_{i,t} = g(\gamma_{\mathcal{C}(\mathbf{x}_{i,t})} + \beta_{\mathcal{C}(\mathbf{x}_{i,t})} p_{i,t} + \boldsymbol{\alpha}_{\mathcal{C}(\mathbf{x}_{i,t})}^{\intercal} \mathbf{x}_{i,t})$. Because the feasible price set $\mathcal{P}$, the feature space $\mathcal{X}$, and the parameter space Θ are all compact, and the function $g(\cdot)$ is continuous, $\tilde{d}_{i,t}$ lies in a closed interval $\mathscr{D}$ for all $i \in \mathcal{N}$ and $t \in \mathcal{T}$. Suppose that $\{\sum_{i=1}^{N} \varepsilon_{i,t} : t = 1, \ldots, T\}$ are i.i.d. with a density $f(\cdot)$ that belongs to the exponential family of distributions, i.e., $f(e; \boldsymbol{\phi}) = B(e) \exp[\boldsymbol{\phi}^{\intercal} \mathbf{T}(e) - A(\boldsymbol{\phi})]$, where $A(\cdot), B(\cdot)$ and $\mathbf{T}(\cdot)$ are known smooth functions with $B(e) > 0$ for all e, and $\boldsymbol{\phi}$ is an unknown parameter chosen from a compact set Φ in the natural parameter space $\{\boldsymbol{\phi} : A(\boldsymbol{\phi}) < \infty\}$. Let $F(\cdot)$ be the cumulative distribution function associated with the density $f(\cdot)$.

Given an expected consumption vector $\tilde{\mathbf{d}}_t = (\tilde{d}_{1,t}, \ldots, \tilde{d}_{N,t})$, the parameters $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, $\varkappa$, and the feature observation $\mathbf{x}_t$, let $\mathscr{R}(\tilde{\mathbf{d}}_t; \boldsymbol{\theta}, \boldsymbol{\phi}, \varkappa, \mathbf{x}_t)$ denote the seller's total expected profit in period t over all customers. For tractability, we consider the case where $g(\cdot)$ is the identity function. In this case, Proposition 6 proves the existence and uniqueness of a solution to a system of non-linear equations, and establishes that this solution is the unconstrained optimizer of $\mathscr{R}(\tilde{\mathbf{d}}_t; \boldsymbol{\theta}, \boldsymbol{\phi}, \varkappa, \mathbf{x}_t)$.

PROPOSITION 6. *Suppose that* $(a_1 - a_2) \bar{f} N |\beta_{\min}| < 2$, *where* $\bar{f} = \sup\{f(e; \boldsymbol{\phi}) : e \in \mathbb{R}, \boldsymbol{\phi} \in \Phi\}$. *If* $g(\cdot)$ *is the identity function, then there exists a unique solution* $\tilde{\mathbf{d}}_t^* \in \mathbb{R}^N$ *to the following system of non-linear equations:*

$$\beta_{\mathcal{C}(\mathbf{x}_{j,t})}^{-1} \left[(g^{-1})'(\tilde{d}_{j,t})\, \tilde{d}_{j,t} + g^{-1}(\tilde{d}_{j,t}) - \gamma_{\mathcal{C}(\mathbf{x}_{j,t})} - \boldsymbol{\alpha}_{\mathcal{C}(\mathbf{x}_{j,t})}^{\intercal} \mathbf{x}_{j,t} \right] - a_2 - (a_1 - a_2) F\left(\tilde{q} - \sum_{i=1}^{N} \tilde{d}_{i,t}; \boldsymbol{\phi} \right) = 0 \tag{10}$$

for $j = 1, \ldots, N$. *This solution is also the unique unconstrained optimizer of* $\mathscr{R}(\tilde{\mathbf{d}}_t; \boldsymbol{\theta}, \boldsymbol{\phi}, \varkappa, \mathbf{x}_t)$.

REMARK 3. If $C(\cdot)$ is convex, then $(a_1 - a_2) \bar{f} N |\beta_{\min}| < 0 < 2$; i.e., the condition in Proposition 6 holds.

We now turn to the analysis of regret. Given the parameters $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, $\varkappa$, and the feature observation $\mathbf{x}_t$, let $\mathscr{R}^*(\boldsymbol{\theta}, \boldsymbol{\phi}, \varkappa, \mathbf{x}_t) = \mathscr{R}(\tilde{\mathbf{d}}_t^*; \boldsymbol{\theta}, \boldsymbol{\phi}, \varkappa, \mathbf{x}_t)$ be the optimal profit in period t, and suppose that $\tilde{\mathbf{d}}_t^*$ is in the interior of $\mathscr{D}^N$. Then, the average expected regret of any admissible policy π is

$$\Delta_{\boldsymbol{\theta}}^{\pi}(N; T) = \frac{1}{NT} \mathbb{E}_{\mathbf{X}, \boldsymbol{\theta}}^{\pi} \left\{ \sum_{t=1}^{T} [\mathscr{R}^*(\boldsymbol{\theta}, \boldsymbol{\phi}, \varkappa, \mathbf{X}_t) - \mathscr{R}(\tilde{\mathbf{d}}_t^{\pi}; \boldsymbol{\theta}, \boldsymbol{\phi}, \varkappa, \mathbf{X}_t)] \right\}, \tag{11}$$

where $\tilde{\mathbf{d}}_t^{\pi}$ is the vector of expected consumption vector chosen by the policy π in period t. The next theorem characterizes an upper bound on the average regret of our policy for piecewise linear cost functions.

THEOREM 4. (UPPER BOUND ON REGRET WITH MORE GENERAL COST FUNCTIONS) *Suppose that the conditions in Proposition 6 are satisfied. Let* $\pi = \hat{\pi}(M, \boldsymbol{\sigma}, \boldsymbol{\lambda})$, *where* $M = \lceil \kappa\sqrt{T} \rceil \vee 2$, $\kappa \geq 1$, $\boldsymbol{\sigma} \in \mathbb{R}^3$, *and* $\boldsymbol{\lambda} = \{(\lambda_{k,1}\ \lambda_{k,2}\ \lambda_{k,3}) : k = 1, \ldots, \bar{\mathcal{K}}\}$ *with* $\lambda_{k,1} = \tilde{c}_{k,1}\sqrt{M \log(d_1 M)}$, $\lambda_{k,2} = \tilde{c}_{k,2}\sqrt{N \log(d_2 N)}$, $\lambda_{k,3} = \tilde{c}_{k,3}\sqrt{NM \log(d_3 NM)}$, *and* $\tilde{c}_{k,1}, \tilde{c}_{k,2}, \tilde{c}_{k,3} > 0$. *Then, there exists a positive constant* K_{12} *such that*

$$\Delta_{\boldsymbol{\theta}}^{\pi}(N,T) \leq \frac{K_{12}\bar{\mathcal{K}}\bar{s}\log(dNT)}{\sqrt{N \wedge T}} \ \textit{for}\ N, T \in \{2, 3, \ldots\}. \tag{12}$$

Theorem 4 shows that our upper bound on average regret in Theorem 2 still holds provided that the cost functions satisfy the condition ensuring the existence and uniqueness of a solution to the above non-linear system of equations. This result can be further extended to piecewise linear cost functions with multiple kinks and even general differentiable cost functions, as long as the corresponding conditions are met; see Appendix K. Moreover, we can also accommodate time-varying cost functions by simply adding a time index to the cost parameter vector $\varkappa$. In this case, our results continue to hold if $(a_{1,t} - a_{2,t})\bar{f}N|\beta_{\min}| < 2$ for all t.

7.3. Lagged Price Effect and Lagged Consumption Effect

In addition to the factors in (1), there might be other factors that impact the customers' energy consumption. For example, a customer's current consumption might be influenced by previous retail electricity prices. This is because, to avoid a high price in a given period, the customer could shift some of her consumption to a later period. This is referred to as the lagged price effect.[8] Similarly, a customer's current consumption level could depend on her previous consumption levels. That is, low consumption in one period might signal high consumption in a later period—this is called the *lagged consumption effect*. In this subsection, we extend our model and algorithm to capture the lagged price effect and lagged consumption effect.

In our model, the lagged price (consumption) effect can be captured by including a past price (past consumption) as an additional variable in (1). However, this extended model is analytically intractable.[9] Thus, instead of theoretical analysis, we design an extended algorithm in a more general setting and numerically test the performance of this algorithm by using our real-life data set on Austin Energy's pricing experiment. To this end, we first generate two variables, *lagged price* and *lagged consumption*. Regarding the former, the data set provides prices for each user in each period. Based on these prices, every 24-hour block is

[8] Lagged price effects may be relevant in different sectors; see, e.g., Cohen et al. (2017) for an application in retail.

[9] In this extension, identifying the optimal pricing policy requires solving a high-dimensional dynamic program. Note that the pricing decisions in a period directly impact the consumption in future periods. To compute the expected future payoff of a pricing decision, the seller needs to construct an empirical distribution of features because the underlying feature distribution is unknown. For this purpose, the seller must keep track of a large number of feature realizations, meaning that that the dynamic program characterizing the optimal policy would have a large number of state variables (e.g., there are 157 features in our real-life data set). This makes the problem intractable in general.

naturally divided into two time intervals: 6:00 a.m.–10:00 p.m. and 10:00 p.m.–6:00 a.m. the next day. This is because the prices from 10:00 p.m.–6:00 a.m. the next day are much lower than the ones in all other hours for pricing trial participants, although they are subject to fluctuating prices. Hence, to capture any possible consumption-shifting behaviors in line with the experiment design, we split each day into two time intervals, and calculate each user's average price during each time interval. We call this variable the lagged price. Then, we extend the model such that in any period, each user's consumption depends not only on the price and features in that period, but also on that user's average price during the previous time interval. The lagged consumption variable and the associated model extension are constructed in a similar fashion.

To calibrate our extended model with the lagged price, we consider a linear model and apply backward elimination as in §6.3.3. Because the lagged price is typically correlated with the current price, we use a two-stage regression to first identify the current price effect and feature coefficients for each cluster, and then regress the residuals over the lagged price to estimate the lagged price effect. We call this model the *lagged price model*. We calibrate the *lagged consumption model* in a similar way, replacing the lagged price by the lagged consumption. Consistent with our initial intuition, we find a positive lagged price effect and a negative lagged consumption effect in both clusters.

To account for lagged effects in our algorithm, we modify the parameter estimation step of our policy. At the end of price experimentation ($t = M$), we first perform parallel partial regression with lasso regularization to estimate the current price effect and feature coefficients using the same set of hyper-parameters as in §6.3.4. We then run a regression of the residuals over the lagged variable to estimate the lagged effect. In the subsequent periods ($t = M+1, \ldots, T$), we implement feature-based pricing based on the estimated parameters, and record the prices and corresponding consumptions. At the end of each time interval, we calculate the average prices and consumptions as the corresponding lagged quantities for the periods in the next time interval. The algorithm proceeds in this manner until the end of the time horizon.

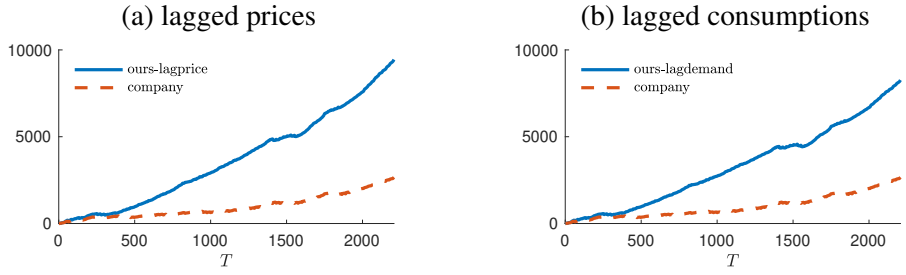


Figure 10 **Cumulative profits with lagged price and lagged consumption effect.**

Figure 10 shows the cumulative profits of our extended policy that accounts for the lagged prices (or lagged consumptions) and the company policy under the lagged price model (or the lagged consumption model). Our policy significantly outperforms the company's historical decisions, further demonstrating the practical value of our policy in the presence of lagged price/consumption effects.

8. Concluding Remarks

In this paper, we investigate a problem where an electric utility company dynamically adjusts its retail electricity prices to serve N customers over a discrete time horizon of T periods. Unique to the energy context, the utility company observes detailed information on features of various forms of heterogeneity pertaining to customers and exogenous factors. These features induce a structure of clusters (customer segments), and the customers' consumption behavior is determined by the features and prices distinctively in each cluster. The utility company knows neither the underlying cluster structure nor the consumption parameters in any cluster, and aims to learn these while earning profits. We design a data-driven policy using spectral clustering and feature-based pricing for this purpose. We prove that the average regret of our policy is at most in the order of $1/\sqrt{NT}$ when all features are fully heterogeneous over time and customers, and in the order of $1/\sqrt{N \wedge T}$ when some features may be invariant over time and customers. Moreover, we conduct multiple case studies, based on simulation experiments as well as a real-life data set from the retail electricity sector, to demonstrate the practical value of our approach. Our results indicate that there is substantial value in jointly using clustering and feature-based pricing. Furthermore, there is also significant value in using active price experimentation and lasso regularization. On top of these, our computational analysis shows that our policy significantly outperforms the real-life decisions of the company in our data set, and is robust to various forms of model misspecification.

Implementing our policy clearly requires data collection by utilities. The good news is that smart meters and other innovative technologies provide real-time (or near real-time) data. These and various publicly available data greatly facilitate the collection of feature data for utilities. For example, in our context, time-heterogenous features consist of weather-related information, such as temperature, humidity, pressure, etc., and they are all publicly available. Thus, utilities can easily incorporate this information into their policies. Moreover, several user-heterogeneous features such as a building's square footage, year of construction, etc. are also publicly available. Hence, utilities can easily access such information as well. In fact, the real-life data set we used in our case study demonstrates that in practice, utilities can access highly granular information about all these types of features. This further emphasizes the feasibility of collecting real-time (or near real-time) feature information for utilities.

It is worth comparing our online learning approach with offline learning. If a utility has access to historical data, then it can try to learn the cluster structure and the consumption parameters in an offline fashion. However, this is often challenging because historical data typically lacks granularity and price variation. In fact, this challenge is faced by the majority of U.S. utilities (see footnote 1 for details). On the other hand, if the utility does not have any historical data, then it needs to accumulate smart meter data with time-varying prices and dynamically learn the cluster structure and the consumption parameters. Without accurately estimating these in an online manner, the utility would incur inflated or even non-vanishing average regret. This further emphasizes the necessity of a carefully designed online learning policy.

References

Allcott, H. (2009). Real Time Pricing and Electricity Markets. Working paper.

Aouad, A., Elmachtoub, A. N., Ferreira, K. J., and McNellis, R. (2019). Market segmentation trees. *arXiv preprint arXiv:1906.01174*.

Arthur, D. and Vassilvitskii, S. (2006). k-means++: The Advantages of Careful Seeding. Technical report, Stanford University.

Atkinson, K. E. (1967). The numerical solution of the eigenvalue problem for compact integral operators. *Transactions of the American Mathematical Society*, 129(3):458–465.

Austin Energy (2014). Austin Energy Annual Report. https://austinenergy.com/wcm/connect/29f18bbd-b1f0-4a88-8fd0-1679564c034f/2014AnnualPerformanceReport.pdf?MOD=AJPERES&CVID=mNOMjLt. Accessed on 6/25/19.

Austin Energy (2018). Austin Energy Annual Report. https://austinenergy.com/wcm/connect/fc5e5028-8309-49f0-aae3-67db46bff892/2018corporate-annual-report.pdf?MOD=AJPERES&CVID=mFlnbMl. Accessed on 7/11/20.

Ban, G.-Y. and Keskin, N. B. (2020). Personalized Dynamic Pricing with Machine Learning: High Dimensional Features and Heterogeneous Elasticity. *Management Science,* forthcoming.

Bernstein, F., Modaresi, S., and Sauré, D. (2019). A dynamic clustering approach to data-driven assortment personalization. *Management Science*, 65(5):2095–2115.

Bertsimas, D. and Mersereau, A. J. (2007). A Learning Approach for Interactive Marketing to a Customer Segment. *Operations Research*, 55(6):1120–1135.

Bertsimas, D. J., Mersereau, A. J., and Patel, N. R. (2003). Dynamic Classification of Online Customers. In *Proceedings of the 2003 SIAM International Conference on Data Mining*, pages 107–118.

Borenstein, S. (2005). The Long-Run Efficiency of Real-Time Electricity Pricing. *The Energy Journal*, 26(3).

Bühler, T. and Salamon, D. (2018). *Functional Analysis*. American Mathematical Society.

CEC (2018). Utility Energy Supply Plans, and Supply Contracts. www.energy.ca.gov/almanac/electricity_data/supply_forms.html, Accessed on 8/8/21.

Cesa-Bianchi, N., Gentile, C., and Zappella, G. (2013). A Gang of Bandits. In *Advances in Neural Information Processing Systems*, pages 737–745.

Cohen, M. C., Leung, N.-H. Z., Panchamgam, K., Perakis, G., and Smith, A. (2017). The impact of linear optimization on promotion planning. *Operations Research*, 65(2):446–468.

ComEd (2020). ComEd's Hourly Pricing Program. https://hourlypricing.comed.com/. Accessed on 7/9/20.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

De Clercq, G. (2018). Run Your Dishwasher When the Sun Shines: Dynamic Power Pricing Grows. *Reuters*. https://www.reuters.com/article/us-europe-electricity-prices-insight/run-your-dishwasher-when-the-sun-shines-dynamic-power-pricing-grows-idUSKBN1KN0L7. Accessed on 7/10/20.

Ding, C., Rusmevichientong, P., and Topaloglu, H. (2014). Balancing Revenues and Repair Costs under Partial Information about Product Reliability. *Production and Operations Management*, 23(11):1899–1918.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104.

Dutta, G. and Mitra, K. (2017). A Literature Review on Dynamic Pricing of Electricity. *Journal of the Operational Research Society*, 68(10):1131–1145.

ECOFYS (2018). Dynamic Electricity Prices. https://asset-ec.eu/wp-content/uploads/2018/10/Dynamic-electricity-prices.pdf. Accessed on 7/9/20.

EIA (2020). Annual Electric Power Industry Report, Form EIA-861 Detailed Data Files. https://www.eia.gov/electricity/data/eia861/. Advanced_Meters_2019. Accessed on 7/22/21.

Eligo Energy (2021). Choose your energy and save. https://www.eligoenergy.com. Accessed on 7/22/21.

Everitt, B., Landau, S., Leese, M., and Stahl, D. (2011). Cluster analysis.

Faruqui, A. and Bourbonnais, C. (2019). A Meta Analysis of Time-Varying Rates. https://brattlefiles.blob.core.windows.net/files/16560_a_meta_analysis_of_time-varying_rates.pdf. Accessed on 7/9/20.

Faruqui, A. and Sergici, S. (2010). Household Response to Dynamic Pricing of Electricity: A Survey of 15 Experiments. *Journal of Regulatory Economics*, 38(2):193–225.

Federal Energy Regulatory Commission (2019). 2019 Assessment of Demand Response and Advanced Metering. https://www.ferc.gov/legal/staff-reports/2019/DR-AM-Report2019.pdf. Accessed on 6/2/20.

Feng, Z., Dawande, M., Janakiraman, G., and Qi, A. (2020). An Asymptotically Tight Learning Algorithm for Mobile-Promotion Platforms. Working Paper, UT Dallas – Naveen Jindal School of Management.

FERC (2007). 18 CFR Part 35 - Wholesale Competition in Regions With Organized Electric Markets; Proposed Rule. www.gpo.gov/fdsys/pkg/FR-2007-07-02/pdf/E7-12550.pdf, Accessed on 8/8/21.

Gentile, C., Li, S., Kar, P., Karatzoglou, A., Zappella, G., and Etrue, E. (2017). On Context-dependent Clustering of Bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1253–1262. JMLR.

Gentile, C., Li, S., and Zappella, G. (2014). Online Clustering of Bandits. In *International Conference on Machine Learning*, pages 757–765.

Harrison, J. M. and Sunar, N. (2015). Investment Timing with Incomplete Information and Multiple Means of Learning. *Operations Research*, 63(2):442–457.

Hausman, E., Hornby, R., and Smith, A. (2008). Bilateral Contracting in Deregulated Electricity Markets - A Report to the American Public Power Association. Synapse Energy Economics Inc., pg. 3.

He, L., Ma, G., Qi, W., and Wang, X. (2019). Charging an Electric Vehicle-Sharing Fleet. *Manufacturing & Service Operations Management*. forthcoming.

Huang, H., Sunar, N., and Swaminathan, J. M. (2020). Do Noisy Customer Reviews Discourage Platform Sellers? Empirical Analysis of an Online Solar Marketplace. Working paper, UNC Kenan-Flagler Business School.

Javanmard, A. and Nazerzadeh, H. (2019). Dynamic Pricing in High-Dimensions. *JMLR*, 20(1):315–363.

Kallus, N. (2017). Recursive partitioning for personalization using observational data. In *International Conference on Machine Learning*, pages 1789–1798. PMLR.

Keener, R. W. (2010). *Theoretical statistics: Topics for a core course*. Springer Science & Business Media.

Keskin, N. B. and Birge, J. R. (2019). Dynamic Selling Mechanisms for Product Differentiation and Learning. *Operations Research*, 67(4):1069–1089.

Keskin, N. B., Li, Y., and Song, J.-S. J. (2021). Data-driven dynamic pricing and ordering with perishable inventory in a changing environment. *Management Science, forthcoming*.

Keskin, N. B. and Zeevi, A. (2014). Dynamic Pricing with an Unknown Demand Model: Asymptotically Optimal Semi-myopic Policies. *Operations Research*, 62(5):1142–1167.

Kim, M. J. (2020). Variance Regularization in Sequential Bayesian Optimization. *Mathematics of Operations Research*, forthcoming.

Lazaar, J. (2016). Electricity Regulation In the US: A Guide. www.raponline.org/wp-content/uploads/2016/07/rap-lazar-electricity-regulation-US-june-2016.pdf, pg. 40 and Figure 7.1 on pg. 41, Accessed on 7/8/21.

MacQueen, J. et al. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. Oakland, Calif., USA.

MCE (2018). MCE 2019 Integrated Resource Plan. www.mcecleanenergy.org/wp-content/uploads/2019/01/MCE-2019-Integrated-Resource-Plan_11-8-2018_V_12-21-18.pdf, pg. 34, Accessed on 8/17/21.

McKinsey & Company (2019). The new way to engage with energy customers: personalization at scale. https://www.mckinsey.com/industries/electric-power-and-natural-gas/our-insights/the-new-way-to-engage-with-energy-customers-personalization-at-scale. Accessed on 7/23/21.

Miao, S., Chen, X., Chao, X., Liu, J., and Zhang, Y. (2019). Context-based Dynamic Pricing with Online Clustering. Working paper, University of Michigan.

Nambiar, M., Simchi-Levi, D., and Wang, H. (2019). Dynamic Learning and Pricing with Model Misspecification. *Management Science*, 65(11):4980–5000.

Navigant Research (2017). Big Data Drives Demand Side Management Innovation. https://blog.aee.net/big-data-drives-demand-side-management-innovation. Accessed on 6/2/20.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On Spectral Clustering: Analysis and an Algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856.

Nguyen, T. T. and Lauw, H. W. (2014). Dynamic Clustering of Contextual Multi-armed Bandits. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1959–1962.

OECD (2018). Personalised Pricing in the Digital Era – Note by the United States. www.ftc.gov/system/files/attachments/us-submissions-oecd-2010-present-other-international-competition-fora/personalized_pricing_note_by_the_united_states.pdf, Accessed on 8/24/21.

OpenEI (2019). U.S. Utility Rate Database: Austin Energy. https://openei.org/apps/USURDB/rate/view/539fcac1ec4f024d2f53fef2#3__Energy. Accessed on 6/25/19.

PCE (2017). PCE 2018 Integrated Resource Plan. www.peninsulacleanenergy.com/wp-content/uploads/2018/01/PCE-FINAL-2017-IRP-Updated.pdf, pg. 19 and 20, Accessed on 8/17/21.

Pecan Street (2019). Dataport – Pecan Street Inc. https://www.pecanstreet.org/dataport/. Accessed on 6/25/19.

Qi, A., Ahn, H.-S., and Sinha, A. (2017). Capacity Investment with Demand Learning. *Operations Research*, 65(1):145–164.

Qi, W. and Shen, Z.-J. M. (2019). A Smart-City Scope of Operations Management. *POM*, 28(2):393–406.

Rajabi, A., Eskandari, M., Ghadi, M. J., Li, L., Zhang, J., and Siano, P. (2020). A Comparative Study of Clustering Techniques for Electrical Load Pattern Segmentation. *Renewable and Sustainable Energy Reviews*, 120:109628.

Reliant Energy (2021). Pick your free electricity plan. https://www.reliant.com/en/public/reliant-pick-your-free.jsp. Accessed on 7/22/21.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., and Lin, C.-T. (2017). A Review of Clustering Techniques and Developments. *Neurocomputing*, 267:664–681.

Sunar, N. and Birge, J. R. (2019). Strategic Commitment to a Production Schedule with Uncertain Supply and Demand: Renewable Energy in Day-Ahead Electricity Markets. *Management Science*, 65(2):714–734.

Sunar, N. and Swaminathan, J. M. (2018). Net-Metered Distributed Renewable Energy: A Peril for Utilities? *Management Science*, forthcoming.

Sunar, N., Yu, S., and Kulkarni, V. (2020). Competitive Investment with Bayesian Learning: Choice of Business Size and Timing. *Operations Research*, forthcoming.

Energy Ogre (2021). Custom Pricing Now Available to Energy Ogre Members. www.energyogre.com/energy-ogre-custom-pricing, Accessed on 8/24/2021.

European Commission (2021). Smart Grids and Meters. https://ec.europa.eu/energy/topics/markets-and-consumers/smart-grids-and-meters_en. Accessed on 7/22/21.

Green Mountain Energy (2021). www.greenmountainenergy.com/business-electricity/get-a-quote/, Accessed on 8/24/2021.

The Edison Foundation (2021). Electric Company Smart Meter Deployments: Foundation for a Smart Grid (2021 Update). https://www.edisonfoundation.net/-/media/Files/IEI/publications/IEI_Smart_Meter_Report_April_2021.ashx, Accessed on 8/20/21.

The E.U. (2019). Directive (EU) 2019/944 of the European Parliament and the Council – Common Rules for the Internal Market for Electricity. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0944. Accessed on 7/9/20.

U.S. EIA (2021a). Annual Electric Power Industry Report, Form EIA-861 detailed data files. www.eia.gov/electricity/data/eia861/, Accessed on 8/24/21.

U.S. EIA (2021b). Electric Sales, Revenue, and Average Price. www.eia.gov/electricity/sales_revenue_price/, Table 1 for 2017, Accessed on 8/20/21.

Xoom Energy (2021). xoomenergy.com/en/big-business-program-a, Accessed on 8/24/2021.

Tropp, J. A. (2011). User-friendly Tail Bounds for Matrix Martingales. Technical report, Caltech.

von Luxburg, U. (2007). A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4):395–416.

von Luxburg, U., Belkin, M., and Bousquet, O. (2008). Consistency of Spectral Clustering. *The Annals of Statistics*, 36(2):555–586.

Wang, Y., Chen, Q., Kang, C., Zhang, M., Wang, K., and Zhao, Y. (2015). Load Profiling and Its Application to Demand Response: A Review. *Tsinghua Science and Technology*, 20(2):117–129.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *JASA*, 58(301):236–244.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, 61(3):439–447.

Zhang, H. (2018). Exact and Approximate Solutions for Dynamic Decision-Making with a Continuous Unknown State. Working Paper, University of British Columbia – Sauder School of Business.

Online Appendix for "Data-driven Clustering and Feature-based Retail Electricity Pricing with Smart Meters"

Appendix A: Illustrating Example of Our Policy

To demonstrate how joint clustering and feature-based pricing policy works, consider the illustrative example in Figure 1. In addition, the consumption parameters $\{\boldsymbol{\theta}_k^\intercal = [\gamma_k \ \beta_k \ \boldsymbol{\alpha}_k^\intercal] : k = 1, \ldots, \mathcal{K}\}$ are as follows:

$$\boldsymbol{\gamma} = \begin{bmatrix} 2.0 \\ 1.2 \\ 1.5 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} -0.2 \\ -0.8 \\ -0.5 \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} 0.6 & -0.4 \\ -0.1 & 0.2 \\ -0.2 & 0.1 \end{bmatrix}.$$

The consumption shocks follow the standard Gaussian distribution, and the expected consumption function is linear, that is, $g(\cdot)$ is chosen to be the identity function in this example. To implement our policy, we first generate a set of $n = 300$ test prices uniformly in the interval $[2, 4]$ and collect the consumption observations under the test prices and the 300 feature vectors illustrated in Figure 1a. As shown in Figure 1b, we estimate the underlying cluster structure using spectral clustering with the Gaussian kernel. We then estimate the consumption parameters for each estimated cluster k. Because the consumption model is linear and all features are fully heterogeneous with dimension 2, parameter estimation reduces to linear regressions. Upon observing new features, we perform spectral clustering using the 300 features together with the new feature, and record the cluster label $\tilde{k}$ for the new feature. Figure 11 depicts 100 new random observations and the corresponding cluster prediction: spectral clustering achieves 100% prediction accuracy in this example, up to the same permutation of cluster labels as in the cluster estimation shown in Figure 1b. Finally, as the consumption model is linear, we set the optimal price based on a new feature observation $\tilde{\mathbf{x}}$ as

$$\varphi(\hat{\boldsymbol{\theta}}_{\tilde{k}}, \tilde{\mathbf{x}}) = -\frac{\hat{\gamma}_{\tilde{k}} + \hat{\boldsymbol{\alpha}}_{\tilde{k}}^\intercal \tilde{\mathbf{x}}}{2\hat{\beta}_{\tilde{k}}},$$

if $\tilde{\mathbf{x}}$ is predicted to reside in cluster $\tilde{k}$ and the corresponding parameter estimate is $\hat{\boldsymbol{\theta}}_{\tilde{k}}^\intercal = [\hat{\gamma}_{\tilde{k}} \ \hat{\beta}_{\tilde{k}} \ \hat{\boldsymbol{\alpha}}_{\tilde{k}}^\intercal]$.

(a) new sample observations (b) cluster prediction

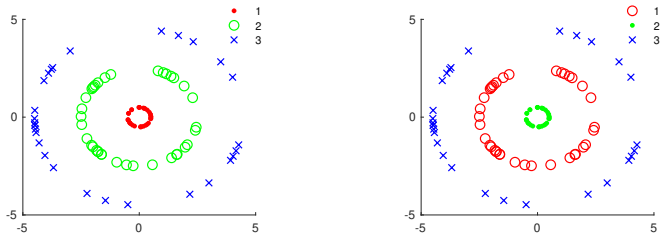


Figure 11 **New sample observations and the cluster prediction**

Appendix B: Proof of Proposition 1.

Given a random sample $\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_n$ of size n generated independently from the distribution $\mathbb{P}_{\mathcal{Z}}$, we define the empirical measure as $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\tilde{\mathbf{Z}}_i}$, where $\delta_{\tilde{\mathbf{Z}}_\ell}$ is the Dirac measure at $\tilde{\mathbf{Z}}_i$. The empirical degree function, $d_n(\cdot)$, and the true degree function, $d(\cdot)$, are then defined as

$$d_n(\mathbf{x}) = \int \psi(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbb{P}_n(\mathbf{y}), \tag{B.1}$$

$$d(\mathbf{x}) = \int \psi(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{y}), \tag{B.2}$$

for $\mathbf{x} \in \mathcal{Z}$. Note that $d_n(\mathbf{x})$ is simply the row sum of the degree matrix corresponding to $\mathbf{x}$, normalized by the number of observations, and $d(\mathbf{x})$ can be viewed as its "limit" as $n \to \infty$. Let $h_n(\mathbf{x},\mathbf{y}) = \psi(\mathbf{x},\mathbf{y})/\sqrt{d_n(\mathbf{x})d_n(\mathbf{y})}$ and $h(\mathbf{x},\mathbf{y}) = \psi(\mathbf{x},\mathbf{y})/\sqrt{d(\mathbf{x})d(\mathbf{y})}$ be the normalized similarity functions. Based on this, we define the following integral operators $\{J_n : n \in \mathbb{N}\}$ and J from the Banach space of continuous functions on $\mathcal{Z}$ equipped with the sup-norm, $(C(\mathcal{Z}), \|\cdot\|_\infty)$, onto itself, given by

$$J_n f(\mathbf{x}) = \int h_n(\mathbf{x},\mathbf{y}) f(\mathbf{y})\,\mathrm{d}\mathbb{P}_n(\mathbf{y}), \tag{B.3}$$

$$Jf(\mathbf{x}) = \int h(\mathbf{x},\mathbf{y}) f(\mathbf{y})\,\mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{y})\,. \tag{B.4}$$

We first show that J is a compact operator. Since $\mathcal{Z}$ is a compact set, for any $f \in C(\mathcal{Z})$, we have

$$\int |f(\mathbf{x})|^2\,\mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{x}) \le \|f\|_\infty^2 < \infty, \tag{B.5}$$

where $\|f\|_\infty$ is the supremum norm of f on $\mathcal{Z}$. Hence, f is in $L^2(\mathcal{Z})$, the space of square-integrable functions on $\mathcal{Z}$. (Strictly speaking, the elements of $L^2(\mathcal{Z})$ are equivalence classes of functions. Viewing f as the representative of its equivalence class, we write $f \in L^2(\mathcal{Z})$ for notational simplicity.) Let $\{f_m\}_{m\in\mathbb{N}}$ be an orthonormal basis of $L^2(\mathcal{Z})$. Such an orthonormal basis always exists because $L^2(\mathcal{Z})$ is a Hilbert space. Then, $\{f_{m,\ell}\}_{m,\ell\in\mathbb{N}}$ is an orthonormal basis of $L^2(\mathcal{Z}\times\mathcal{Z})$, where $f_{m,\ell}(\mathbf{x},\mathbf{y}) = f_m(\mathbf{x})\overline{f_\ell(\mathbf{y})}$ for $\mathbf{x},\mathbf{y} \in \mathcal{Z}$, and $\overline{f_\ell(\mathbf{y})}$ denotes the complex conjugate of $f_\ell(\mathbf{y})$. Since h is continuous on $\mathcal{Z}\times\mathcal{Z}$, and $\mathcal{Z}$ is compact, we have $h \in L^2(\mathcal{Z}\times\mathcal{Z})$. Thus, we can write $h(\mathbf{x},\mathbf{y}) = \sum_{m=1}^{\infty}\sum_{\ell=1}^{\infty} \phi_{m,\ell} f_m(\mathbf{x})\overline{f_\ell(\mathbf{y})}$ for $\mathbf{x},\mathbf{y} \in \mathcal{Z}$, where

$$\phi_{m,\ell} = \langle f_{m,\ell}, h\rangle = \iint \overline{f_m(\mathbf{x})} f_\ell(\mathbf{y}) h(\mathbf{x},\mathbf{y})\,\mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{x})\mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{y}). \tag{B.6}$$

Consequently, J is a bounded operator because

$$\begin{aligned}
\|Jf\|_2^2 &= \int \left(\int h(\mathbf{x},\mathbf{y}) f(\mathbf{y})\,\mathrm{d}\mathbb{P}_{\mathcal{Z}}(y)\right)^2 \mathrm{d}\mathbb{P}_{\mathcal{Z}}(x) \\
&\le \int \left(\int |h(\mathbf{x},\mathbf{y})|^2\,\mathrm{d}\mathbb{P}_{\mathcal{Z}}(y)\right)\cdot\left(\int |f(\mathbf{y})|^2\,\mathrm{d}\mathbb{P}_{\mathcal{Z}}(y)\right)\mathrm{d}\mathbb{P}_{\mathcal{Z}}(x) \\
&= \|h\|_2^2 \cdot \|f\|_2^2\,.
\end{aligned} \tag{B.7}$$

Define $h_{\mathfrak{N}} : \mathcal{Z}\times\mathcal{Z} \to \mathbb{C}$ as

$$h_{\mathfrak{N}}(\mathbf{x},\mathbf{y}) = \sum_{m=1}^{\mathfrak{N}}\sum_{\ell=1}^{\infty} \phi_{m,\ell} f_m(\mathbf{x})\overline{f_\ell(\mathbf{y})}, \tag{B.8}$$

and $J_{\mathfrak{N}} : C(\mathcal{Z}) \to C(\mathcal{Z})$ as

$$J_{\mathfrak{N}} f(\mathbf{x}) = \int h_{\mathfrak{N}}(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) \, \mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{y}). \tag{B.9}$$

Note that $h_{\mathfrak{N}} \in L^2(\mathcal{Z} \times \mathcal{Z})$ because

$$\begin{aligned} \|h_{\mathfrak{N}}\|_2^2 &= \left\langle \sum_{m=1}^{\mathfrak{N}} \sum_{\ell=1}^{\infty} \phi_{m,\ell} f_{m,\ell} \, , \sum_{m'=1}^{\mathfrak{N}} \sum_{\ell'=1}^{\infty} \phi_{m',\ell'} f_{m',\ell'} \right\rangle \\ &= \sum_{m=1}^{\mathfrak{N}} \sum_{\ell=1}^{\infty} \overline{\phi_{m,\ell}} \left\langle f_{m,\ell} \, , \sum_{m'=1}^{\mathfrak{N}} \sum_{\ell'=1}^{\infty} \phi_{m',\ell'} f_{m',\ell'} \right\rangle \\ &= \sum_{m=1}^{\mathfrak{N}} \sum_{\ell=1}^{\infty} |\phi_{m,\ell}|^2 \leq \sum_{m=1}^{\infty} \sum_{\ell=1}^{\infty} |\phi_{m,\ell}|^2 = \|h\|_2^2 < \infty. \end{aligned} \tag{B.10}$$

Furthermore, writing $f = \sum_{m=1}^{\infty} \alpha_m f_m$ with $\alpha_m = \langle f_m, f \rangle$, we have

$$\begin{aligned} J_{\mathfrak{N}} f(\mathbf{x}) &= \int h_{\mathfrak{N}}(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) \, \mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{y}) \\ &= \int \sum_{m=1}^{\mathfrak{N}} \sum_{\ell=1}^{\infty} \phi_{m,\ell} f_m(\mathbf{x}) \overline{f_\ell(\mathbf{y})} \sum_{m'=1}^{\infty} \alpha_{m'} f_{m'}(\mathbf{y}) \, \mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{y}) \\ &= \sum_{m=1}^{\mathfrak{N}} \sum_{\ell=1}^{\infty} \phi_{m,\ell} f_m(\mathbf{x}) \int \overline{f_\ell(\mathbf{y})} \sum_{m'=1}^{\infty} \alpha_{m'} f_{m'}(\mathbf{y}) \, \mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{y}) \\ &= \sum_{m=1}^{\mathfrak{N}} \left(\sum_{\ell=1}^{\infty} \phi_{m,\ell} \alpha_\ell \right) f_m(\mathbf{x}), \end{aligned} \tag{B.11}$$

where the interchange of integration and the limit follows from the dominated convergence theorem and the Cauchy-Schwarz inequality because

$$\int |\overline{f_\ell(\mathbf{y})} f(\mathbf{y})| \, \mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{y}) \leq \left(\int |f_\ell(\mathbf{y})|^2 \, \mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{y}) \right)^{1/2} \cdot \left(\int |f(\mathbf{y})|^2 \, \mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{y}) \right)^{1/2} = \|f_\ell\|_2 \cdot \|f\|_2 < \infty \,. \tag{B.12}$$

The infinite sum $\sum_{\ell=1}^{\infty} \phi_{m,\ell} \alpha_\ell$ is finite for all m by the Cauchy-Schwarz inequality for series because

$$\left| \sum_{\ell=1}^{\infty} \phi_{m,\ell} \alpha_\ell \right| \leq \left(\sum_{\ell=1}^{\infty} |\phi_{m,\ell}|^2 \right)^{1/2} \cdot \left(\sum_{\ell=1}^{\infty} |\alpha_\ell|^2 \right)^{1/2} \leq \|h\|_2 \cdot \|f\|_2 < \infty \,. \tag{B.13}$$

Additionally, by applying the argument that shows the boundedness of J, we can show that $J_{\mathfrak{N}}$ is also a bounded operator. Since $J_{\mathfrak{N}} f$ is a finite linear combination of the first $\mathfrak{N}$ elements of the orthonormal basis $\{f_m\}_{m \in \mathbb{N}}$, it follows that $J_{\mathfrak{N}}$ is of finite rank and thus compact. Finally, by the Cauchy-Schwarz inequality, we deduce that for any $f \in C(\mathcal{Z})$,

$$\begin{aligned} \|(J - J_{\mathfrak{N}}) f\|_2^2 &= \int \left(\int (h(\mathbf{x}, \mathbf{y}) - h_{\mathfrak{N}}(\mathbf{x}, \mathbf{y})) f(\mathbf{y}) \, \mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{y}) \right)^2 \mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{x}) \\ &\leq \int \left(\int |h(\mathbf{x}, \mathbf{y}) - h_{\mathfrak{N}}(\mathbf{x}, \mathbf{y})|^2 \, \mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{y}) \right) \mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{x}) \cdot \left(\int |f(\mathbf{y})|^2 \, \mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{y}) \right) \\ &= \iint \left| \sum_{m=\mathfrak{N}+1}^{\infty} \sum_{\ell=1}^{\infty} \phi_{m,\ell} f_{m,\ell}(\mathbf{x}, \mathbf{y}) \right|^2 \mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{x}) \, \mathrm{d}\mathbb{P}_{\mathcal{Z}}(\mathbf{y}) \cdot \|f\|_2^2 \end{aligned}$$

$$= \left\langle \sum_{m=\mathfrak{N}+1}^{\infty}\sum_{\ell=1}^{\infty}\phi_{m,\ell}f_{m,\ell}\,,\ \sum_{m'=\mathfrak{N}+1}^{\infty}\sum_{\ell'=1}^{\infty}\phi_{m',\ell'}f_{m',\ell'}\right\rangle \cdot \|f\|_2^2$$
$$= \sum_{m=\mathfrak{N}+1}^{\infty}\sum_{\ell=1}^{\infty}|\phi_{m,\ell}|^2\cdot\|f\|_2^2, \tag{B.14}$$

where the equivalence of the iterated integral and the double integral follows from Fubini's theorem because $\iint|h|^2=\|h\|_2^2<\infty$, which in turn implies that the tail sum must converge to 0 as $\mathfrak{N}\to\infty$, and consequently,

$$\|J-J_{\mathfrak{N}}\|=\sup_{f\in C(\mathcal{Z})\setminus\{0\}}\frac{\|(J-J_{\mathfrak{N}})f\|_2}{\|f\|_2}\le\left(\sum_{m=\mathfrak{N}+1}^{\infty}\sum_{\ell=1}^{\infty}|\phi_{m,\ell}|^2\right)^{1/2}\to 0 \tag{B.15}$$

as $\mathfrak{N}\to\infty$. Thus, by Theorem 4.2.10(ii) of Bühler and Salamon (2018), J is a compact operator as the limit of a sequence of compact operators in the norm topology. By Theorem 5.2.8(ii) of Bühler and Salamon (2018), all non-zero eigenvalues of J are isolated.

Let $V_n=I-J_n$ and $V=I-J$, where I is the identity operator. According to Proposition 9 of von Luxburg et al. (2008), ρ is an eigenvalue of V_n with eigenfunction f if and only if ρ is an eigenvalue of $\mathcal{L}_n$ with eigenvector v_n obtained by evaluating f at the observed feature vectors. Since $\psi(\mathbf{x},\mathbf{y})$ is a positive definite kernel and so is $h(\mathbf{x},\mathbf{y})$, the eigenvalues of J are all positive and simple, and thus isolated. Hence, all the eigenvalues of V are isolated and different from 1, because ρ is an eigenvalue of V if and only if $1-\rho$ is an eigenvalue of J. Denote by $\rho^{(k)}$ the k^{th} eigenvalue of V in the ascending order. By Theorem 15 of von Luxburg et al. (2008), the k^{th} eigenfunction $f_n^{(k)}$ of V_n converges to the k^{th} eigenfunction $f^{(k)}$ of V up to a change of sign. More precisely, there exists a sequence $a_n^{(k)}\in\{+1,-1\}$ such that $\|a_n^{(k)}f_n^{(k)}-f^{(k)}\|_\infty\to 0$ almost surely. Moreover, by Theorems 16 and 19 of von Luxburg et al. (2008), there exist positive constants c_1 and c_2 such that

$$\mathbb{P}_{\mathbf{Z}}\left\{\|a_n^{(k)}f_n^{(k)}-f^{(k)}\|_\infty\le c_1\sqrt{\frac{\log n}{n}}\right\}\ge 1-\frac{c_2}{n}. \tag{B.16}$$

Note that if $\|a_n^{(k)}f_n^{(k)}-f^{(k)}\|_\infty\le c_1\sqrt{\log n/n}$, then for any $b\in\mathbb{R}$ there exists a constant $K_1>0$ such that

$$\frac{1}{n}\left|\{i:(a_n^{(k)}f_n^{(k)}(\tilde{\mathbf{Z}}_i)-b)(f^{(k)}(\tilde{\mathbf{Z}}_i)-b)<0\}\right|\le K_1\sqrt{\frac{\log n}{n}}, \tag{B.17}$$

where $|A|$ denotes the cardinality of A for any given set A. If there are at most $\bar{\mathcal{K}}$ true clusters and the spectral clustering algorithm makes use of the first $\bar{\mathcal{K}}$ eigenfunctions of V_n to estimate the underlying cluster structure by thresholding at $b\in\mathbb{R}$, then

$$\mathbb{P}_{\mathbf{Z}}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{\mathcal{C}(\tilde{\mathbf{Z}}_i)\neq\hat{\mathcal{C}}(\tilde{\mathbf{Z}}_i)\}\ge K_1\bar{\mathcal{K}}\sqrt{\frac{\log n}{n}}\right\}$$
$$\le\mathbb{P}_{\mathbf{Z}}\left\{\bigcup_{k=1}^{\bar{\mathcal{K}}}\left\{\frac{1}{n}\left|\{i:(a_n^{(k)}f_n^{(k)}(\tilde{\mathbf{Z}}_i)-b)(f^{(k)}(\tilde{\mathbf{Z}}_i)-b)<0\}\right|\ge K_1\sqrt{\frac{\log n}{n}}\right\}\right\}$$
$$\le\sum_{k=1}^{\bar{\mathcal{K}}}\mathbb{P}_{\mathbf{Z}}\left\{\frac{1}{n}\left|\{i:(a_n^{(k)}f_n^{(k)}(\tilde{\mathbf{Z}}_i)-b)(f^{(k)}(\tilde{\mathbf{Z}}_i)-b)<0\}\right|\ge K_1\sqrt{\frac{\log n}{n}}\right\}$$

$$\leq \sum_{k=1}^{\bar{\mathcal{K}}} \mathbb{P}_{\mathbf{Z}} \left\{ \|a_n^{(k)} f_n^{(k)} - f^{(k)}\|_\infty \geq K_1 \sqrt{\frac{\log n}{n}} \right\} \leq \frac{K_2 \bar{\mathcal{K}}}{n}. \tag{B.18}$$

This completes the proof. Q.E.D.

Appendix C: Proofs of Results in §5.1

Proof of Proposition 2. We let $m = (t-1)N + i$ for $i \in \mathcal{N}$ and $t \in \mathcal{T}$, and based on this, we write $D_m = D_{i,t}$, $\mathbf{X}_m = \mathbf{X}_{i,t}$, $p_m = p_{i,t}$, and $\mathbf{U}_m^\intercal = [1 \; p_m \; \mathbf{X}_m^\intercal]$. Because all features are fully heterogeneous, it suffices to choose a single lasso-regularization penalty λ_k for each $k \in \{1, \ldots, \bar{\mathcal{K}}\}$. Given $K_3 > 0$, let $\mathcal{A}_k = \{\|\hat{\boldsymbol{\vartheta}}_k(\lambda_k) - \boldsymbol{\theta}_k\|_2^2 \geq K_3 \bar{s} \log(dNM)/(NM)\}$. We assume that $K_3 \bar{s} \log(dNM)/(NM) \leq 1$ because otherwise we can simply take $K_4 \geq K_3$ to complete the proof trivially.

Since the quasi-likelihood (3) is strictly concave with respect to $\tilde{\boldsymbol{\theta}}$ given λ_k and $\mathbf{U} = \{\mathbf{U}_m : m = 1, \ldots, NM\}$, there exists $\tilde{\boldsymbol{\theta}}_k \in \Theta$ such that $\|\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\|_2^2 = K_3 \bar{s} \log(dNM)/(NM)$ and $Q_k(\boldsymbol{\theta}_k; \lambda_k, \mathbf{U}) < Q_k(\tilde{\boldsymbol{\theta}}_k; \lambda_k, \mathbf{U})$. For notational simplicity, we suppress the dependency of $Q_k(\boldsymbol{\theta}_k; \lambda_k, \mathbf{U})$ on $\mathbf{U}$ in the rest of the proof. We thus have $\mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}(\mathcal{A}_k) \leq \mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\{Q_k(\boldsymbol{\theta}_k; \lambda_k) < Q_k(\tilde{\boldsymbol{\theta}}_k; \lambda_k)\}$. Let $\hat{\chi}_m(k) = \mathbb{I}\{\hat{\mathcal{C}}(\mathbf{X}_m) = k\}$ and $\chi_m(k) = \mathbb{I}\{\mathcal{C}(\mathbf{X}_m) = k\}$, and define $\tilde{Q}_k(\tilde{\boldsymbol{\theta}}) = Q_k(\tilde{\boldsymbol{\theta}}; \tilde{\lambda}) + \tilde{\lambda}\|\tilde{\boldsymbol{\theta}}\|_1$ for generic values of $\tilde{\boldsymbol{\theta}}$ and $\tilde{\lambda}$. Note that

$$\begin{aligned}
\nabla \tilde{Q}_k(\tilde{\boldsymbol{\theta}}) &= \sum_{m=1}^{NM} \hat{\chi}_m(k) g'(\tilde{\boldsymbol{\theta}}^\intercal \mathbf{U}_m) \frac{D_m - g(\tilde{\boldsymbol{\theta}}^\intercal \mathbf{U}_m)}{\nu(g(\tilde{\boldsymbol{\theta}}^\intercal \mathbf{U}_m))} \mathbf{U}_m \\
&= \sum_{m=1}^{NM} \hat{\chi}_m(k) [D_m - g(\tilde{\boldsymbol{\theta}}^\intercal \mathbf{U}_m)] \mathbf{U}_m, \qquad \text{(C.1)} \\
\nabla^2 \tilde{Q}_k(\tilde{\boldsymbol{\theta}}) &= -\sum_{m=1}^{NM} \hat{\chi}_m(k) g'(\tilde{\boldsymbol{\theta}}^\intercal \mathbf{U}_m) \mathbf{U}_m \mathbf{U}_m^\intercal, \qquad \text{(C.2)}
\end{aligned}$$

since $\nu = g' \circ g^{-1}$. Based on the second-order Taylor expansion of $\tilde{Q}_k(\cdot)$ at $\boldsymbol{\theta}_k$, we have

$$\begin{aligned}
\tilde{Q}_k(\tilde{\boldsymbol{\theta}}_k) &= \tilde{Q}_k(\boldsymbol{\theta}_k) + (\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)^\intercal \nabla \tilde{Q}_k(\boldsymbol{\theta}_k) + \frac{1}{2}(\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)^\intercal \nabla^2 \tilde{Q}_k(\bar{\boldsymbol{\theta}}_k)(\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k) \\
&= \tilde{Q}_k(\boldsymbol{\theta}_k) + (\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)^\intercal \sum_{m=1}^{NM} \hat{\chi}_m(k) [D_m - g(\boldsymbol{\theta}_k^\intercal \mathbf{U}_m)] \mathbf{U}_m \\
&\quad - \frac{1}{2}(\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)^\intercal \sum_{m=1}^{NM} \hat{\chi}_m(k) g'(\bar{\boldsymbol{\theta}}_k^\intercal \mathbf{U}_m) \mathbf{U}_m \mathbf{U}_m^\intercal (\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k), \qquad \text{(C.3)}
\end{aligned}$$

where $\bar{\boldsymbol{\theta}}_k$ is on the line segment connecting $\boldsymbol{\theta}_k$ and $\tilde{\boldsymbol{\theta}}_k$. Note that $D_m = g(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_m)}^\intercal \mathbf{U}_m) + \varepsilon_m$ and $\mathcal{C}(\mathbf{X}_m) = k$ if and only if $\hat{\chi}_m(k) = \chi_m(k) = 1$. Let $\Lambda_k = \{m = 1, \ldots, NM : \hat{\chi}_m(k) = 1, \chi_m(k) = 0\}$. We thus have

$$\begin{aligned}
&Q_k(\tilde{\boldsymbol{\theta}}_k; \lambda_k) - Q_k(\boldsymbol{\theta}_k; \lambda_k) \\
&\quad = \tilde{Q}_k(\tilde{\boldsymbol{\theta}}_k) - \tilde{Q}_k(\boldsymbol{\theta}_k) + \lambda_k \|\boldsymbol{\theta}_k\|_1 - \lambda_k \|\tilde{\boldsymbol{\theta}}_k\|_1 \\
&\quad = (\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)^\intercal \sum_{m=1}^{NM} \hat{\chi}_m(k) \varepsilon_m \mathbf{U}_m + (\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)^\intercal \sum_{m \in \Lambda_k} [g(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_m)}^\intercal \mathbf{U}_m) - g(\boldsymbol{\theta}_k^\intercal \mathbf{U}_m)] \mathbf{U}_m \\
&\qquad - \frac{1}{2}(\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)^\intercal \sum_{m=1}^{NM} \hat{\chi}_m(k) g'(\bar{\boldsymbol{\theta}}_k^\intercal \mathbf{U}_m) \mathbf{U}_m \mathbf{U}_m^\intercal (\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k) + \lambda_k \|\boldsymbol{\theta}_k\|_1 - \lambda_k \|\tilde{\boldsymbol{\theta}}_k\|_1. \qquad \text{(C.4)}
\end{aligned}$$

Let $\mathcal{V}_k = \sum_{m=1}^{NM} \hat{\chi}_m(k)\mathbf{U}_m\mathbf{U}_m^\intercal$, $\mathcal{M}_k(\ell) = \sum_{m=1}^{\ell} \hat{\chi}_m(k)\varepsilon_m\mathbf{U}_m$, and $\boldsymbol{\zeta}_k = \sum_{m\in\Lambda_k}[g(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{x}_m)}^\intercal\mathbf{U}_m) - g(\boldsymbol{\theta}_k^\intercal\mathbf{U}_m)]\mathbf{U}_m$. Based on this, we deduce that

$$\mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\{\mathcal{A}_k\} \le \mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\bigg\{\frac{1}{2}g'_{\min}(\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)^\intercal\mathcal{V}_k(\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k) < (\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)^\intercal\mathcal{M}_k(NM) + (\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)^\intercal\boldsymbol{\zeta}_k + \lambda_k\|\boldsymbol{\theta}_k\|_1-\lambda_k\|\tilde{\boldsymbol{\theta}}_k\|_1\bigg\}, \tag{C.5}$$

where $g'_{\min} = \min\{g'(\boldsymbol{\vartheta}^\intercal\mathbf{u}) : \mathbf{u}^\intercal = [1\ p\ \mathbf{x}^\intercal], \boldsymbol{\vartheta}\in\Theta, \mathbf{x}\in\mathcal{X}, p\in\mathcal{P}\}$. To provide a lower bound on the term $(\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)^\intercal\mathcal{V}_k(\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)$ and upper bounds on $(\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)^\intercal\mathcal{M}_k(NM)$ and $(\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)^\intercal\boldsymbol{\zeta}_k$, we first state the following two lemmas and prove them at the end of this subsection.

LEMMA 1. *There exist positive constants C_1 and C_2 such that*

$$\mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\{\mu_{\min}(\mathcal{V}_k) \le C_1NM\} \le \frac{C_2\bar{s}\log(dNM)}{NM} \tag{C.6}$$

for all $k\in\{1,\dots,\bar{\mathcal{K}}\}$ and all $M\ge 2$, where $\mu_{\min}(\mathcal{V}_k)$ is the minimum eigenvalue of $\mathcal{V}_k$.

LEMMA 2. *There exists a positive constant C_3 such that*

$$\mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\{\|\mathcal{M}_k(NM)\|_\infty > \lambda_k\} \le \frac{C_3\bar{s}\log(dNM)}{NM} \tag{C.7}$$

for all $k\in\{1,\dots,\bar{\mathcal{K}}\}$ and all $M\ge 2$.

Using Lemmas 1 and 2, the Rayleigh-Ritz theorem, and Hölder's inequality, we deduce that

$$\begin{aligned}
&\mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\{\mathcal{A}_k\}\\
&\le \mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\bigg\{\frac{1}{2}g'_{\min}\mu_{\min}(\mathcal{V}_k)\|\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k\|_2^2 < (\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)^\intercal\mathcal{M}_k(NM) + (\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)^\intercal\boldsymbol{\zeta}_k + \lambda_k\|\boldsymbol{\theta}_k\|_1-\lambda_k\|\tilde{\boldsymbol{\theta}}_k\|_1,\\
&\qquad \mu_{\min}(\mathcal{V}_k) > C_1NM\bigg\} + \mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\{\mu_{\min}(\mathcal{V}_k) \le C_1NM\}\\
&\le \mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\bigg\{\frac{1}{2}g'_{\min}C_1NM\|\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k\|_2^2 < (\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)^\intercal\mathcal{M}_k(NM) + (\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)^\intercal\boldsymbol{\zeta}_k + \lambda_k\|\boldsymbol{\theta}_k\|_1-\lambda_k\|\tilde{\boldsymbol{\theta}}_k\|_1\bigg\}\\
&\qquad + \frac{C_2\bar{s}\log(dNM)}{NM}\\
&\le \mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\bigg\{\frac{1}{2}g'_{\min}C_1NM\|\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k\|_2^2 < \|\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k\|_1\|\mathcal{M}_k(NM)\|_\infty + (\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)^\intercal\boldsymbol{\zeta}_k + \lambda_k\|\boldsymbol{\theta}_k\|_1-\lambda_k\|\tilde{\boldsymbol{\theta}}_k\|_1,\\
&\qquad \|\mathcal{M}_k(NM)\|_\infty \le \lambda_k\bigg\} + \mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\{\|\mathcal{M}_k(NM)\|_\infty > \lambda_k\} + \frac{C_2\bar{s}\log(dNM)}{NM}\\
&\le \mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\bigg\{\frac{1}{2}g'_{\min}C_1NM\|\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k\|_2^2 < \lambda_k\|\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k\|_1 + \lambda_k\|\boldsymbol{\theta}_k\|_1-\lambda_k\|\tilde{\boldsymbol{\theta}}_k\|_1 + (\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)^\intercal\boldsymbol{\zeta}_k\bigg\}\\
&\qquad + \frac{(C_2+C_3)\bar{s}\log(dNM)}{NM}.
\end{aligned} \tag{C.8}$$

Let $\tilde{\boldsymbol{\theta}}_k^{\mathcal{S}}$ and $\boldsymbol{\theta}_k^{\mathcal{S}}$ be the vectors consisting of components of $\tilde{\boldsymbol{\theta}}_k$ and $\boldsymbol{\theta}_k$ whose indices are in the support $\mathcal{S}$ of $\boldsymbol{\theta}_k$, respectively. Let $\tilde{\boldsymbol{\theta}}_k^{\mathcal{S}^c}$ be the vector consisting of components of $\tilde{\boldsymbol{\theta}}_k$ whose indices are not in $\mathcal{S}$. We then have

$$\|\tilde{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k\|_1+\|\boldsymbol{\theta}_k\|_1-\|\tilde{\boldsymbol{\theta}}_k\|_1 = \|\tilde{\boldsymbol{\theta}}_k^{\mathcal{S}}-\boldsymbol{\theta}_k^{\mathcal{S}}\|_1+\|\tilde{\boldsymbol{\theta}}_k^{\mathcal{S}^c}\|_1+\|\boldsymbol{\theta}_k^{\mathcal{S}}\|_1-\|\tilde{\boldsymbol{\theta}}_k^{\mathcal{S}}\|_1-\|\tilde{\boldsymbol{\theta}}_k^{\mathcal{S}^c}\|_1$$

$$\leq 2\|\tilde{\boldsymbol{\theta}}_k^{\mathcal{S}} - \boldsymbol{\theta}_k^{\mathcal{S}}\|_1 \leq 2\sqrt{\bar{s}}\|\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\|_2, \tag{C.9}$$

where the last two steps follow from Minkowski's inequality and the Cauchy-Schwarz inequality.

Finally, let $\mathcal{B}_k = \{\sum_{m=1}^{NM} \mathbb{I}\{\chi_m(k) \neq \hat{\chi}_m(k)\} \geq 2K_1\sqrt{NM\log(NM)}\}$. On the event $\mathcal{B}_k^c$, we have by the Cauchy-Schwarz inequality that

$$|(\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)^\intercal \boldsymbol{\zeta}_k| \leq \|\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\|_2 \|\boldsymbol{\zeta}_k\|_2 \leq 4\bar{\nu}K_1\sqrt{NM\log(NM)}\|\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\|_2, \tag{C.10}$$

where $\bar{\nu} = \max\{|g(\boldsymbol{\vartheta}^\intercal \mathbf{u})| \cdot \|\mathbf{u}\|_2 \colon \mathbf{u}^\intercal = [1\ p\ \mathbf{x}^\intercal], \boldsymbol{\vartheta} \in \Theta, \mathbf{x} \in \mathcal{X}, p \in \mathcal{P}\}$. It follows by Proposition 1 that

$$\begin{aligned}
\mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\{\mathcal{A}_k\} &\leq \mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\left\{\frac{1}{2}g'_{\min}C_1 NM\|\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\|_2^2 < 2(\lambda_k\sqrt{\bar{s}} + 2\bar{\nu}K_1\sqrt{NM\log(NM)})\|\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\|_2,\ \mathcal{B}_k^c\right\} \\
&\quad + \mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\{\mathcal{B}_k\} + \frac{(C_2 + C_3)\bar{s}\log(dNM)}{NM} \\
&\leq \mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\|_2 < \xi_1\sqrt{\frac{\bar{s}\log(dNM)}{NM}}\right\} + \frac{(2K_2 + C_2 + C_3)\bar{s}\log(dNM)}{NM},
\end{aligned} \tag{C.11}$$

where $\xi_1 = 4(\tilde{c}_k + 2\bar{\nu}K_1)/(g'_{\min}C_1)$. Note that if $K_3 \geq \xi_1^2$, then the first term on the right-hand side vanishes because $\|\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\|_2^2 = K_3\bar{s}\log(dNM)/(NM)$. Note that $\|\hat{\boldsymbol{\theta}}_k(\lambda_k) - \boldsymbol{\theta}_k\|_2 \leq \|\hat{\boldsymbol{\vartheta}}_k(\lambda_k) - \boldsymbol{\theta}_k\|_2$, where $\hat{\boldsymbol{\theta}}_k(\lambda_k)$ is the projection of $\hat{\boldsymbol{\vartheta}}_k(\lambda_k)$ onto Θ. We thus complete the proof by taking $K_3 = \xi_1^2$ and $K_4 = 2K_2 + C_2 + C_3$. Q.E.D.

Proof of Theorem 1. For any $t = M+1, \ldots, T$, let $\hat{N}(t)$ be the total number of misclassified features in the first t periods, i.e., $\hat{N}(t) = \sum_{s=1}^t \sum_{i=1}^N \mathbb{I}\{\mathcal{C}(\mathbf{X}_{i,s}) \neq \hat{\mathcal{C}}(\mathbf{X}_{i,s})\}$. Hence, by Proposition 1, we have that

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}[\hat{N}(t)] &= \mathbb{E}_{\mathbf{X}}\left\{\hat{N}(t) \,\middle|\, \hat{N}(t) \geq K_1\bar{\mathcal{K}}\sqrt{Nt\log(Nt)}\right\} \cdot \mathbb{P}_{\mathbf{X}}\left\{\hat{N}(t) \geq K_1\bar{\mathcal{K}}\sqrt{Nt\log(Nt)}\right\} \\
&\quad + \mathbb{E}_{\mathbf{X}}\left\{\hat{N}(t) \,\middle|\, \hat{N}(t) < K_1\bar{\mathcal{K}}\sqrt{Nt\log(Nt)}\right\} \cdot \mathbb{P}_{\mathbf{X}}\left\{\hat{N}(t) < K_1\bar{\mathcal{K}}\sqrt{Nt\log(Nt)}\right\} \\
&\leq Nt\,\frac{K_2\bar{\mathcal{K}}}{Nt} + K_1\bar{\mathcal{K}}\sqrt{Nt\log(Nt)} \\
&\leq (K_1 + K_2)\bar{\mathcal{K}}\sqrt{Nt\log(Nt)}\,.
\end{aligned} \tag{C.12}$$

Now, consider the event $\{\hat{\mathcal{C}}(\mathbf{X}_{i,t}) = \mathcal{C}(\mathbf{X}_{i,t})\}$ for any $t = M+1, \ldots, T$ and any $i = 1, \ldots, N$. By the Taylor expansion at $\varphi(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{x}_{i,t})}, \mathbf{x}_{i,t})$, we have that

$$\mathcal{R}^*(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{x}_{i,t})}, \mathbf{x}_{i,t}) - \mathcal{R}(p^\pi_{i,t}; \boldsymbol{\theta}_{\mathcal{C}(\mathbf{x}_{i,t})}, \mathbf{x}_{i,t}) \leq R''_{\max}[\varphi(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{x}_{i,t})}, \mathbf{x}_{i,t}) - p^\pi_{i,t}]^2\,, \tag{C.13}$$

where $R''_{\max} = \max\{|\frac{\partial^2\mathcal{R}}{\partial p^2}(p; \boldsymbol{\vartheta}, \mathbf{x})| \colon p \in \mathcal{P}, \boldsymbol{\vartheta} \in \Theta, \mathbf{x} \in \mathcal{X}\}$. Moreover, by the implicit function theorem, we further have

$$|\varphi(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{x}_{i,t})}, \mathbf{x}_{i,t}) - p^\pi_{i,t}| \leq \xi_2\|\boldsymbol{\theta}_{\mathcal{C}(\mathbf{x}_{i,t})} - \hat{\boldsymbol{\theta}}_{\mathcal{C}(\mathbf{x}_{i,t})}\|_2\,, \tag{C.14}$$

$\xi_2 = \max\{|\frac{\partial^2\mathcal{R}}{\partial p\partial\boldsymbol{\vartheta}}(p; \boldsymbol{\vartheta}, \mathbf{x})/\frac{\partial^2\mathcal{R}}{\partial p^2}(p; \boldsymbol{\vartheta}, \mathbf{x})| \colon p \in \mathcal{P}, \boldsymbol{\vartheta} \in \Theta, \mathbf{x} \in \mathcal{X}\}$. By taking $M = \lceil \kappa\sqrt{T/N}\rceil \vee 2$ with $\kappa > 1$, we deduce from (C.12), (C.13), (C.14), and Proposition 2 that

$$\begin{aligned}
&\Delta^\pi_{\boldsymbol{\theta}}(N, T) \\
&\leq \frac{1}{NT}\sum_{t=1}^M \sum_{i=1}^N \mathbb{E}^\pi_{\mathbf{X},\boldsymbol{\theta}}\left\{[\mathcal{R}^*(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t}) - \mathcal{R}(p^\pi_{i,t}; \boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t})]\right\}
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{NT}\sum_{t=M+1}^{T}\sum_{i=1}^{N}\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{[\mathcal{R}^*(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})},\mathbf{X}_{i,t})-\mathcal{R}(p^{\pi}_{i,t};\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})},\mathbf{X}_{i,t})]\mathbb{I}\{\hat{\mathcal{C}}(\mathbf{X}_{i,t})\neq\mathcal{C}(\mathbf{X}_{i,t})\}\right\}\\
&+\frac{1}{NT}\sum_{t=M+1}^{T}\sum_{i=1}^{N}\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{[\mathcal{R}^*(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})},\mathbf{X}_{i,t})-\mathcal{R}(p^{\pi}_{i,t};\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})},\mathbf{X}_{i,t})]\mathbb{I}\{\hat{\mathcal{C}}(\mathbf{X}_{i,t})=\mathcal{C}(\mathbf{X}_{i,t})\}\right\}\\
&\leq\frac{M}{T}R_{\text{rg}}+\frac{1}{NT}R_{\text{rg}}\mathbb{E}_{\mathbf{X}}[\hat{N}(T)]+\frac{T-M}{T}R''_{\max}\xi_2^2\frac{(K_3+K_4\theta_{\max})\bar{s}\log(dNM)}{NM}\\
&\leq\frac{2\kappa R_{\text{rg}}}{\sqrt{NT}}+R_{\text{rg}}\bar{\mathcal{K}}(K_1+K_2)\sqrt{\frac{\log(NT)}{NT}}+(K_3+K_4\theta_{\max})R''_{\max}\xi_2^2\bar{s}\frac{\log(dNT)}{\kappa N\sqrt{T}}\\
&\leq\frac{K_5\bar{\mathcal{K}}\bar{s}\log(dNT)}{\sqrt{NT}},
\end{aligned}
\tag{C.15}
$$

where $R_{\text{rg}}=\max\{\mathcal{R}(p;\boldsymbol{\vartheta},\mathbf{x}):p\in\mathcal{P},\boldsymbol{\vartheta}\in\Theta,\mathbf{x}\in\mathcal{X}\}-\min\{\mathcal{R}(p;\boldsymbol{\vartheta},\mathbf{x}):p\in\mathcal{P},\boldsymbol{\vartheta}\in\Theta,\mathbf{x}\in\mathcal{X}\}$, $\theta_{\max}=\max\{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2^2\colon\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta\}$, and $K_5=R_{\text{rg}}(2\kappa+K_1+K_2)+(K_3+K_4\theta_{\max})R''_{\max}\xi_2^2$. Q.E.D.

Proof of Lemma 1. We first note that it is sufficient to consider the case where $C_2\bar{s}\log(dNM)\leq NM$, because otherwise the result is trivial. Fix a cluster k. Let $N_k=\sum_{m=1}^{NM}\chi_m(k)$ be the number of observations that truly belong to class k and $\hat{N}_k=\sum_{m=1}^{NM}\hat{\chi}_m(k)$ be the number of observations that are classified to cluster k by spectral clustering, where $\chi_m(k)=\mathbb{I}\{\mathcal{C}(\mathbf{X}_m)=k\}$ and $\hat{\chi}_m(k)=\mathbb{I}\{\hat{\mathcal{C}}(\mathbf{X}_m)=k\}$, respectively. Based on our assumption that $\mathbb{P}_{\mathcal{X}}(R_k)\geq c_{\min}>0$ for all k, where R_k is the feature space of the true cluster k, we have $\mathbb{E}_{\mathbf{X}}(N_k)\geq c_{\min}NM$. Define $\Lambda'_k=\{m=1,\ldots,NM:\hat{\chi}_m(k)=0,\chi_m(k)=1\}$ and $O_k=\{m=1,\ldots,NM:\hat{\chi}_m(k)=1,\chi_m(k)=1\}$. On the events $\{N_k\geq c_{\min}NM/2\}$ and $\{|\Lambda'_k|<2K_1\sqrt{NM\log(NM)}\}$, where $|\cdot|$ denotes the cardinality of a set, we have $|O_k|=N_k-|\Lambda'_k|\geq\lceil\varrho_1 c_{\min}NM\rceil$ for some positive constant ϱ_1. We choose $\tilde{N}_k=\lceil\varrho_1 c_{\min}NM\rceil$ observations arbitrarily from O_k to form a subset $\tilde{O}_k$. It then follows that

$$
\begin{aligned}
\mu_{\min}(\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}[\mathcal{V}_k])&\geq\mu_{\min}\left(\sum_{m\in\tilde{O}_k}\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left[\mathbf{U}_m\mathbf{U}_m^{\intercal}\right]\right)\\
&=\mu_{\min}\left(\sum_{m\in\tilde{O}_k}\begin{bmatrix}1 & p_m & \boldsymbol{\mu}_{k,3}^{\intercal}\\ p_m & p_m^2 & p_m\boldsymbol{\mu}_{k,3}^{\intercal}\\ \boldsymbol{\mu}_{k,3} & p_m\boldsymbol{\mu}_{k,3} & \boldsymbol{\Sigma}_{k,3}+\boldsymbol{\mu}_{k,3}\boldsymbol{\mu}_{k,3}^{\intercal}\end{bmatrix}\right).
\end{aligned}
\tag{C.16}
$$

The inequality follows because $\mu_{\min}(\boldsymbol{A}+\boldsymbol{B})\geq\mu_{\min}(\boldsymbol{A})+\mu_{\min}(\boldsymbol{B})$ for symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, and $\mathbf{U}_m\mathbf{U}_m^{\intercal}$ is positive semi-definite for all m. To provide a lower bound on the right hand side of (C.16), we state the following lemma and defer the proof after the proof of Lemma 2.

LEMMA 3. *Let $\boldsymbol{\Sigma}$ be an $\ell\times\ell$ positive definite matrix. Then for any integer $n>1$,*

$$
\mu_{\min}\left(\sum_{m=1}^{n}\begin{bmatrix}1 & p_m & \boldsymbol{\mu}^{\intercal}\\ p_m & p_m^2 & p_m\boldsymbol{\mu}^{\intercal}\\ \boldsymbol{\mu} & p_m\boldsymbol{\mu} & \boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^{\intercal}\end{bmatrix}\right)\geq\frac{1}{(\|\boldsymbol{\mu}\|_2+1)^2+1}\min\left\{\mu_{\min}\left(\sum_{m=1}^{n}\begin{bmatrix}1 & p_m\\ p_m & p_m^2\end{bmatrix}\right),\ n\mu_{\min}(\boldsymbol{\Sigma})\right\},
\tag{C.17}
$$

if at least two elements of the set $\{p_m:m=1,\ldots,n\}$ are distinct, where $\mu_{\min}(\boldsymbol{A})$ denotes the minimum eigenvalue of matrix $\boldsymbol{A}$.

Based on Lemma 3, we further deduce that

$$\mu_{\min}(\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}[\mathcal{V}_k]) \geq \frac{1}{(\|\boldsymbol{\mu}_{k,3}\|_2+1)^2+1} \min\left\{ \mu_{\min}\left(\sum_{m=1}^{\tilde{N}_k} \begin{bmatrix} 1 & p_m \\ p_m & p_m^2 \end{bmatrix} \right),\ \tilde{N}_k \mu_{\min}(\boldsymbol{\Sigma}_{k,3}) \right\}$$
$$\geq \frac{1}{(\|\boldsymbol{\mu}_{k,3}\|_2+1)^2+1} \min\left\{ \tilde{\tau} \sum_{m\in\tilde{O}_k} (p_m - \bar{p}_{\tilde{N}_k})^2,\ \tilde{N}_k \mu_{\min}(\boldsymbol{\Sigma}_{k,3}) \right\}, \tag{C.18}$$

where $\bar{p}_{\tilde{N}_k} = \sum_{m\in\tilde{O}_k} p_m/\tilde{N}_k$ and $\tilde{\tau} = 2/(1+2p_{\max}-p_{\min})^2$. The second inequality follows directly from Lemma 2 of Keskin and Zeevi (2014). Since the prices are uniformly chosen from $\mathcal{P}$ and fixed before the selling season, there exists some $\tilde{p}_k > 0$ such that $\sum_{m\in\tilde{O}_k}(p_m - \bar{p}_{\tilde{N}_k})^2 \geq \tilde{p}_k$ and thus $\mu_{\min}(\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}[\mathcal{V}_k]) \geq \tau_{\min} NM$, where $\tau_{\min} = \varrho_1 c_{\min} \min\{\tilde{\tau}\tilde{p}_k, \mu_{\min}(\boldsymbol{\Sigma}_{k,3})\}/(\|\boldsymbol{\mu}_{k,3}\|_2+1)^2+1)$.

On the other hand, we observe that the maximum eigenvalue $\mu_{\max}(\mathbf{U}_m\mathbf{U}_m^{\intercal})$ of $\mathbf{U}_m\mathbf{U}_m^{\intercal}$ is upper bounded by a constant almost surely for all m, that is, $\mu_{\max}(\mathbf{U}_m\mathbf{U}_m^{\intercal}) \leq \mathrm{tr}(\mathbf{U}_m\mathbf{U}_m^{\intercal}) = \mathbf{U}_m^{\intercal}\mathbf{U}_m = 1+\mathbf{X}_m^{\intercal}\mathbf{X}_m + p_m^2 \leq 1 + x_{\max}^2 + p_{\max}^2 = R$, where $x_{\max} = \max\{\|\mathbf{x}\|_2 \colon \mathbf{x}\in\mathcal{X}\}$. By Theorem 3.1 of Tropp (2011) with $\delta = \frac{1}{2}$, we have that

$$\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{ \mu_{\min}(\mathcal{V}_k) \leq \frac{1}{2}\tau_{\min}NM,\ N_k \geq \frac{1}{2}c_{\min}NM,\ |\Lambda_k'| < 2K_1\sqrt{NM\log(NM)} \right\} \leq (d+2)e^{-\varrho_2 NM}, \tag{C.19}$$

where $\varrho_2 = \frac{1}{2R}(1-\log 2)\tau_{\min}$. Hence,

$$\begin{aligned}
&\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{ \mu_{\min}(\mathcal{V}_k) \leq \frac{1}{2}\tau_{\min}NM \right\} \\
&\quad\leq \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{ \mu_{\min}(\mathcal{V}_k) \leq \frac{1}{2}\tau_{\min}NM,\ N_k \geq \frac{1}{2}c_{\min}NM,\ |\Lambda_k'| < 2K_1\sqrt{NM\log(NM)} \right\} \\
&\qquad + \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{ N_k - \mathbb{E}_{\mathbf{X}}[N_k] < -\frac{1}{2}c_{\min}NM \right\} + \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{ |\Lambda_k'| \geq 2K_1\sqrt{NM\log(NM)} \right\} \\
&\quad\leq (d+2)e^{-\varrho_2 NM} + e^{-\frac{1}{2}c_{\min}^2 NM} + \frac{2K_2}{NM} \\
&\quad\leq 3de^{-\varrho_2 C_2 \bar{s}\log(dNM)} + \frac{2(c_{\min}^{-2}+K_2)}{NM}.
\end{aligned} \tag{C.20}$$

The first inequality follows from the union bound and $\mathbb{E}_{\mathbf{X}}[N_k] \geq c_{\min}NM$. The second inequality follows from the Hoeffding's inequality and Proposition 1. The last inequality follows because $C_2\bar{s}\log(dNM) \leq NM$ and $e^{-x} \leq \frac{1}{x}$ for all $x > 0$. We finish the proof by choosing $C_1 = \frac{1}{2}\tau_{\min}$ and $C_2 = (\varrho_2^{-1}\vee 3) + 2(c_{\min}^{-2} + K_2)$. Q.E.D.

Proof of Lemma 2. We prove this lemma by establishing bounds on the tail probability of the j^{th} component $\mathcal{M}_k^j(NM)$ of $\mathcal{M}_k(NM)$ for all $j \in \{1,\ldots,d+2\}$. For any given $k \in \{1,\ldots,\bar{\mathcal{K}}\}$ and $j \in \{1,\ldots,d+2\}$, define $Y_{k,\tilde{\psi}}^j(\ell) = \exp(\tilde{\psi}v_k\mathcal{M}_k^j(\ell) - \frac{1}{2}\tilde{\psi}v_k^2\ell)$ for $\ell = 2,3,\ldots,$ with $Y_{k,\tilde{\psi}}^j(1) = 1$ for some $\tilde{\psi} \geq 2/\underline{c}^2$, where $v_k = \frac{\lambda_k}{NM}$ and $\underline{c} = \min\{\tilde{c}_k : k = 1,\ldots,\bar{\mathcal{K}}\}$. Let $\mathcal{F}_\ell = \sigma(p_1,\ldots,p_\ell,\varepsilon_1,\ldots,\varepsilon_\ell,\mathbf{X}_1,\ldots,\mathbf{X}_{\ell+1})$, where $\ell = a_1N + a_2$ for some $a_1 \in \{0,\ldots,M-1\}$ and $a_2 \in \{0,\ldots,N-1\}$. Based on our assumptions on consumption shocks, there are positive constants $\bar{v}$ and η_0 such that $\mathbb{E}[\exp(\eta\varepsilon_\ell)|\mathcal{F}_{\ell-1}] \leq \exp(\bar{v}\sigma_0^2\eta^2)$ for all η satisfying $|\eta| \leq \eta_0$. Let $\bar{u} = \sup\{\|\mathbf{u}\|_\infty \colon \mathbf{u}^{\intercal} = [1\ p\ \mathbf{x}^{\intercal}], \mathbf{x}\in\mathcal{X}, p\in\mathcal{P}\}$ and $\tilde{c} = \max\{\tilde{c}_k : k = 1,\ldots,\bar{\mathcal{K}}\}$.

Recall that it suffices to consider the case where $C_3\bar{s}\log(dNM)/(NM) \leq 1$ since otherwise the result is trivial. Hence, by choosing $C_3 \geq (\tilde{\psi}\tilde{c}\overline{u}/\eta_0)^2$, we then have

$$|\tilde{\psi}v_k\hat{\chi}_m(k)U_m^j| \leq \tilde{\psi}\frac{\lambda_k}{NM}\overline{u} = \tilde{\psi}\tilde{c}_k\sqrt{\frac{\log(dNM)}{NM}}\overline{u} \leq \tilde{\psi}\tilde{c}_k\overline{u}C_3^{-1/2} \leq \eta_0, \tag{C.21}$$

for all $j \in \{1,\ldots,d+2\}$, where U_m^j is the j^{th} component of $\mathbf{U}_m$. Therefore,

$$\begin{aligned}\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}[Y^j_{k,\tilde{\psi}}(\ell)|\mathcal{F}_{\ell-1}] &= Y^j_{k,\tilde{\psi}}(\ell-1)e^{-\frac{1}{2}\tilde{\psi}v_k^2}\cdot\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left[\exp\left(\tilde{\psi}v_k\hat{\chi}_\ell(k)\varepsilon_\ell U_\ell^j\right)\middle|\mathcal{F}_{\ell-1}\right]\\ &\leq Y^j_{k,\tilde{\psi}}(\ell-1)\exp\left(-\frac{1}{2}\tilde{\psi}v_k^2+\overline{v}\sigma_0^2\tilde{\psi}^2v_k^2\hat{\chi}_\ell^2(k)\overline{u}^2\right) \leq Y^j_{k,\tilde{\psi}}(\ell-1)\end{aligned} \tag{C.22}$$

as long as $\tilde{\psi} \leq \frac{1}{2\overline{v}\sigma_0^2\overline{u}^2}$. This shows that $(Y^j_{k,\tilde{\psi}}(\ell),\mathcal{F}_\ell)$ is a supermartingale. As a result,

$$\begin{aligned}\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\mathcal{M}_k^j(NM) > \lambda_k\right\} &\leq \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\mathcal{M}_k^j(NM) > v_kNM\right\}\\ &\leq \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\exp\left(\tilde{\psi}v_k\mathcal{M}_k^j(NM)-\frac{1}{2}\tilde{\psi}v_k^2NM\right) \geq e^{\frac{1}{2}\tilde{\psi}v_k^2NM}\right\}\\ &\leq e^{-\frac{1}{2}\tilde{\psi}v_k^2NM} = e^{-\frac{1}{2}\tilde{\psi}\tilde{c}_k^2\log(dNM)},\end{aligned} \tag{C.23}$$

where the last inequality follows from the Markov inequality and the fact that $Y^j_{k,\tilde{\psi}}(1)=1$. Choosing $v_k = -\frac{\lambda_k}{NM}$ and using a similar argument yields the same upper bound on $\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\{\mathcal{M}_k^j(NM) < -\lambda_k\}$. Combining the above results and invoking the union bound, we deduce that

$$\begin{aligned}\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\mathcal{M}_k(NM)\|_\infty > \lambda_k\right\} &= \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\max_{j\in\{1,\ldots,d+2\}}\{|\mathcal{M}_k^j(NM)|\} > \lambda_k\right\}\\ &\leq \sum_{j=1}^{d+2}\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{|\mathcal{M}_k^j(NM)| > \lambda_k\right\}\\ &\leq 2(d+2)e^{-\frac{1}{2}\tilde{\psi}\tilde{c}_k^2\log(dNM)}\\ &\leq \frac{6\bar{s}\log(dNM)}{NM},\end{aligned} \tag{C.24}$$

since $\tilde{c}_k \geq \sqrt{2}\tilde{\psi}^{-1/2}$ by the choice of $\tilde{\psi}$. Choosing $C_3 = 6 \vee (\tilde{\psi}\tilde{c}\overline{u}/\eta_0)^2$ completes the proof. Q.E.D.

Proof of Lemma 3. Fix a vector $\mathbf{z} \in \mathbb{R}^{\ell+2}$ with $\|\mathbf{z}\|_2 = 1$ and partition it as $\mathbf{z}^\intercal = [\mathbf{z}_1^\intercal\ \mathbf{z}_2^\intercal]$, where $\mathbf{z}_1 \in \mathbb{R}^2$ and $\mathbf{z}_2 \in \mathbb{R}^\ell$. Let

$$\mathcal{J}_n = \sum_{m=1}^n\begin{bmatrix}1 & p_m & \boldsymbol{\mu}^\intercal\\ p_m & p_m^2 & p_m\boldsymbol{\mu}^\intercal\\ \boldsymbol{\mu} & p_m\boldsymbol{\mu} & \boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^\intercal\end{bmatrix} = \begin{bmatrix}n & \sum_{m=1}^n p_m & n\boldsymbol{\mu}^\intercal\\ \sum_{m=1}^n p_m & \sum_{m=1}^n p_m^2 & \sum_{m=1}^n p_m\boldsymbol{\mu}^\intercal\\ n\boldsymbol{\mu} & \sum_{m=1}^n p_m\boldsymbol{\mu} & n(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^\intercal)\end{bmatrix} \tag{C.25}$$

and

$$\mathscr{P}_n = \sum_{m=1}^n\begin{bmatrix}1 & p_m\\ p_m & p_m^2\end{bmatrix} = \begin{bmatrix}n & \sum_{m=1}^n p_m\\ \sum_{m=1}^n p_m & \sum_{m=1}^n p_m^2\end{bmatrix}. \tag{C.26}$$

Since $n > 1$ and at least two elements of $\{p_m : m = 1,\ldots,n\}$ are distinct, it follows that

$$\begin{aligned}\det(\mathscr{P}_n) &= n\sum_{m=1}^n p_m^2 - \left(\sum_{m=1}^n p_m\right)^2\\ &= n\sum_{m=1}^n p_m^2 - \left(\sum_{m=1}^n p_m^2 + 2\sum_{m=1}^n\sum_{m'=m+1}^n p_mp_{m'}\right)\end{aligned}$$

$$= \sum_{m=1}^{n} \sum_{m'=m+1}^{n} (p_m - p_{m'})^2 > 0\,, \tag{C.27}$$

and

$$\mathscr{P}_n^{-1} = \frac{1}{\det(\mathscr{P}_n)} \begin{bmatrix} \sum_{m=1}^n p_m^2 & -\sum_{m=1}^n p_m \\ -\sum_{m=1}^n p_m & n \end{bmatrix}. \tag{C.28}$$

Consequently,

$$\begin{aligned}
\mathbf{z}^\top \mathcal{J}_n \mathbf{z} &= \mathbf{z}_1^\top \mathscr{P}_n \mathbf{z}_1 + \mathbf{z}_1^\top \begin{bmatrix} n\mu^\top \\ \sum_{m=1}^n p_m \boldsymbol{\mu}^\top \end{bmatrix} \mathbf{z}_2 + \mathbf{z}_2^\top \begin{bmatrix} n\boldsymbol{\mu} & \sum_{m=1}^n p_m \boldsymbol{\mu} \end{bmatrix} \mathbf{z}_1 + n\mathbf{z}_2^\top (\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top)\mathbf{z}_2 \\
&= \left(\mathbf{z}_1^\top + \mathbf{z}_2^\top \begin{bmatrix} n\boldsymbol{\mu} & \sum_{m=1}^n p_m \boldsymbol{\mu} \end{bmatrix} \mathscr{P}_n^{-1}\right) \mathscr{P}_n \mathbf{z}_1 + \mathbf{z}_1^\top \begin{bmatrix} n\boldsymbol{\mu}^\top \\ \sum_{m=1}^n p_m \boldsymbol{\mu}^\top \end{bmatrix} \mathbf{z}_2 \\
&\quad + \mathbf{z}_2^\top \begin{bmatrix} n\boldsymbol{\mu} & \sum_{m=1}^n p_m \boldsymbol{\mu} \end{bmatrix} \mathscr{P}_n^{-1} \begin{bmatrix} n\boldsymbol{\mu}^\top \\ \sum_{m=1}^n p_m \boldsymbol{\mu}^\top \end{bmatrix} \mathbf{z}_2 \\
&\quad + n\mathbf{z}_2^\top \left(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top - \frac{1}{n} \begin{bmatrix} n\boldsymbol{\mu} & \sum_{m=1}^n p_m \boldsymbol{\mu} \end{bmatrix} \mathscr{P}_n^{-1} \begin{bmatrix} n\boldsymbol{\mu}^\top \\ \sum_{m=1}^n p_m \boldsymbol{\mu}^\top \end{bmatrix} \right) \mathbf{z}_2 \\
&= \left(\mathbf{z}_1^\top + \mathbf{z}_2^\top \begin{bmatrix} n\boldsymbol{\mu} & \sum_{m=1}^n p_m \boldsymbol{\mu} \end{bmatrix} \mathscr{P}_n^{-1}\right) \mathscr{P}_n \left(\mathbf{z}_1 + \mathscr{P}_n^{-1} \begin{bmatrix} n\boldsymbol{\mu}^\top \\ \sum_{m=1}^n p_m \boldsymbol{\mu}^\top \end{bmatrix} \mathbf{z}_2 \right) \\
&\quad + n\mathbf{z}_2^\top \left(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top - \frac{1}{n} \begin{bmatrix} n\boldsymbol{\mu} & \sum_{m=1}^n p_m \boldsymbol{\mu} \end{bmatrix} \mathscr{P}_n^{-1} \begin{bmatrix} n\boldsymbol{\mu}^\top \\ \sum_{m=1}^n p_m \boldsymbol{\mu}^\top \end{bmatrix} \right) \mathbf{z}_2 \\
&= (\mathbf{z}_1^\top + \mathbf{z}_2^\top [\boldsymbol{\mu}\ \mathbf{0}])\, \mathscr{P}_n\, (\mathbf{z}_1 + [\boldsymbol{\mu}\ \mathbf{0}]^\top \mathbf{z}_2) + n\mathbf{z}_2^\top \boldsymbol{\Sigma} \mathbf{z}_2\,.
\end{aligned} \tag{C.29}$$

If $\|\mathbf{z}_1\|_2 \geq \frac{\|\boldsymbol{\mu}\|_2+1}{\sqrt{(\|\boldsymbol{\mu}\|_2+1)^2+1}}$, then it follows by $\|\mathbf{z}\|_2 = 1$ that $\|\mathbf{z}_2\|_2 \leq \frac{1}{\sqrt{(\|\boldsymbol{\mu}\|_2+1)^2+1}}$. In this case, we further have that

$$\begin{aligned}
\mathbf{z}^\top \mathcal{J}_n \mathbf{z} &\geq (\mathbf{z}_1^\top + \mathbf{z}_2^\top [\boldsymbol{\mu}\ \mathbf{0}])\, \mathscr{P}_n\, (\mathbf{z}_1 + [\boldsymbol{\mu}\ \mathbf{0}]^\top \mathbf{z}_2) \\
&\geq \mu_{\min}(\mathscr{P}_n) \cdot \|\mathbf{z}_1 + [\boldsymbol{\mu}\ \mathbf{0}]^\top \mathbf{z}_2\|_2^2 \\
&\geq \mu_{\min}(\mathscr{P}_n)\, (\|\mathbf{z}_1\|_2 - \|\mathbf{z}_2\|_2 \cdot \|\boldsymbol{\mu}\|_2)^2 \\
&\geq \frac{\mu_{\min}(\mathscr{P}_n)}{(\|\boldsymbol{\mu}\|_2+1)^2+1}\,.
\end{aligned} \tag{C.30}$$

If $\|\mathbf{z}_1\|_2 < \frac{\|\boldsymbol{\mu}\|_2+1}{\sqrt{(\|\boldsymbol{\mu}\|_2+1)^2+1}}$, then it follows by $\|\mathbf{z}\|_2 = 1$ that $\|\mathbf{z}_2\|_2 \geq \frac{1}{\sqrt{(\|\boldsymbol{\mu}\|_2+1)^2+1}}$. In this case, we have that

$$\mathbf{z}^\top \mathcal{J}_n \mathbf{z} \geq n\mathbf{z}_2^\top \boldsymbol{\Sigma} \mathbf{z}_2 \geq n\mu_{\min}(\boldsymbol{\Sigma}) \cdot \|\mathbf{z}_2\|_2^2 \geq \frac{n\mu_{\min}(\boldsymbol{\Sigma})}{(\|\boldsymbol{\mu}\|_2+1)^2+1}\,. \tag{C.31}$$

In sum, we have

$$\mathbf{z}^\top \mathcal{J}_n \mathbf{z} \geq \frac{1}{(\|\boldsymbol{\mu}\|_2+1)^2+1} \min\left\{\mu_{\min}(\mathscr{P}_n), n\mu_{\min}(\boldsymbol{\Sigma})\right\}. \tag{C.32}$$

Since it holds for any $\mathbf{z} \in \mathbb{R}^{\ell+2}$ satisfying $\|\mathbf{z}\|_2 = 1$, the desired result follows by the Rayleigh-Ritz theorem.

Q.E.D.

Appendix D: Proofs of Results in §5.2

Proof of Proposition 3. Let $\mathbf{U}_{t,1}^\intercal = [1\ \mathbf{X}_{t,1}^\intercal]$, $\mathbf{U}_{i,2}^\intercal = [1\ \mathbf{X}_{i,2}^\intercal]$ and $\mathbf{U}_{i,t,3}^\intercal = [1\ p_{i,t}\ \mathbf{X}_{i,t,3}^\intercal]$. Thus,

$$\mathbf{U}_{t,1}\mathbf{U}_{t,1}^\intercal = \begin{bmatrix} 1 & \mathbf{X}_{t,1}^\intercal \\ \mathbf{X}_{t,1} & \mathbf{X}_{t,1}\mathbf{X}_{t,1}^\intercal \end{bmatrix} \tag{D.1}$$

$$\mathbf{U}_{i,2}\mathbf{U}_{i,2}^\intercal = \begin{bmatrix} 1 & \mathbf{X}_{i,2}^\intercal \\ \mathbf{X}_{i,2} & \mathbf{X}_{i,2}\mathbf{X}_{i,2}^\intercal \end{bmatrix} \tag{D.2}$$

$$\mathbf{U}_{i,t,3}\mathbf{U}_{i,t,3}^\intercal = \begin{bmatrix} 1 & p_{i,t} & \mathbf{X}_{i,t,3}^\intercal \\ p_{i,t} & p_{i,t}^2 & p_{i,t}\mathbf{X}_{i,t,3}^\intercal \\ \mathbf{X}_{i,t,3} & p_{i,t}\mathbf{X}_{i,t,3} & \mathbf{X}_{i,t,3}\mathbf{X}_{i,t,3}^\intercal \end{bmatrix}. \tag{D.3}$$

Fix a cluster, and denote by k_1, k_2, and k_3 the cluster labels corresponding to the three types of features, respectively. For each $b \in \{1,2,3\}$, we create dummy clusters to make $\mathcal{K}_b = \bar{\mathcal{K}}_b$. Let k be the linear index of the cluster label vector $[k_1, k_2, k_3]^\intercal$, which is given by $k = \bar{\mathcal{K}}_2\bar{\mathcal{K}}_3(k_1-1) + \bar{\mathcal{K}}_3(k_2-1) + k_3$. For each $b \in \{1,2,3\}$, define a mapping $\mathcal{C}_b : \mathcal{X}_b \mapsto \{1,\ldots,\bar{\mathcal{K}}_b\}$ such that $\mathcal{C}_b(\mathbf{X}_{i,t,b}) = \text{Proj}_b(\mathcal{C}(\mathbf{X}_{i,t}))$ for all $i \in \mathcal{N}$ and $t \in \mathcal{T}$, where $\text{Proj}_b(\cdot)$ is the projection operator to the b^{th} component of the cluster label vector. The mapping $\hat{\mathcal{C}}_b(\cdot)$ that represents the estimated cluster labels is defined in the same way, for all $b \in \{1,2,3\}$.

We first consider the regression over the time-heterogeneous features $\mathbf{X}_{t,1} = \mathbf{X}_{i,t,1}$ in cluster k (for which the first component of the cluster label vector is k_1), and denote the coefficients as $\boldsymbol{\theta}_{k,1}^\intercal = [\gamma_{k,1}\ \boldsymbol{\alpha}_{k,1}^\intercal]$. As in the proof of Proposition 2, we define $\mathcal{V}_{k_1} = \sum_{t=1}^M \hat{\chi}_t(k_1)\mathbf{U}_{t,1}\mathbf{U}_{t,1}^\intercal$, $\mathcal{M}_{k_1}(\ell) = \sum_{t=1}^\ell \hat{\chi}_t(k_1)\varepsilon_t\mathbf{U}_{t,1}$, and $\boldsymbol{\zeta}_{k_1} = \sum_{t\in\Lambda_{k_1}}[g(\boldsymbol{\theta}_{\mathcal{C}_1(\mathbf{X}_{t,1})}^\intercal\mathbf{U}_{t,1}) - g(\boldsymbol{\theta}_{k_1}^\intercal\mathbf{U}_{t,1})]\mathbf{U}_{t,1}$, where $\hat{\chi}_t(k_1) = \mathbb{I}\{\hat{\mathcal{C}}_1(\mathbf{X}_{t,1}) = k_1\}$, $\chi_t(k_1) = \mathbb{I}\{\mathcal{C}_1(\mathbf{X}_{t,1}) = k_1\}$, and $\Lambda_{k_1} = \{t = 1,\ldots,M : \hat{\chi}_t(k_1) = 1,\ \chi_t(k_1) = 0\}$. Thus, following the same procedure as in the proof of Lemma 1, we conclude that for any $k_1 \in \{1,\ldots,\bar{\mathcal{K}}_1\}$, there exist positive constants C_5 and C_6 such that $\mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\{\mu_{\min}(\mathcal{V}_{k_1}) \le C_5 M\} \le C_6\bar{s}\log(d_1M)/M$. Since the proof of Lemma 2 does not involve the i.i.d. assumption on the feature vectors, the result follows directly by choosing $\lambda_{k,1} = \tilde{c}_{k,1}\sqrt{M\log(d_1M)}$ for some $\tilde{c}_{k,1} > 0$ and by replacing NM with M. Therefore, there exist positive constants C_7 and C_8 such that

$$\mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\hat{\boldsymbol{\alpha}}_{k,1}(\lambda_{k,1}) - \boldsymbol{\alpha}_{k,1}\|_2^2 \ge \frac{C_7\bar{s}\log(d_1M)}{M}\right\} \le \frac{C_8\bar{s}\log(d_1M)}{M} \tag{D.4}$$

for any $k_1 \in \{1,\ldots,\bar{\mathcal{K}}_1\}$ and all $M \ge 2$. Similarly, we consider the regression over the user-heterogeneous features $\mathbf{X}_{i,2} = \mathbf{X}_{i,t,2}$ in cluster k (for which the second component of the cluster label vector is k_2) and denote the coefficients as $\boldsymbol{\theta}_{k,2}^\intercal = [\gamma_{k,2}\ \boldsymbol{\alpha}_{k,2}^\intercal]$. In this case, we conclude that there exist positive constants C_9 and C_{10} such that

$$\mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\hat{\boldsymbol{\alpha}}_{k,2}(\lambda_{k,2}) - \boldsymbol{\alpha}_{k,2}\|_2^2 \ge \frac{C_9\bar{s}\log(d_2N)}{N}\right\} \le \frac{C_{10}\bar{s}\log(d_2N)}{N} \tag{D.5}$$

for all $k_2 \in \{1,\ldots,\bar{\mathcal{K}}_2\}$ and all $N \ge 2$, where $\lambda_{k,2} = \tilde{c}_{k,2}\sqrt{N\log(d_2N)}$ for some $\tilde{c}_{k,2} > 0$. Moreover, the result of Lemma 2 is directly applicable to the regression over the fully heterogeneous features and prices. Therefore, denoting the coefficients for this regression as $\boldsymbol{\theta}_{k,3}^\intercal = [\gamma_{k,3}\ \boldsymbol{\alpha}_{k,3}^\intercal\ \beta_k]$, we deduce that there exist positive constants C_{11} and C_{12} such that

$$\mathbb{P}^\pi_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\hat{\boldsymbol{\alpha}}_{k,3}(\lambda_{k,3}) - \boldsymbol{\alpha}_{k,3}\|_2^2 + (\hat{\beta}_k(\lambda_{k,3}) - \beta_k)^2 \ge \frac{C_{11}\bar{s}\log(d_3MN)}{MN}\right\} \le \frac{C_{12}\bar{s}\log(d_3MN)}{MN} \tag{D.6}$$

for all $k_3 \in \{1,\dots,\bar{\mathcal{K}}_3\}$ and all $M \geq 2$, where $\lambda_{k,3} = \tilde{c}_{k,3}\sqrt{MN\log(d_3MN)}$ for some $\tilde{c}_{k,3} > 0$. Finally, the estimator $\hat{\gamma}_k$ of the intercept γ_k in cluster k is given by

$$\hat{\gamma}_k(\boldsymbol{\lambda}_k) = \frac{1}{N_k}\sum_{t=1}^{M}\sum_{i=1}^{N}\hat{\chi}_{i,t}(k)[\hat{\gamma}_{k,3}(\lambda_{k,3}) - \hat{\boldsymbol{\alpha}}_{k,1}^{\intercal}(\lambda_{k,1})\mathbf{X}_{t,1} - \hat{\boldsymbol{\alpha}}_{k,2}^{\intercal}(\lambda_{k,2})\mathbf{X}_{i,2}],$$

where $\boldsymbol{\lambda}_k = (\lambda_{k,1},\lambda_{k,2},\lambda_{k,3})$, $\hat{\chi}_{i,t}(k) = \mathbb{I}\{\hat{\mathcal{C}}(\mathbf{X}_{i,t}) = k\}$, and $N_k = \sum_{t=1}^{M}\sum_{i=1}^{N}\hat{\chi}_{i,t}(k)$. By the triangle inequality and the Cauchy-Schwarz inequality, we deduce that for all $i \in \{1,\dots,N\}$ and $t \in \{1,\dots,M\}$,

$$\begin{aligned}
&|\hat{\gamma}_{k,3}(\lambda_{k,3}) - (\hat{\boldsymbol{\alpha}}_{k,1}(\lambda_{k,1})^{\intercal}\mathbf{X}_{t,1} + \hat{\boldsymbol{\alpha}}_{k,2}(\lambda_{k,2})^{\intercal}\mathbf{X}_{i,2}) - \gamma_k| \\
&\quad\leq \left|\hat{\gamma}_{k,3}(\lambda_{k,3}) - (\gamma_k + \boldsymbol{\alpha}_{k,1}^{\intercal}\mathbf{X}_{t,1} + \boldsymbol{\alpha}_{k,2}^{\intercal}\mathbf{X}_{i,2})\right| + B_1\left\|\hat{\boldsymbol{\alpha}}_{k,1}(\lambda_{k,1}) - \boldsymbol{\alpha}_{k,1}\right\|_2 + B_2\left\|\hat{\boldsymbol{\alpha}}_{k,2}(\lambda_{k,2}) - \boldsymbol{\alpha}_{k,2}\right\|_2,
\end{aligned}$$

where $B_1 = \max\{\|\mathbf{x}_1\|_2 \colon \mathbf{x}_1 \in \mathcal{X}_1\}$ and $B_2 = \max\{\|\mathbf{x_2}\|_2 \colon \mathbf{x}_2 \in \mathcal{X}_2\}$. As a result, there exist positive constants C_{13} and C_{14} such that

$$\begin{aligned}
&\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\left|\hat{\gamma}_{k,3}(\lambda_{k,3}) - \hat{\boldsymbol{\alpha}}_{k,1}^{\intercal}(\lambda_{k,1})\mathbf{X}_{t,1} - \hat{\boldsymbol{\alpha}}_{k,2}^{\intercal}(\lambda_{k,2})\mathbf{X}_{i,2} - \gamma_k\right|^2 \geq \frac{C_{13}\bar{s}\log(dNM)}{N\wedge M}\right\} \\
&\quad\leq \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\left|\hat{\gamma}_{k,3}(\lambda_{k,3}) - (\gamma_k + \boldsymbol{\alpha}_{k,1}^{\intercal}\mathbf{X}_{t,1} + \boldsymbol{\alpha}_{k,2}^{\intercal}\mathbf{X}_{i,2})\right|^2 \geq \frac{C_{11}\bar{s}\log(d_3NM)}{NM}\right\} \\
&\qquad + \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\hat{\boldsymbol{\alpha}}_{k,1}(\lambda_{k,1}) - \boldsymbol{\alpha}_{k,1}\|_2^2 \geq \frac{C_7\bar{s}\log(d_1M)}{M}\right\} \\
&\qquad + \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\hat{\boldsymbol{\alpha}}_{k,2}(\lambda_{k,2}) - \boldsymbol{\alpha}_{k,2}\|_2^2 \geq \frac{C_9\bar{s}\log(d_2N)}{N}\right\} \\
&\quad\leq \frac{C_{14}\bar{s}\log(dNM)}{N\wedge M},
\end{aligned} \tag{D.7}$$

for all N and M such that $N \wedge M \geq 2$. Hence,

$$\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{|\hat{\gamma}_k(\boldsymbol{\lambda}_k) - \gamma_k|^2 \geq \frac{C_{13}\bar{s}\log(dNM)}{N\wedge M}\right\} \leq \frac{C_{14}\bar{s}\log(dNM)}{N\wedge M}. \tag{D.8}$$

Combining the results in (D.4), (D.5), (D.6), and (D.8) completes the proof. Q.E.D.

Proof of Theorem 2. Since there are three types of features, misclassification error of a time-heterogeneous feature transfers to every user in that period, and misclassification error of a user-heterogeneous feature carries through every time period. Therefore, by Proposition 1, the expected total number of misclassified features for the first t periods satisfies

$$\mathbb{E}_{\mathbf{X}}[\hat{N}(t)] \leq (K_1+K_2)\bar{\mathcal{K}}\left(N\sqrt{t\log t} + t\sqrt{N\log N} + \sqrt{Nt\log(Nt)}\right). \tag{D.9}$$

By taking $M = \lceil\kappa\sqrt{T}\rceil \vee 2$ with $\kappa > 1$, we deduce from (D.9), (C.13), (C.14), and Proposition 3 that

$$\begin{aligned}
&\Delta^{\pi}_{\boldsymbol{\theta}}(N,T) \\
&\quad\leq \frac{1}{NT}\sum_{t=1}^{M}\sum_{i=1}^{N}\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{[\mathcal{R}^*(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})},\mathbf{X}_{i,t}) - \mathcal{R}(p^{\pi}_{i,t};\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})},\mathbf{X}_{i,t})]\right\} \\
&\qquad + \frac{1}{NT}\sum_{t=M+1}^{T}\sum_{i=1}^{N}\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{[\mathcal{R}^*(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})},\mathbf{X}_{i,t}) - \mathcal{R}(p^{\pi}_{i,t};\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})},\mathbf{X}_{i,t})]\mathbb{I}\{\hat{\mathcal{C}}(\mathbf{X}_{i,t}) \neq \mathcal{C}(\mathbf{X}_{i,t})\}\right\} \\
&\qquad + \frac{1}{NT}\sum_{t=M+1}^{T}\sum_{i=1}^{N}\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{[\mathcal{R}^*(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})},\mathbf{X}_{i,t}) - \mathcal{R}(p^{\pi}_{i,t};\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})},\mathbf{X}_{i,t})]\mathbb{I}\{\hat{\mathcal{C}}(\mathbf{X}_{i,t}) = \mathcal{C}(\mathbf{X}_{i,t})\}\right\}
\end{aligned}$$

$$\leq \frac{M}{T}R_{\rm rg} + \frac{1}{NT}R_{\rm rg}\mathbb{E}_{\mathbf{X}}[\hat{N}(T)] + \frac{T-M}{T}R''_{\max}\xi_2^2\frac{(K_6+K_7\theta_{\max})\bar{s}\log(dNM)}{N\wedge M}$$
$$\leq \frac{2\kappa R_{\rm rg}}{\sqrt{T}} + 3R_{\rm rg}\bar{\mathcal{K}}(K_1+K_2)\sqrt{\frac{\log(NT)}{N\wedge T}} + (K_6+K_7\theta_{\max})R''_{\max}\xi_2^2\bar{s}\frac{\log(dNT)}{N\wedge(\kappa\sqrt{T})}$$
$$\leq \frac{K_8\bar{\mathcal{K}}\bar{s}\log(dNT)}{\sqrt{N\wedge T}}, \tag{D.10}$$

where $K_8 = R_{\rm rg}(2\kappa+3K_1+3K_2)+(K_6+K_7\theta_{\max})R''_{\max}\xi_2^2$. Q.E.D.

Appendix E: Proofs of Results in §5.3

Proof of Proposition 4. We first summarize some useful properties about $\{J_n : n\in\mathbb{N}\}$ and J:

(a) In the proof of Proposition 1, we have shown that $\{J_n : n\in\mathbb{N}\}$ and J are all compact integral operators. By Theorem 5.2.8(ii) of Bühler and Salamon (2018), 0 is the only possible accumulation point in the spectrum of J, and all non-zero eigenvalues of J are isolated.

(b) J is also self-adjoint, so its spectrum $\sigma(J)\subset\mathbb{R}$, see e.g., Theorem 5.3.16(i) of Bühler and Salamon (2018).

(c) As a direct consequence of being self-adjoint, J is a normal operator, so the spectral radius $\sup_{\rho\in\sigma(J)}|\rho|$ equals its operator norm $\|J\|$, see e.g. Theorem 5.3.15(ii) of Bühler and Salamon (2018).

(d) The set $\mathcal{J}=\{J\}\cup\{J_n : n\in\mathbb{N}\}$ is uniformly bounded, that is, $\sup\{\|\tilde{J}\|\colon \tilde{J}\in\mathcal{J}\}<\infty$. This is because for any $f\in C(\mathcal{Z})$,

$$\sup_{\tilde{J}\in\mathcal{J}}\|\tilde{J}f\|_\infty \leq \frac{\psi_{\max}}{\psi_{\min}}\|f\|_\infty < \infty, \tag{E.1}$$

where $\psi_{\max}=\max\{\psi(\mathbf{x},\mathbf{y}) : (\mathbf{x},\mathbf{y})\in\mathcal{X}\times\mathcal{X}\}$ and $\psi_{\min}=\min\{\psi(\mathbf{x},\mathbf{y}) : (\mathbf{x},\mathbf{y})\in\mathcal{X}\times\mathcal{X}\}$. The uniform upper bound is $r_{\mathcal{J}}=\psi_{\max}/\psi_{\min}$, which follows immediately by the Banach-Steinhaus theorem.

According to Theorem 16 of von Luxburg et al. (2008), there exists a constant $\tilde{C}>0$ and a sequence $\{s_n : n\in\mathbb{N}\}$ of signs with $s_n\in\{-1,1\}$ such that

$$\|s_nu_n - u\|_\infty \leq \tilde{C}\sup_{f\in\mathscr{F}}|\mathbb{P}_nf - \mathbb{P}_{\mathcal{Z}}f|, \tag{E.2}$$

where $\mathscr{F}$ is defined as in Definition 10 of von Luxburg et al. (2008). Fix a non-zero simple eigenvalue ρ of J. Choose a positive constant $r_\rho<\operatorname{dist}(\rho,\sigma(J)\setminus\{\rho\})$, and define $\Gamma_1=\{\rho'\in\mathbb{C} : |\rho-\rho'|=r_\rho\}$ and Γ_2 as a simple closed rectifiable curve with $\Gamma_1\cap\Gamma_2=\varnothing$, containing $\sigma(J)\setminus\{\rho\}$ in the interior and excluding ρ in the exterior. Let $F=B_{r_{\mathcal{J}}}(0)\setminus(\operatorname{int}(\Gamma_1)\cup\operatorname{int}(\Gamma_2))$, where $B_{r_{\mathcal{J}}}(0)$ is the ball of radius $r_{\mathcal{J}}$ in $\mathbb{C}$ centered at 0, and $\operatorname{int}(\cdot)$ denotes the interior of a closed curve. Clearly, F is a compact subset of the resolvent set $R(J)=\mathbb{C}\setminus\sigma(J)$. Let $r_0=\operatorname{dist}(0,F)$ and R_F be a uniform upper bound on the operator norm of the resolvent operator $R_{\rho'}(J)=(\rho'\mathbf{1}-J)^{-1}$ associated with ρ' over all $\rho'\in F$, that is, $\sup_{\rho'\in F}\|R_{\rho'}(J)\|\leq R_F$. Then, by Theorem 3(b) of Atkinson (1967), we have $\tilde{C}=2C(3r_{\mathcal{J}}+1)$, where $C=2R_F^2r_\rho/r_0$.

Note that $R_{\rho'}(J)$ is bijective for any $\rho'\in F$ by definition of the resolvent set, so the adjoint of $R_{\rho'}(J)$ is $((\rho'\mathbf{1}-J)^{-1})^*=((\rho'\mathbf{1}-J)^*)^{-1}=(\bar{\rho}'\mathbf{1}-J)^{-1}=R_{\bar{\rho}'}(J)$ by Lemma 5.3.9 of Bühler and Salamon (2018), where $\bar{\rho}'$ is the complex conjugate of ρ'. By the resolvent identity, we have $R_{\rho'}(J)R_{\bar{\rho}'}(J)=R_{\bar{\rho}'}(J)R_{\rho'}(J)$,

implying that $R_{\rho'}(J)$ is a normal operator. Hence, $\|R_{\rho'}(J)\| = \sup\{|\tilde{\rho}|\colon \tilde{\rho} \in \sigma(R_{\rho'}(J))\}$. For any $\rho_0 \neq \rho'$, observe that $(\rho' - \rho_0)^{-1}(\rho_0 \mathbf{1} - J) = ((\rho' - \rho_0)^{-1}\mathbf{1} - R_{\rho'}(J))(\rho' \mathbf{1} - J)$. Since $\rho' \in F \subset R(J)$, it follows that $\rho' \mathbf{1} - J$ is bijective, so $\rho_0 \mathbf{1} - J$ is bijective if and only if $(\rho' - \rho_0)^{-1}\mathbf{1} - R_{\rho'}(J)$ is bijective if and only if $(\rho' - \rho_0)^{-1} \in R(R_{\rho'}(J))$, that is, ρ_0 is in the resolvent set of J if and only if $(\rho' - \rho_0)^{-1}$ is in the resolvent set of $R_{\rho'}(J)$. Therefore, $\sigma(R_{\rho'}(J)) = \{(\rho' - \rho_0)^{-1} : \rho_0 \in \sigma(J)\}$. It then follows that

$$\|R_{\rho'}(J)\| = \sup_{\rho_0 \in \sigma(J)} |(\rho' - \rho_0)^{-1}| = \operatorname{dist}(\rho', \sigma(J))^{-1}. \tag{E.3}$$

Since $\sup_{\rho' \in F} \|R_{\rho'}(J)\| \leq R_F$, we then have $\inf_{\rho' \in F} \operatorname{dist}(\rho', \sigma(J)) \geq R_F^{-1}$. Hence, as r_ρ goes to 0, $\inf_{\rho' \in F} \operatorname{dist}(\rho', \sigma(J))$ also goes to 0 and R_F goes to ∞. Since $r_0 \to \rho$ as $r_\rho \to 0$, we deduce that $C \to \infty$ as $r_\rho \to 0$. On the other hand, as $r_\rho \to \operatorname{dist}(\rho, \sigma(J) \setminus \{\rho\})$, $r_0 \to \rho - \operatorname{dist}(\rho, \sigma(J) \setminus \{\rho\})$ and R_F also goes to ∞. Therefore, to find the smallest $\tilde{C}$ such that (E.2) holds, it suffices to find the tightest upper bound R_F on $\sup_{\rho' \in F} \|R_{\rho'}(J)\|$ by choosing the set F, or equivalently, the two closed curves Γ_1 and Γ_2. Based on the above reasoning, the best choice of F such that R_F is the smallest is to let Γ_1 and Γ_2 be tangent at the midpoint between ρ and the closest point to it in $\sigma(J) \setminus \{\rho\}$. Moreover, the distance between the origin and any point on Γ_2 is at least $\rho - \operatorname{dist}(\rho, \sigma(J) \setminus \{\rho\})/2$ because otherwise r_0 would be smaller than $\rho - \operatorname{dist}(\rho, \sigma(J) \setminus \{\rho\})/2$ and C would be larger. Under this choice of F, the tightest upper bound on $\sup_{\rho' \in F} \|R_{\rho'}(J)\|$ is given by

$$R_\rho = \frac{2}{\operatorname{dist}(\rho, \sigma(J) \setminus \{\rho\})} = \frac{1}{r_\rho}. \tag{E.4}$$

Furthermore, we have $r_0 = \rho - r_\rho$, so the best choice of the constant C is given by $2R_\rho^2/(R_\rho \rho - 1)$. The smallest $\tilde{C}$ for (E.2) to hold is then obtained by plugging in the best choice of C.

Appendix F: Model Parameters Used in §6

F.1. Parameter Values for §6.1.1

The means of the mixture of Gaussian distributions are

$$\begin{aligned}
\boldsymbol{\mu}_1 &= [\ 0.6\ \ 1\ \ -0.6\ \ 1.2\ \ 1.3\ \ -1\ \ 2\ \ 1.5\ \ 0.7\ \ 1.6\ ]^\intercal, \\
\boldsymbol{\mu}_2 &= [\ 1.5\ \ 0\ \ -1.5\ \ 0.2\ \ -0.5\ \ 0.4\ \ 3\ \ -0.3\ \ -1\ \ -0.2\ ]^\intercal, \\
\boldsymbol{\mu}_3 &= [\ -0.8\ \ 2.1\ \ 0.9\ \ -1\ \ -1.2\ \ 1.3\ \ 0.8\ \ -1.2\ \ 0\ \ -0.9\ ]^\intercal,
\end{aligned}$$

and common covariance is $0.05\boldsymbol{I}_{10}$. The feature coefficients are given by

$$\boldsymbol{\alpha} = \begin{bmatrix} 0.8 & 0 & -0.2 & 0 & 0 & 0.4 & 0 & 0.6 & -1 & 0 \\ 0 & 0.3 & 0 & 0 & 0.5 & 0.9 & 0.6 & 0 & 0 & 0.1 \\ 0 & -1 & 0 & 0 & -0.8 & -0.5 & 0 & 0.9 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{3 \times 10},$$

the price coefficients are $\boldsymbol{\beta} = [-1.5\ \ -1\ \ -0.8]^\intercal$, and the intercept parameters are $\boldsymbol{\gamma} = [20\ 12\ 15]^\intercal$. The consumption model is linear (i.e., the link function $g(\cdot)$ is the identity function), and the consumption shocks $\{\varepsilon_t\}$ are independently and identically drawn from the normal distribution with standard deviation 0.01.

F.2. Parameter Values for §6.1.2

The means of the mixture of Gaussian distributions are

$$\boldsymbol{\mu}_1 = [\ 1.5 \quad 1 \quad\ \ 0.4 \quad 1.2\ ]^\intercal,$$
$$\boldsymbol{\mu}_2 = [\ 0.6 \ -0.2 \ \ -1.5 \quad 0.2\ ]^\intercal,$$

and common covariance is $0.05\boldsymbol{I}_4$. The feature coefficients are given by

$$\boldsymbol{\alpha} = \begin{bmatrix} -0.2 & 0 & -0.8 & 0 \\ 0 & -0.3 & 0 & 0.9 \end{bmatrix} \in \mathbb{R}^{2\times 4},$$

the price coefficients are $\boldsymbol{\beta} = [-1.5 \ -1]^\intercal$, and the intercept parameters are $\boldsymbol{\gamma} = [20\ 12]^\intercal$.

For all the expanded dimensions, the means are 0.5 for the features in cluster 1 and -0.5 for the features in cluster 2, and the feature coefficients are all 0. Other model parameters stay unchanged. In this way, we expand the features homogeneously and keep the feature sparsity fixed for all problem settings.

F.3. Parameter Values for §6.1.3

The feature coefficients are given by

$$\boldsymbol{\alpha} = \begin{bmatrix} -0.2 & 0 & -0.8 & 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & \mathbf{0}_{10}^\intercal \\ 0 & -0.3 & 0 & 0.9 & -0.1 & -0.1 & -0.1 & -0.1 & -0.1 & -0.1 & \mathbf{0}_{10}^\intercal \end{bmatrix} \in \mathbb{R}^{2\times 20}.$$

Recall that $\bar{s}$ is defined as the upper bound on the *union* of the non-zeros indices over all clusters. Thus, we have $\bar{s} = 10$ in the base setting. We then create a sequence of problems by changing the feature coefficients on dimension 11 through dimension 16 to 0.1 (-0.1) for the first (second) cluster, one dimension at a time. The feature dimension $d = 20$ stays unchanged.

F.4. Parameter Values for §6.2

The time-heterogeneous features are 10-dimensional and randomly drawn from a mixture of three multivariate Gaussian distributions with the means

$$\boldsymbol{\mu}_1 = [\ \ 0.8 \quad 1.1 \ -0.6 \quad 1.2 \quad 1.3 \ -1 \quad\ 2 \quad\ \ 1.5 \quad 0.7 \quad\ 1.6\ ]^\intercal,$$
$$\boldsymbol{\mu}_2 = [\ \ 1.5 \quad\ 0 \quad\ -1.4 \quad 0.5 \ -0.5 \quad 0.2 \ 3 \quad\ -0.3 \ -1 \quad\ -0.2\ ]^\intercal,$$
$$\boldsymbol{\mu}_3 = [\ -0.2 \ -1 \quad\ \ 0.4 \ -0.3 \quad 0.5 \quad 1 \quad 0.9 \quad 0.6 \ -0.2 \quad 0.7\ ]^\intercal,$$

and the common covariance matrix $0.05\boldsymbol{I}_{10}$. The mixture weight for each of the three components is $1/3$. The user-heterogeneous features are 8-dimensional and randomly generated around two concentric 8-dimensional spheres of radii 1 and 3 centered at origin. The mixture weight for each of the two components is $1/2$. Because a feature vector is the concatenation of time- and user-heterogeneous features, there are a total of 6 underlying clusters. To summarize the consumption parameters of all underlying clusters, let

$$\begin{bmatrix} \boldsymbol{\theta}_1^\intercal \\ \vdots \\ \boldsymbol{\theta}_6^\intercal \end{bmatrix} = [\boldsymbol{\gamma}\ \boldsymbol{\beta}\ \boldsymbol{\alpha}], \text{ where } \boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_6 \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_6 \end{bmatrix}, \text{ and } \boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha}_1^\intercal \\ \vdots \\ \boldsymbol{\alpha}_6^\intercal \end{bmatrix}.$$

The price coefficients are given by $\boldsymbol{\beta} = [-1.5\ -1.2\ -0.8\ -1\ -0.4\ -1.8]^\intercal$, and the intercept parameters are given by $\boldsymbol{\gamma} = [20\ 12\ 9\ 14\ 8\ 15]^\intercal$. The feature coefficients are given by $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_{\text{time}}\ \boldsymbol{\alpha}_{\text{user}}]$, where

$$\boldsymbol{\alpha}_{\text{time}} = \begin{bmatrix} 0.8 & 0 & -0.2 & 0 & 0 & 0.4 & 0 & 0.6 & -1 & 0 \\ 0 & 0.3 & 0 & 0 & 0.5 & 0.9 & 0.6 & 0 & 0 & 0.1 \\ 0 & 1 & 0 & 0.8 & 0 & 0.5 & 0 & 0.2 & 0 & -0.3 \\ 0.2 & 0 & -0.4 & 0 & -1 & 0.3 & 0.1 & 0 & 0 & -0.1 \\ 0.6 & 0 & 0.1 & 1 & 0 & 0 & 0 & 0.5 & 0.3 & 0 \\ 0 & -0.2 & 0.1 & -0.4 & 0 & 0.4 & 0 & 0 & 0.5 & 0.2 \end{bmatrix} \in \mathbb{R}^{6\times 10},$$

$$\boldsymbol{\alpha}_{\text{user}} = \begin{bmatrix} 0.6 & 1.2 & 0 & 0 & 0.2 & -0.4 & 0 & 0.3 \\ 1 & 0 & -0.5 & 0 & 0.3 & 0 & 0.2 & -0.1 \\ 0.2 & -0.3 & 0.1 & 0 & 0 & -0.1 & 0.4 & 0.6 \\ 0 & 0 & -0.8 & 0.3 & 0 & -0.2 & 0.1 & 0 \\ 0.1 & 0 & 0 & -0.2 & 0 & 0.4 & -0.1 & 0.2 \\ 0 & -0.3 & 0 & 0.2 & 0.4 & 0 & -0.1 & 0.1 \end{bmatrix} \in \mathbb{R}^{6\times 8}.$$

Here, $\boldsymbol{\alpha}_{\text{time}}$ and $\boldsymbol{\alpha}_{\text{user}}$ contain the coefficients of time-heterogeneous and user-heterogeneous features, respectively. Each of the six rows corresponds to one of the six underlying clusters, and each column corresponds to a different feature. Note that the feature coefficient vector of each cluster contains about 8 zeros out of 18 entries. Thus, the model sparsity has a non-negligible effect on the estimation of the parameters.

The policy parameters are as follows. The set of feasible prices is $[2, 10]$. The prices in the experimentation periods are chosen from this region in the same way as in §6.1. The regularization parameters of our policy are $\lambda_{\text{time}} = 0.5\sqrt{M\log(d_1 M)}$ and $\lambda_{\text{user}} = 0.5\sqrt{N\log(d_2 N)}$ for the time-heterogeneous and user-heterogeneous features, respectively, where $M = \lceil \kappa\sqrt{T} \rceil$, $\kappa = 10$, $d_1 = 10$, and $d_2 = 8$.

F.5. Calibrated Parameter Values for §6.3.3

The model calibration results described in §6.3.3 are as follows: the calibrated coefficients of features for the two clusters are

$$\begin{aligned} \boldsymbol{\alpha}_1 = [&-0.04\ 0.02\ -0.72\ -0.01\ -0.05\ -0.04\ -1.81\ 0.53\ 1.38\ 0\ -1.01\ -0.95\ 7.59\ 8.60\ 10.70\ 0 \\ &-1.66\ -0.53\ -1.81\ 2.55\ -0.32\ -0.65\ 0\ 1.67\ -0.67\ -2.70\ 0\ -0.14\ 0.52\ 0\ 0.41\ -1.29 \\ &-0.04\ -1.08\ 6.51\ -0.28\ -2.14\ 0.67\ 2.02\ 1.79\ 0.11\ 0.12\ -0.60\ -0.45\ -1.13\ 5.64\ -3.66 \\ &-5.45\ 4.39\ 2.17\ 2.64\ 2.93\ -16.44\ 4.76\ -8.98\ -8.12\ -8.09\ -1.64\ 0.20\ 3.41\ 1.61\ 0\ 3.32 \\ &4.07\ 5.30\ 8.70\ -0.27\ -0.42\ -0.89\ -2.97\ 0\ -4.88\ 5.51\ -1.63\ 0.73\ 2.55\ 1.72\ -2.95 \\ &-2.56\ -2.34\ -0.87\ 1.09\ 0\ -1.87\ 1.36\ 1.07\ -2.45\ -3.20\ -1.59\ -0.94\ -0.59\ -0.79\ 0.84 \\ &-0.08\ 1.65\ 4.98\ 0\ 1.36\ -1.03\ 0\ 1.39\ 3.47\ 5.07\ -1.64\ -0.53\ 0\ 0.63\ -8.17\ -0.61\ 4.81\ 0 \\ &-5.86\ -9.83\ -4.09\ 2.02\ 5.72\ 6.60\ -3.43\ 5.62\ -5.85\ 0\ 8.56\ 2.61\ \mathbf{0}_2^\intercal\ 3.82\ -3.74\ 4.44 \\ &4.26\ -1.91\ 0\ 1.57\ 6.45\ 7.70\ \mathbf{0}_2^\intercal\ 2.26\ 0\ 1.02\ \mathbf{0}_{16}^\intercal\ 3.98\ 0]^\intercal, \end{aligned}$$

$$\boldsymbol{\alpha}_2 = [0\ 0.02\ 0.12\ 0.02\ -0.04\ -0.04\ -0.32\ 0.49\ 2.75\ 0.72\ -0.79\ -0.41\ 2.94\ 4.05\ 6.52\ 0\ -0.80$$
$$0.32\ 0.28\ 3.08\ -3.64\ 1.01\ -0.95\ -1.51\ -1.10\ 1.61\ -3.20\ -0.40\ -0.58\ -0.41\ 0.66$$
$$-0.35\ -0.20\ 0\ 3.63\ -2.20\ 0.50\ -0.21\ 1.72\ 1.53\ 0.93\ -0.42\ -2.50\ 0.77\ -2.39\ 2.64\ 3.69$$
$$0.62\ 8.62\ 7.31\ 6.88\ 7.07\ -9.42\ -3.11\ -5.08\ -4.24\ -3.12\ -1.92\ -0.35\ 2.84\ 2.24\ 0$$
$$-4.78\ -3.82\ -3.07\ -2.00\ -2.74\ -4.81\ -5.24\ -3.41\ \mathbf{0}_2^\intercal\ 3.64\ 1.41\ 2.61\ 2.22\ 2.80\ -1.14$$
$$-0.07\ 0\ -0.14\ 2.12\ 0.85\ 0.87\ -4.62\ -2.93\ 5.64\ 8.10\ 6.28\ 7.80\ -1.81\ -1.74\ 1.50\ 0$$
$$2.24\ 5.96\ 0\ 1.56\ -0.41\ -1.58\ -0.33\ 1.64\ 2.56\ -1.42\ -0.58\ -2.70\ -1.95\ -4.97\ 2.06$$
$$5.84\ -1.93\ 1.28\ -5.13\ -5.76\ -2.16\ 1.29\ 0\ -2.85\ -5.80\ 0\ 1.26\ 4.09\ 3.85\ 7.14\ 0.92$$
$$2.69\ -9.30\ 3.38\ 0.47\ -0.36\ 3.65\ 2.11\ 6.55\ \mathbf{0}_5^\intercal\ 2.19\ \mathbf{0}_2^\intercal\ 4.50\ \mathbf{0}_7^\intercal\ -3.83\ \mathbf{0}_7^\intercal]^\intercal.$$

The calibrated price-sensitivity coefficients and intercept parameters for the two clusters are $\beta_1 = -4.58$, $\beta_2 = -2.97$, $\gamma_1 = 42.70$, and $\gamma_2 = 42.58$.

Appendix G: Summary of Features in the Real-life Data Set

Table 3 Summary of Features in Section 6.3

Feature	Type	Description	Summary statistics
temperature	continuous	temperature (°F)	mean: 66.13, std: 12.62, range: $22.92 - 93.44$
dew_point	continuous	temperature (°F)	mean: 51.71, std: 13.45, range: $14.43 - 74.56$
humidity	continuous	relative humidity	mean: 64.21%, std: 21.61%, range: $11\% - 98\%$
visibility	continuous	visibility (miles)	mean: 9.12, std: 1.89, range: $0.3 - 10$
pressure	continuous	air pressure (mbar)	mean: 1015.0, std: 5.81, range: $996.7 - 1030.4$
wind_speed	continuous	wind speed (mph)	mean: 8.59, std: 4.21, range: $0 - 23.45$
precip_intensity	continuous	precipitation intensity (inch/min)	mean: 0.0025, std: 0.0234, range: $0 - 0.5683$
precip_prob	continuous	precipitation probability	mean: 0.0309, std: 0.1531, range: $0 - 1$
house_constr	categorical	house construction year	5 levels, from 1924 to 2014
square_foot	categorical	total square footage	5 levels, from 0 to 5600
foundation	categorical	foundation type	3 levels, pier beam, slab, or other
weekday_home	categorical	whether customers stay at home on any weekdays	5 levels, Monday through Friday (multiple choices)
education	categorical	highest level of education among all residents	4 levels, from vocational school to postgraduate degree
income	categorical	total annual income	8 levels, from 1,000 to 1,000,000
pv_size	continuous	size of the solar panel (kW)	mean: 3.19, std: 3.17, range: $0 - 10$
pv_satisfaction	categorical	how satisfied with the solar panel	5 levels, from very dissatisfied to very satisfied
pv_pos_factor	categorical	positive factors of solar panel	5 levels, independence from utility, emission-free, etc. (multiple choices)
pv_reason	continuous	reasons of having a solar panel	principal components of the matrix of word counts
retrofits	categorical	any changes to the house since 2012	3 levels: yes, no, N/A
irrigation	categorical	any irrigation system	3 levels: yes, no, N/A
ceil_fans	continuous	counts of ceiling fans	mean: 4.33, std: 2.10, range: $0 - 8$
hvac_type	categorical	type of HVAC system	8 levels: central air, two-way heat pump, etc.
temp_sum_weekday	categorical	thermostat temperature on summer weekdays	4 levels, from 70 °F to 90 °F
temp_sum_morning	categorical	thermostat temperature in summer mornings	4 levels, from 60 °F to 83 °F
temp_sum_evening	categorical	thermostat temperature in summer evenings	3 levels, from 65 °F to 79 °F

Feature	Type	Description	Summary statistics
temp_sum_sleeping	categorical	thermostat temperature in summer sleeping hours	3 levels, from 70 °F to 84 °F
temp_sum_weekend	categorical	thermostat temperature in summer weekends	3 levels, from 70 °F to 84 °F
temp_win_weekday	categorical	thermostat temperature on winter weekdays	5 levels, from 60 °F to 85 °F
temp_win_morning	categorical	thermostat temperature in winter mornings	3 levels, from 64 °F to 75 °F
temp_win_evening	categorical	thermostat temperature in winter evenings	2 levels, from 68 °F to 75 °F
temp_win_sleeping	categorical	thermostat temperature in winter sleeping hours	4 levels, from 60 °F to 75 °F
temp_win_weekend	categorical	thermostat temperature in winter weekends	3 levels, from 64 °F to 75 °F
thermo_prog	categorical	whether the thermostat is programmed	3 levels: yes, no, N/A
thermo_diff	categorical	how difficulty to program the thermostat	4 levels, from easy to very difficult
ac_package	categorical	whether the customer buys an annual AC package	3 levels: yes, no, N/A
cable_box	categorical	number of cable or satellite TV boxes	5 levels, from 0 through 5 or more, including N/A
dvr	categorical	number of DVRs	5 levels, from 0 through 5 or more, including N/A
wifi	categorical	number of WIFI routers	4 levels, from 0 through 5 or more, including N/A
game	categorical	number of game systems	5 levels, from 0 through 5 or more, including N/A

Note: All categorical variables are converted into indicator variables. The first 8 features are time-heterogeneous and the rest are user-heterogeneous.

Appendix H: Robustness Checks

We conduct a robustness check of the performance of our policy under a model as described in §6.3.3 except that the parameters are calibrated using forward addition. The following table shows the percentage increase in average regret if our policy is replaced by these alternative policies.

Table 4 Percentage Increase in Average Regret Relative to Joint Clustering and Feature-based Pricing

	$T = 1200$	$T = 1700$	$T = 2200$
No clustering	-2.47%	22.06%	74.79%
No feature-based pricing	61.93%	64.54%	110.59%
Greedy (no experimentation)	397.45%	459.19%	1516.43%
No regularization	37.05%	25.39%	42.70%
Company	76.07%	90.80%	162.10%

The following figures illustrate the cumulative profits of our policy versus the company's historical decisions. Our policy significantly increases the three-month total profits by at least 146% under the model calibrated using forward addition.

(a) cumulative profit over T periods ($) (b) three-month total profit ($)

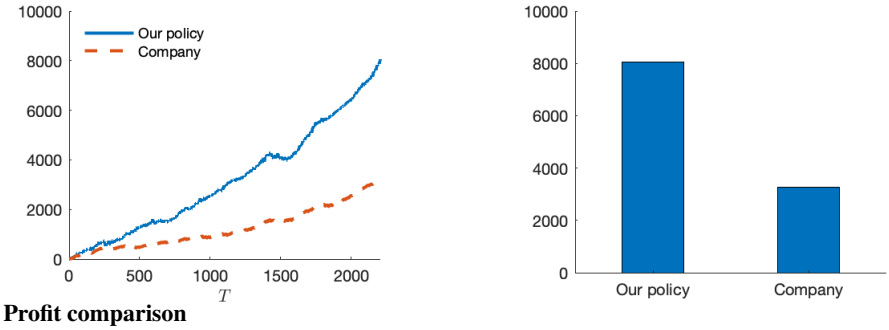


Figure 12 Profit comparison

Appendix I: Proofs of Results in §7.1

Proof of Proposition 5.

(a) We first consider the regret due to delayed true detection. For any fixed $j=1,\ldots,\mathscr{C}$, we assume that there exists at least one early false detection between the j^{th} and the $(j+1)^{\text{th}}$ true change-points, that is, $\tau_j^* < \tau_j^+$; otherwise, there is nothing to prove. By definition, $\hat{\tau}_j^+$ is the first cycle of claimed detection after τ_j^*, so there is no claimed detection between cycle $L(\tau_j^*)$ and cycle τ_j^*, where $L(\tau_j^*)$ is the latest detection cycle before τ_j^*. We first consider the event that at least one cycle k between $L(\tau_j^*)$ and τ_j^* contains no change-points and that the feature mean in cycle k is different from that after τ_j^*, that is,

$$\mathscr{B}_j = \bigcup_{k=L(\tau_j^*)}^{\tau_j^*} \left\{ \boldsymbol{\mu}_t = \boldsymbol{\mu}_{t'} \neq \boldsymbol{\mu}_{(\tau_j^*+1)\ell+1} \ \forall t,t' \in \{k\ell+1,\ldots,(k+1)\ell\} \right\}. \tag{I.1}$$

Note that the event $\{\hat{\tau}_j^+ - \tau_j^* \geq \tau\}$ implies the event $\{\chi_{\tau_j^*+\tau-1}=0\}$, that is, there is no claimed detection in cycle $\hat{k}=\tau_j^*+\tau-2$. Assume that $\hat{\tau}_j^+ - \tau_j^* \geq 3$; otherwise, the average regret between cycle τ_j^* and $\hat{\tau}_j^+$ is at most $2\ell R_{\text{rg}}/T$, which is in the order of $\log T/T$. Hence, for any $\hat{k}=\tau_j^*+1,\ldots,\tau_{j+1}^*-2$, we have

$$\begin{aligned}
\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\{\hat{\tau}_j^+ - \tau_j^* \geq \tau, \mathscr{B}_j\} &\leq \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\{\chi_{\hat{k}+1}=0, \mathscr{B}_j\} \\
&= \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{ \sup\left\{ \|\bar{\mathbf{X}}_{\hat{k}} - \bar{\mathbf{X}}_k\|_2 \colon k = L(\tau_j^*),\ldots,\hat{k}-1 \right\} \leq \eta, \mathscr{B}_j \right\} \\
&\leq \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{ \|\bar{\mathbf{X}}_{\hat{k}} - \bar{\mathbf{X}}_{\tilde{k}}\|_2 \leq \eta \text{ for some } \tilde{k} \in \{L(\tau_j^*),\ldots,\tau_j^*\}, \mathscr{B}_j \right\} \\
&\leq \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{ \frac{1}{\ell}\| \sum_{t=\ell\hat{k}+1}^{\ell(\hat{k}+1)} (\boldsymbol{\mu}_t + \boldsymbol{\xi}_t) - \sum_{t=\ell\tilde{k}+1}^{\ell(\tilde{k}+1)} (\boldsymbol{\mu}_t + \boldsymbol{\xi}_t)\|_2 \leq \eta, \right. \\
&\qquad\qquad \left. \text{for some } \tilde{k} \in \{L(\tau_j^*),\ldots,\tau_j^*\}, \mathscr{B}_j \right\} \\
&\leq \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{ \|\bar{\boldsymbol{\xi}}_{\hat{k}} - \bar{\boldsymbol{\xi}}_{\tilde{k}}\|_2 \geq \|\boldsymbol{\mu}_t - \boldsymbol{\mu}_{(\tau_j^*+1)\ell+1}\|_2 - \eta, \right. \\
&\qquad\qquad \left. \text{for some } \tilde{k} \in \{L(\tau_j^*),\ldots,\tau_j^*\}, t \in \{\ell\tilde{k}+1,\ldots,\ell(\tilde{k}+1)\} \right\}
\end{aligned} \tag{I.2}$$

By choosing $\eta = \delta_{\min}/2$, we further deduce that

$$\begin{aligned}\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\{\hat{\tau}_j^+ - \tau_j^* \geq \tau, \mathscr{B}_j\} &\leq \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\bar{\boldsymbol{\xi}}_{\hat{k}} - \bar{\boldsymbol{\xi}}_{\tilde{k}}\|_2 \geq \eta, \text{ for some } \tilde{k} \in \{L(\tau_j^*), \ldots, \tau_j^*\}\right\} \\ &\leq \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\bar{\boldsymbol{\xi}}_{\hat{k}}\|_2 \geq \frac{1}{2}\eta\right\} + \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\bar{\boldsymbol{\xi}}_{\tilde{k}}\|_2 \geq \frac{1}{2}\eta\right\} \\ &\leq 4T^{-5/2}\,. \end{aligned} \tag{I.3}$$

Therefore, the average regret for periods in the optimization phase between τ_j^* and $\hat{\tau}_j^+$ is

$$\begin{aligned}&\frac{1}{NT}\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\sum_{t=\ell\tau_j^*+1}^{\ell\hat{\tau}_j^+}\sum_{i=1}^{N}[\mathcal{R}^*(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t}) - \mathcal{R}(p^{\pi}_{i,t}; \boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t})]\mathbb{I}\{t \notin \mathscr{E}, \mathscr{B}_j\}\right\} \\ &\leq \frac{1}{NT}R_{\text{rg}}N\ell \cdot \mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}[(\hat{\tau}_j^+ - \tau_j^*)\mathbb{I}(\mathscr{B}_j)] \\ &\leq \frac{2\ell R_{\text{rg}}}{T} + \frac{R_{\text{rg}}}{T}\ell\sum_{\tau=3}^{\tau_{j+1}^*-\tau_j^*}\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\{\hat{\tau}_j^+ - \tau_j^*, \mathscr{B}_j\} \\ &\leq 4R_{\text{rg}}(\kappa_0 + \kappa_1 + 1)T^{-1}\log T\,. \end{aligned} \tag{I.4}$$

We now switch to the event $\mathscr{B}_j^c$, in which case there exists at least one change-point in every cycle between $L(\tau_j^*)$ and τ_j^*. By the second order Taylor expansion and the implicit function theorem, the average regret for periods in the optimization phase between τ_j^* and $\hat{\tau}_j^+$ can be expressed as

$$\begin{aligned}&\frac{1}{NT}\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\sum_{t=\ell\tau_j^*+1}^{\ell\hat{\tau}_j^+}\sum_{i=1}^{N}[\mathcal{R}^*(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t}) - \mathcal{R}(p^{\pi}_{i,t}; \boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t})]\mathbb{I}\{t \notin \mathscr{E}, \mathscr{B}_j^c\}\right\} \\ &\leq \frac{1}{NT}\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\sum_{t=\ell\tau_j^*+1}^{\ell\hat{\tau}_j^+}\sum_{i=1}^{N}\sum_{k=1}^{\bar{\mathcal{K}}}\varrho\bar{R}_2^2\|\hat{\boldsymbol{\theta}}_k(\boldsymbol{\lambda}_k) - \boldsymbol{\theta}_k\|_2^2\mathbb{I}\{t \notin \mathscr{E}, \mathscr{B}_j^c\}\right\}, \end{aligned} \tag{I.5}$$

where $\varrho = \sup\{|2\beta g'(\boldsymbol{\theta}^{\mathsf{T}}\mathbf{u}^*) + \beta^2\varphi(\boldsymbol{\theta}, \mathbf{x})g''(\boldsymbol{\theta}^{\mathsf{T}}\mathbf{u}^*)|^{-1}\colon \boldsymbol{\theta} \in \Theta, \mathbf{x} \in \mathcal{X}, p \in \mathcal{P}\}$ and $\bar{R}_2 = \sup\{\max\{|\frac{\partial^2\mathcal{R}}{\partial p^2}|, |\frac{\partial^2\mathcal{R}}{\partial p \partial \boldsymbol{\theta}}|\} : \boldsymbol{\theta} \in \Theta, \mathbf{x} \in \mathcal{X}, p \in \mathcal{P}\}$. For each estimated cluster $k_1 = 1, \ldots, \bar{\mathcal{K}}_1$, let

$$\tilde{\boldsymbol{\zeta}}_{k_1} = \sum_{t=\ell L(\tau_j^*)+1}^{\ell(\tau_j^*+1)}[g(\boldsymbol{\theta}_{\mathcal{C}_1(\mathbf{X}_t)}^{\mathsf{T}}\mathbf{U}_t) - g(\boldsymbol{\theta}_{k_1}^{\mathsf{T}}\mathbf{U}_t)]\mathbf{U}_t\mathbb{I}\{t \in \mathscr{E}, \mathcal{C}_1(\mathbf{X}_t) \neq k_1\}\,, \tag{I.6}$$

which relates to the quantity that corrects the effect due to misclassification from cycle $L(\tau_j^*)$ through τ_j^*, where $\mathbf{U}_t^{\mathsf{T}} = [1\ \mathbf{X}_t^{\mathsf{T}}\ p_t]$. Since there are at most $\mathscr{C}$ change-points and each cycle from $L(\tau_j^*)$ through τ_j^* contains at least one change-point on $\mathscr{B}_j^c$, it follows that $||\tilde{\boldsymbol{\zeta}}_{k_1}||_2 \leq 2\mathscr{C}\ell\bar{\nu} \leq 4\mathscr{C}(\kappa_0 + \kappa_1)\bar{\nu}\log T$, where $\bar{\nu} = \max\{g(\boldsymbol{\vartheta}^{\mathsf{T}}\mathbf{u}) \cdot \|\mathbf{u}\|_2\colon \mathbf{u}^{\mathsf{T}} = [1\ \mathbf{x}^{\mathsf{T}}\ p], \boldsymbol{\vartheta} \in \Theta, \mathbf{x} \in \mathcal{X}, p \in \mathcal{P}\}$. (This adds a term in the order of $\log T$ to the misclassification error of spectral clustering using $M = \lceil\kappa\sqrt{T}\rceil$ periods of time-heterogeneous features, which is in the order of $\sqrt{M\log M}$, dominating the added term.) The rest then

follows a similar argument as in the proof of Proposition 3 and Theorem 2. Combining the results on the two events $\mathscr{B}_j$ and $\mathscr{B}_j^c$ completes the proof.

(b) We now consider the regret due to early false alarm. Fix $j = 0, \ldots, \mathscr{C}$. Again, we assume that there exists at least one early false detection between the j^{th} and the $(j+1)^{\text{th}}$ true change-points, that is, $\tau_j^* < \tau_j^+ < \tau_j^- < \tau_{j+1}^*$; otherwise, there is nothing to prove. This implies that $\tau_{j+1}^* - \hat{\tau}_j^- \le \tau_{j+1}^* - \tau_j^* - 2$. Note that the event $\{\tau_{j+1}^* - \hat{\tau}_j^- = \tau\}$ implies that there is a claimed detection in cycle $\hat{k} = \tau_{j+1}^* - \tau - 1$ and the latest detection cycle before cycle $\hat{k}$ is no earlier than cycle $\tau_j^* + 1$, that is, $\{\chi_{\hat{k}+1} = 1\} \cap \{L(\hat{k}) \ge \tau_j^* + 1\}$, we then have

$$\begin{aligned}
\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\{\tau_{j+1}^* - \hat{\tau}_j^- = \tau\} &\le \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\{\chi_{\hat{k}+1} = 1, L(\hat{k}) \ge \tau_j^* + 1\} \\
&\le \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\sup\left\{\|\bar{\mathbf{X}}_{\hat{k}} - \bar{\mathbf{X}}_k\|_2 \colon k = \tau_j^* + 1, \ldots, \hat{k} - 1\right\} > \eta\right\} \\
&\le \sum_{k=\tau_j^*+1}^{\hat{k}-1} \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\bar{\mathbf{X}}_{\hat{k}} - \bar{\mathbf{X}}_k\|_2 > \eta\right\}. \qquad \text{(I.7)}
\end{aligned}$$

Since there are no change-points between cycle $\tau_j^* + 1$ and $\hat{k}$, we further deduce that

$$\begin{aligned}
\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\bar{\mathbf{X}}_{\hat{k}} - \bar{\mathbf{X}}_k\|_2 > \eta\right\} &= \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\frac{1}{\ell}\sum_{t=\ell\hat{k}+1}^{\ell(\hat{k}+1)}(\boldsymbol{\mu}_t + \boldsymbol{\xi}_t) - \frac{1}{\ell}\sum_{t=\ell k+1}^{\ell(k+1)}(\boldsymbol{\mu}_t + \boldsymbol{\xi}_t)\|_2 > \eta\right\} \\
&= \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\bar{\boldsymbol{\xi}}_{\hat{k}} - \bar{\boldsymbol{\xi}}_k\|_2 > \eta\right\}, \qquad \text{(I.8)}
\end{aligned}$$

for $k = \tau_j^* + 1, \ldots, \hat{k} - 1$. We now state a useful lemma and prove it at the end of this subsection.

LEMMA 4. *For any cycle k, we have*

$$\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\bar{\boldsymbol{\xi}}_k\|_2 \ge \frac{1}{2}\eta\right\} \le 2T^{-5/2}. \qquad \text{(I.9)}$$

By Lemma 4, it follows that

$$\begin{aligned}
\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\{\tau_{j+1}^* - \hat{\tau}_j^- = \tau\} &\le \sum_{k=\tau_j^*+1}^{\hat{k}-1}\left(\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\bar{\boldsymbol{\xi}}_{\hat{k}}\|_2 > \frac{1}{2}\eta\right\} + \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\|\bar{\boldsymbol{\xi}}_k\|_2 > \frac{1}{2}\eta\right\}\right) \\
&\le T/\ell \cdot 4T^{-5/2} \le 4T^{-3/2}. \qquad \text{(I.10)}
\end{aligned}$$

Therefore, by the tail sum for expectation, we can express the regret in the optimization phase between cycle $\hat{\tau}_j^-$ and τ_{j+1}^* as

$$\begin{aligned}
&\frac{1}{NT}\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\left\{\sum_{t=\ell\hat{\tau}_j^-+1}^{\ell\tau_{j+1}^*}\sum_{i=1}^{N}[\mathcal{R}^*(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t}) - \mathcal{R}(p_{i,t}^{\pi}; \boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t})]\mathbb{I}\{t \notin \mathscr{E}\}\right\} \\
&\le \frac{1}{NT}R_{\text{rg}}\ell N \cdot \mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}[\tau_{j+1}^* - \hat{\tau}_j^-]
\end{aligned}$$

$$= \frac{1}{T} R_{\text{rg}} \ell \sum_{m=1}^{\tau^*_{j+1}-\tau^*_j-2} \sum_{\ell=m}^{\tau^*_{j+1}-\tau^*_j-2} \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\{\tau^*_{j+1} - \hat{\tau}^-_j = \ell\}$$
$$\leq \frac{1}{T} R_{\text{rg}} \ell (T/\ell)^2 \cdot 4T^{-3/2} \leq 4R_{\text{rg}} T^{-1/2}\,. \tag{I.11}$$

Choosing $K_{10} = 4R_{\text{rg}}$ completes the proof.

Q.E.D.

Proof of Theorem 3. Finally, since the environment is stationary and there is no claimed detection between cycle $\hat{\tau}^+_j$ and $\hat{\tau}^-_j$, the upper bound on the average regret is provided in Theorem 2. In sum, the overall average regret in the presence of change-points in the time-heterogeneous features is

$$\begin{aligned}
\Delta^{\pi}_{\boldsymbol{\theta}}(N,T) &= \frac{1}{NT}\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}} \left\{ \sum_{t=1}^{T}\sum_{i=1}^{N} [\mathcal{R}^*(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t}) - \mathcal{R}(p^{\pi}_{i,t}; \boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t})] \mathbb{I}\{t \in \mathscr{E}\} \right\} \\
&\quad + \sum_{j=0}^{\mathscr{C}} \left(\sum_{t=\ell\tau^*_j+1}^{\ell\hat{\tau}^+_j} + \sum_{t=\ell\hat{\tau}^+_j+1}^{\ell\hat{\tau}^-_j} + \sum_{t=\ell\hat{\tau}^-_j+1}^{\ell\tau^*_{j+1}} \right) \sum_{i=1}^{N} [\mathcal{R}^*(\boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t}) - \mathcal{R}(p^{\pi}_{i,t}; \boldsymbol{\theta}_{\mathcal{C}(\mathbf{X}_{i,t})}, \mathbf{X}_{i,t})] \mathbb{I}\{t \notin \mathscr{E}\} \\
&\leq \frac{M}{T} R_{\text{rg}} + 3R_{\text{rg}} \bar{\mathcal{K}} (K_1 + K_2) \sqrt{\frac{\log(NT)}{N \wedge T}} \\
&\quad + (\mathscr{C}+1) \left(\frac{K_9 \bar{\mathcal{K}} \bar{s} \log(dNT)}{\sqrt{N \wedge T}} + \frac{K_8 \bar{\mathcal{K}} \bar{s} \log(dNT)}{\sqrt{N \wedge T}} + K_{10} T^{-1/2} \right) \\
&\leq \frac{K_{11} \bar{\mathcal{K}} \bar{s} \log(dNT)}{\sqrt{N \wedge T}}\,,
\end{aligned} \tag{I.12}$$

where $K_{11} = R_{\text{rg}}(2\kappa + 3K_1 + 3K_2) + (\mathscr{C}+1)(K_8 + K_9 + K_{10})$. Q.E.D.

Proof of Lemma 4. Let $\mathbf{S}_t = \sum_{s=1}^{t} \boldsymbol{\xi}_s$. For any scalar ρ satisfying $|\rho| = \eta/2$ and any $t = 1, \ldots, T, j = 1, \ldots, d$, define

$$Z_{t,j} = \exp\left\{ \nu \left(\rho S_{t,j} - \frac{1}{2}\rho^2 \frac{t}{\sqrt{d}} \right) \right\} \tag{I.13}$$

with $Z_{0,j} = 1$, where $\nu = (2\lambda_0/\eta) \wedge (v_0 \sigma^2 \sqrt{d})^{-1}$. Clearly, $Z_{t,j}$ is integrable for all $t = 1, \ldots, T, j = 1, \ldots, d$. Let $\mathcal{F}_{t,j} = \sigma(\boldsymbol{\xi}_{1,j}, \ldots, \boldsymbol{\xi}_{t,j})$. Then,

$$\begin{aligned}
\mathbb{E}[Z_{t,j} | \mathcal{F}_{t-1,j}] &= \exp\left\{ \nu \left(\rho S_{t-1,j} - \frac{1}{2}\rho^2 \frac{t}{\sqrt{d}} \right) \right\} \mathbb{E}\left[\exp\left(\nu\rho\xi_{t,j} \right) | \, \mathcal{F}_{t-1,j} \right] \\
&\leq \exp\left\{ \nu \left(\rho S_{t-1,j} - \frac{1}{2}\rho^2 \frac{t-1}{\sqrt{d}} \right) \right\} \exp\left(-\frac{\nu}{2}\rho^2 \frac{1}{\sqrt{d}} \right) \exp\left(\frac{1}{2} v_0 \sigma^2 \rho^2 \nu^2 \right) \\
&\leq Z_{t-1}\,,
\end{aligned} \tag{I.14}$$

where the first inequality follows because $|\rho\nu| = \eta\nu/2 \leq \lambda_0$ and the second inequality follows because $\nu \geq (v_0 \sigma^2 \sqrt{d})^{-1}$. Thus, $(Z_{t,j}, \mathcal{F}_{t,j})$ is a supermartingale. For any cycle k, we have

$$\mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}} \left\{ \|\bar{\boldsymbol{\xi}}_k\|_2 \geq \frac{1}{2}\eta \right\} = \mathbb{P}^{\pi}_{\mathbf{X},\boldsymbol{\theta}} \left\{ \|\mathbf{S}_\ell\|_2 \geq \frac{1}{2}\eta\ell \right\}$$

$$
\begin{aligned}
&= \mathbb{P}^{\pi}_{\mathbf{x},\boldsymbol{\theta}} \left\{ \sum_{j=1}^{d} S_{\ell,j}^2 \geq \frac{1}{4}\eta^2\ell^2 \right\} \\
&\leq \mathbb{P}^{\pi}_{\mathbf{x},\boldsymbol{\theta}} \left\{ \bigcup_{j=1}^{d} \left\{ |S_{\ell,j}| \geq \frac{1}{2\sqrt{d}}\eta\ell \right\} \right\} \\
&\leq \sum_{j=1}^{d} \mathbb{P}^{\pi}_{\mathbf{x},\boldsymbol{\theta}} \left\{ |S_{\ell,j}| \geq \frac{1}{2\sqrt{d}}\eta\ell \right\} . \qquad \text{(I.15)}
\end{aligned}
$$

Choosing $\rho = \eta/2$, it follows that

$$
\begin{aligned}
\mathbb{P}^{\pi}_{\mathbf{x},\boldsymbol{\theta}} \left\{ S_{\ell,j} \geq \frac{1}{2\sqrt{d}}\eta\ell \right\} &= \mathbb{P}^{\pi}_{\mathbf{x},\boldsymbol{\theta}} \left\{ \nu\rho S_{\ell,j} - \frac{1}{2}\nu\rho^2\frac{\ell}{\sqrt{d}} \geq \frac{1}{2}\nu\rho^2\frac{\ell}{\sqrt{d}} \right\} \\
&= \mathbb{P}^{\pi}_{\mathbf{x},\boldsymbol{\theta}} \left\{ Z_{\ell,j} \geq \exp\left(\frac{1}{2}\nu\rho^2\frac{\ell}{\sqrt{d}}\right) \right\} \\
&\leq \mathbb{E}[Z_{\ell,j}] \exp\left(-\frac{1}{2}\nu\rho^2\frac{\ell}{\sqrt{d}}\right) \\
&\leq \exp\left(-\frac{1}{8}\nu\eta^2\frac{\ell}{\sqrt{d}}\right) . \qquad \text{(I.16)}
\end{aligned}
$$

Choosing $\rho = -\eta/2$ and repeating the above arguments yields

$$
\mathbb{P}^{\pi}_{\mathbf{x},\boldsymbol{\theta}} \left\{ S_{\ell,j} \leq -\frac{1}{2\sqrt{d}}\eta\ell \right\} \leq \exp\left(-\frac{1}{8}\nu\eta^2\frac{\ell}{\sqrt{d}}\right) . \qquad \text{(I.17)}
$$

By choosing $\ell = \lceil \kappa_0 + \kappa_1 \log T \rceil$ with $\kappa_0 = (8\sqrt{d}\log d)/(\nu\eta^2)$ and $\kappa_1 = 20\sqrt{d}/(\nu\eta^2)$, we further have

$$
\mathbb{P}^{\pi}_{\mathbf{x},\boldsymbol{\theta}} \left\{ \|\bar{\boldsymbol{\xi}}_k\|_2 \geq \frac{1}{2}\eta \right\} \leq 2d \exp\left(-\frac{1}{8}\nu\eta^2\frac{\ell}{\sqrt{d}}\right) \leq 2T^{-5/2} . \qquad \text{(I.18)}
$$

Q.E.D.

Appendix J: Proofs of Results in §7.2

Proof of Proposition 6. We first write the single-period profit function when the cost function is piecewise linear with a single kink. Given $\boldsymbol{\theta}, \boldsymbol{\phi}, \kappa, \mathbf{x}_t$, the total profit in period t corresponding to the vector of expected consumptions $\tilde{\mathbf{d}}_t$ is given by

$$
\begin{aligned}
\mathscr{R}(\tilde{\mathbf{d}}_t; \boldsymbol{\theta}, \boldsymbol{\phi}, \varkappa, \mathbf{x}_t) = &\sum_{i=1}^{N} [g^{-1}(\tilde{d}_{i,t}) - \gamma_{\mathcal{C}(\mathbf{x}_{i,t})} - \boldsymbol{\alpha}_{\mathcal{C}(\mathbf{x}_{i,t})}^{\intercal}\mathbf{x}_{i,t}]\beta_{\mathcal{C}(\mathbf{x}_{i,t})}^{-1}\tilde{d}_{i,t} \\
&- \mathbb{E}_{\boldsymbol{\phi}} \left[\mathbb{I}\left\{ \sum_{i=1}^{N} (\tilde{d}_{i,t} + \varepsilon_{i,t}) < \tilde{q} \right\} \left(a_1 \sum_{i=1}^{N} (\tilde{d}_{i,t} + \varepsilon_{i,t}) + b_1 \right) \right] \qquad \text{(J.1)} \\
&- \mathbb{E}_{\boldsymbol{\phi}} \left[\mathbb{I}\left\{ \sum_{i=1}^{N} (\tilde{d}_{i,t} + \varepsilon_{i,t}) \geq \tilde{q} \right\} \left(a_2 \sum_{i=1}^{N} (\tilde{d}_{i,t} + \varepsilon_{i,t}) + b_2 \right) \right] ,
\end{aligned}
$$

where $\mathbb{E}_{\boldsymbol{\phi}}$ is the expectation taken with respect to $\{\sum_{i=1}^{N}\varepsilon_{i,t}: t=1,\ldots,T\}$. For each period t, treating $\tilde{\mathbf{d}}_t$ as the decision variable, the first order condition of profit maximization is given by

$$0=\beta_{\mathcal{C}(\mathbf{x}_{j,t})}^{-1}[(g^{-1})'(\tilde{d}_{j,t})\cdot\tilde{d}_{j,t}+g^{-1}(\tilde{d}_{j,t})-\gamma_{\mathcal{C}(\mathbf{x}_{j,t})}-\boldsymbol{\alpha}_{\mathcal{C}(\mathbf{x}_{j,t})}^{\intercal}\mathbf{x}_{j,t}]-a_2-(a_1-a_2)\cdot F\left(\tilde{q}-\sum_{i=1}^{N}\tilde{d}_{i,t};\boldsymbol{\phi}\right), \tag{J.2}$$

for all $j=1,\ldots,N$, where we use the fact that $a_1\tilde{q}+b_1=a_2\tilde{q}+b_2$. Note that the first-order condition constitutes a system of nonlinear equations. When $g(\cdot)$ takes form of the identity function, the first-order condition forms a simplified system of non-linear equations, which is given by

$$\beta_{\mathcal{C}(\mathbf{x}_{j,t})}^{-1}(2\tilde{d}_{j,t}-\gamma_{\mathcal{C}(\mathbf{x}_{j,t})}-\boldsymbol{\alpha}_{\mathcal{C}(\mathbf{x}_{j,t})}^{\intercal}\mathbf{x}_{j,t})-a_2-(a_1-a_2)\cdot F\left(\tilde{q}-\sum_{i=1}^{N}\tilde{d}_{i,t};\boldsymbol{\phi}\right)=0\,,\quad\forall j=1,\ldots,N. \tag{J.3}$$

Dividing $2\beta_{\mathcal{C}(\mathbf{x}_{j,t})}^{-1}$ on both sides of (J.3) and summing all the N equations, we get

$$\sum_{i=1}^{N}\tilde{d}_{i,t}-c_{1,t}-c_{2,t}F\left(\tilde{q}-\sum_{i=1}^{N}\tilde{d}_{i,t};\boldsymbol{\phi}\right)=0\,, \tag{J.4}$$

where $c_{1,t}=\frac{1}{2}\sum_{i=1}^{N}(\gamma_{\mathcal{C}(\mathbf{x}_{j,t})}+\boldsymbol{\alpha}_{\mathcal{C}(\mathbf{x}_{j,t})}^{\intercal}\mathbf{x}_{j,t}+\beta_{\mathcal{C}(\mathbf{x}_{i,t})}a_2)$ and $c_{2,t}=\frac{1}{2}(a_1-a_2)\sum_{i=1}^{N}\beta_{\mathcal{C}(\mathbf{x}_{i,t})}$. Since $F(-\infty;\boldsymbol{\phi})=0$ and $F(+\infty;\boldsymbol{\phi})=1$, the left hand side of (J.4) goes to $+\infty$ (resp. $-\infty$) as $\sum_{i=1}^{N}\tilde{d}_{i,t}$ tends to $+\infty$ (resp. $-\infty$). Hence, (J.4) has a solution. Plugging the solution back to the N equations, we get the solutions for each $\tilde{d}_{j,t}$ by solving the linear system of equations.

Moreover, the second-order derivatives of $\mathscr{R}$ with respect to $\tilde{\mathbf{d}}_t$ are given by

$$\frac{\partial^2\mathscr{R}}{\partial\tilde{d}_{j,t}\partial\tilde{d}_{k,t}}(\tilde{\mathbf{d}}_t;\boldsymbol{\theta},\boldsymbol{\phi},\varkappa,\mathbf{x}_t)=2\beta_{\mathcal{C}(\mathbf{x}_{j,t})}^{-1}\mathbb{I}\{j=k\}+(a_1-a_2)\cdot f\left(\tilde{q}-\sum_{i=1}^{N}\tilde{d}_{i,t};\boldsymbol{\phi}\right),\quad\forall j,k=1,\ldots,N\,. \tag{J.5}$$

These are exactly the entries of the Hessian matrix $\mathscr{H}$ of $\mathscr{R}$ at $\tilde{\mathbf{d}}_t$. Recall that $\mathscr{H}$ is negative definite if and only if $(-1)^j\det(\mathscr{H}_j)>0$ for all $j=1,\ldots,N$, where $\mathscr{H}_j$ is the sub-matrix of $\mathscr{H}$ formed by the first j rows and j columns. It can be shown by induction that

$$\det(\mathscr{H}_j)=2^{j-1}\prod_{i=1}^{j}\beta_{\mathcal{C}(\mathbf{x}_{i,t})}^{-1}\left[(a_1-a_2)\cdot f\left(\tilde{q}-\sum_{i=1}^{N}\tilde{d}_{i,t};\boldsymbol{\phi}\right)\sum_{i=1}^{j}\beta_{\mathcal{C}(\mathbf{x}_{i,t})}+2\right]. \tag{J.6}$$

Since $\beta_{\mathcal{C}(\mathbf{x}_{i,t})}<0$ for all i,t and $(a_1-a_2)\bar{f}N|\beta_{\min}|<2$, where $\bar{f}=\sup\{f(e;\boldsymbol{\phi}):e\in\mathbb{R},\boldsymbol{\phi}\in\Phi\}$, we deduce that $(-1)^j\det(\mathscr{H}_j)>0$ for all $j=1,\ldots,N$ and thus $\mathscr{H}$ is negative definite. This shows that the unconstrained system of non-linear equations (J.3) has a unique solution that maximizes (J.1). Q.E.D.

Proof of Theorem 4. For any $t=M+1,\ldots,T$ and any $i=1,\ldots,N$, consider the event $\{\hat{\mathcal{C}}(\mathbf{X}_{i,t})=\mathcal{C}(\mathbf{X}_{i,t})\}$. By the implicit function theorem, there exists a unique continuously differentiable function $G(\cdot)$ defined on an open neighborhood U of $\boldsymbol{\theta}_t,\boldsymbol{\phi},\varkappa,\mathbf{x}_t$ such that $\tilde{\mathbf{d}}_t^*=G(\boldsymbol{\theta}_t,\boldsymbol{\phi},\varkappa,\mathbf{x}_t)$ and

$$\nabla G_{\boldsymbol{\theta}}(\boldsymbol{\theta}_t,\boldsymbol{\phi},\varkappa,\mathbf{x}_t)=-[\mathscr{H}(\tilde{\mathbf{d}}_t^*;\boldsymbol{\theta}_t,\boldsymbol{\phi},\varkappa,\mathbf{x}_t)]^{-1}\frac{\partial^2\mathscr{R}}{\partial\tilde{\mathbf{d}}\partial\boldsymbol{\theta}^{\intercal}}(\tilde{\mathbf{d}}_t^*;\boldsymbol{\theta}_t,\boldsymbol{\phi},\varkappa,\mathbf{x}_t) \tag{J.7}$$

$$\nabla G_{\boldsymbol{\phi}}(\boldsymbol{\theta}_t, \boldsymbol{\phi}, \varkappa, \mathbf{x}_t) = -[\mathscr{H}(\tilde{\mathbf{d}}_t^*; \boldsymbol{\theta}_t, \boldsymbol{\phi}, \varkappa, \mathbf{x}_t)]^{-1} \frac{\partial^2 \mathscr{R}}{\partial \tilde{\mathbf{d}} \partial \boldsymbol{\phi}^\intercal}(\tilde{\mathbf{d}}_t^*; \boldsymbol{\theta}_t, \boldsymbol{\phi}, \varkappa, \mathbf{x}_t)\,, \tag{J.8}$$

where $\nabla G_{\boldsymbol{\theta}}$ and $\nabla G_{\boldsymbol{\phi}}$ are the gradient of G with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, respectively.

Fix a period t and all the parameters $\boldsymbol{\theta}_t, \boldsymbol{\phi}, \varkappa, \mathbf{x}_t$. By the multivariate Taylor expansion, we have that

$$\mathscr{R}(\tilde{\mathbf{d}}_t^\pi) = \mathscr{R}(\tilde{\mathbf{d}}_t^*) + \nabla \mathscr{R}(\tilde{\mathbf{d}}_t^*)^\intercal (\tilde{\mathbf{d}}_t^\pi - \tilde{\mathbf{d}}_t^*) + \frac{1}{2}(\tilde{\mathbf{d}}_t^\pi - \tilde{\mathbf{d}}_t^*)^\intercal \mathscr{H}(\bar{\mathbf{d}}_t)(\tilde{\mathbf{d}}_t^\pi - \tilde{\mathbf{d}}_t^*) \tag{J.9}$$

for some $\bar{\mathbf{d}}_t$ that lies in the line segment connecting $\tilde{\mathbf{d}}_t^\pi$ and $\tilde{\mathbf{d}}_t^*$. Since $\nabla \mathscr{R}(\tilde{\mathbf{d}}_t^*) = \mathbf{0}$ and $\mathscr{H}$ is negative definite, we further have

$$\mathscr{R}(\tilde{\mathbf{d}}_t^*) - \mathscr{R}(\tilde{\mathbf{d}}_t^\pi) \le \bar{\mathscr{R}}'' \|\tilde{\mathbf{d}}_t^\pi - \tilde{\mathbf{d}}_t^*\|_2^2, \tag{J.10}$$

where $\bar{\mathscr{R}}'' = -\frac{1}{2}\mathscr{H}_{\min}$ and $\mathscr{H}_{\min} = \min\{\mu_{\min}(\mathscr{H}(\tilde{\mathbf{d}}; \boldsymbol{\theta}, \boldsymbol{\phi}, \varkappa, \mathbf{x})) : \tilde{\mathbf{d}} \in \mathscr{D}^N, \boldsymbol{\theta} \in \Theta, \boldsymbol{\phi} \in \Phi, \varkappa \in \mathscr{K}, \mathbf{x} \in \mathcal{X}^N\}$, because the minimum eigenvalue of a matrix depending on certain parameters is a continuous function in these parameters.

Let the estimates of $\boldsymbol{\theta}$ be $\hat{\boldsymbol{\theta}}$. For any $t = 1, \ldots, M$ and any $i = 1, \ldots, N$, we have $\hat{\varepsilon}_{i,t} = D_{i,t} - \hat{\boldsymbol{\theta}}_{\mathcal{C}(\mathbf{x}_{i,t})}^\intercal \mathbf{u}_{i,t}$, where $\hat{\boldsymbol{\theta}}_{\mathcal{C}(\mathbf{x}_{i,t})}^\intercal = [\hat{\gamma}_{\mathcal{C}(\mathbf{x}_{i,t})}\ \hat{\beta}_{\mathcal{C}(\mathbf{x}_{i,t})}\ \hat{\boldsymbol{\alpha}}_{\mathcal{C}(\mathbf{x}_{i,t})}^\intercal]$ and $\mathbf{u}_{i,t}^\intercal = [1\ p_{i,t}\ \mathbf{x}_{i,t}^\intercal]$. According to the demand function, we immediately have that

$$\left| \sum_{i=1}^N \hat{\varepsilon}_{i,t} - \sum_{i=1}^N \varepsilon_{i,t} \right| \le R \sum_{i=1}^N \|\hat{\boldsymbol{\theta}}_{\mathcal{C}(\mathbf{x}_{i,t})} - \boldsymbol{\theta}_{\mathcal{C}(\mathbf{x}_{i,t})}\|_2\,, \tag{J.11}$$

where $R = \sqrt{1 + p_{\max}^2 + x_{\max}^2}$ and $x_{\max} = \max\{\|\mathbf{x}\|_2 \colon \mathbf{x} \in \mathcal{X}\}$. Denote by $\check{\boldsymbol{\phi}}$ the maximum likelihood estimator of $\boldsymbol{\phi}$ based on $\{\sum_{i=1}^N \hat{\varepsilon}_{i,t} : t = 1, \ldots, M\}$. By differentiating the log-likelihood $\sum_{t=1}^M \log f(\sum_{i=1}^N \hat{\varepsilon}_{i,t}; \boldsymbol{\phi})$ with respect to $\boldsymbol{\phi}$ and setting it to $\mathbf{0}$, we have that

$$\nabla A(\check{\boldsymbol{\phi}}) = \frac{1}{M} \sum_{t=1}^M \mathbf{T}\left(\sum_{i=1}^N \hat{\varepsilon}_{i,t} \right). \tag{J.12}$$

Since the density $f(e; \boldsymbol{\phi}) = B(e) \exp[\boldsymbol{\phi}^\intercal \mathbf{T}(e) - A(\boldsymbol{\phi})]$ integrates to 1 for any $\boldsymbol{\phi} \in \{\boldsymbol{\phi} : A(\boldsymbol{\phi}) < \infty\}$, we have

$$A(\boldsymbol{\phi}) = \log\left(\int_{-\infty}^{\infty} B(e) \exp[\boldsymbol{\phi}^\intercal \mathbf{T}(e)]\, \mathrm{d}e \right). \tag{J.13}$$

Taking derivative with respect to $\boldsymbol{\phi}$, we obtain

$$\begin{aligned} \nabla A(\boldsymbol{\phi}) &= e^{-A(\boldsymbol{\phi})} \frac{\partial}{\partial \boldsymbol{\phi}} \left(\int_{-\infty}^{\infty} B(e) \exp[\boldsymbol{\phi}^\intercal \mathbf{T}(e)]\, \mathrm{d}e \right) \\ &= \int_{-\infty}^{\infty} B(e) e^{-A(\boldsymbol{\phi})} \exp[\boldsymbol{\phi}^\intercal \mathbf{T}(e)] \mathbf{T}(e)\, \mathrm{d}e \\ &= \int_{-\infty}^{\infty} f(e; \boldsymbol{\phi}) \mathbf{T}(e)\, \mathrm{d}e = \mathbb{E}_{\boldsymbol{\phi}}[\mathbf{T}(e)]. \end{aligned} \tag{J.14}$$

Taking another derivative of $\nabla A(\boldsymbol{\phi})$ with respect to $\boldsymbol{\phi}$, the Hessian matrix of $A(\boldsymbol{\phi})$ is

$$\boldsymbol{H}_A(\boldsymbol{\phi}) = \int_{-\infty}^{\infty} f(e; \boldsymbol{\phi}) \mathbf{T}(e) \frac{\partial}{\partial \boldsymbol{\phi}^\intercal} [\boldsymbol{\phi}^\intercal \mathbf{T}(e) - A(\boldsymbol{\phi})]\, \mathrm{d}e$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} f(e;\boldsymbol{\phi})\left[\mathbf{T}(e)-\nabla A(\boldsymbol{\phi})\right]\left[\mathbf{T}(e)-\nabla A(\boldsymbol{\phi})\right]^{\mathsf{T}} \mathrm{d}e \\
&\quad + \nabla A(\boldsymbol{\phi}) \int_{-\infty}^{\infty} f(e;\boldsymbol{\phi})\left[\mathbf{T}(e)-\nabla A(\boldsymbol{\phi})\right]^{\mathsf{T}} \mathrm{d}e \\
&= \mathrm{Var}_{\boldsymbol{\phi}}[\mathbf{T}(e)]. \qquad \text{(J.15)}
\end{aligned}
$$

The last equality follows by (J.14) and the passage of differentiation under the integral sign follows from the dominated convergence theorem, see e.g. Theorem 2.4 in Keener (2010). Fix a $\boldsymbol{\phi} \in \{\phi : A(\boldsymbol{\phi}) < \infty\}$ and a $\mathbf{v} \in \mathbb{R}^{\ell}$, where ℓ is the dimension of $\boldsymbol{\phi}$. For any set of i.i.d. observations $e_1, \ldots, e_n$, $\mathbf{v}^{\mathsf{T}} \boldsymbol{H}_A(\boldsymbol{\phi})\mathbf{v} = 0$ implies that $w(e_i) := v^{\mathsf{T}}[\mathbf{T}(e_i) - \nabla A(\boldsymbol{\phi})] = 0$ for all $i = 1, \ldots, n$. For the exponential family of minimal representation, the components of the sufficient statistics $\mathbf{T}(\cdot)$ are not redundant, that is, the set of vectors $\{\mathbf{T}(e_i) : i = 1, \ldots, n\}$ spans $\mathbb{R}^{\ell}$. (One can always eliminate redundant components of sufficient statistics, if any, to obtain a minimal representation.) Thus, there exists $\mathbf{a} = (a_1, \ldots, a_{\ell})^{\mathsf{T}}$ such that $\mathbf{v} = \sum_{i=1}^{n} a_i[\mathbf{T}(e_i) - \nabla A(\boldsymbol{\phi})]$, and hence, $\mathbf{v}^{\mathsf{T}}\mathbf{v} = \sum_{i=1}^{n} a_i w(e_i) = 0$, i.e., $\mathbf{v} = \mathbf{0}$. This shows that $\boldsymbol{H}_A(\boldsymbol{\phi})$ is positive definite for any $\boldsymbol{\phi} \in \{\boldsymbol{\phi} : A(\boldsymbol{\phi}) < \infty\}$. Therefore, by the inverse function theorem, there exists a continuously differentiable function $(\nabla A)^{-1}$ defined on an open neighborhood V of $\check{\boldsymbol{\phi}}$ to an open neighborhood W of $\nabla(A)(\check{\boldsymbol{\phi}})$ such that

$$
\check{\boldsymbol{\phi}} = (\nabla A)^{-1}\left(\frac{1}{M}\sum_{t=1}^{M} \mathbf{T}\left(\sum_{i=1}^{N} \hat{\varepsilon}_{i,t}\right)\right) \qquad \text{(J.16)}
$$

and $\nabla(\nabla A)^{-1}(\mathbf{z}) = [\boldsymbol{H}_A((\nabla A)^{-1}(\mathbf{z}))]^{-1}$ for all $\mathbf{z} \in W$. Let $\hat{\boldsymbol{\phi}}$ be the projection of $\check{\boldsymbol{\phi}}$ onto Φ and $\tilde{\boldsymbol{\phi}}$ be the ML estimate based on $\{\sum_{i=1}^{N} \varepsilon_{i,t} : t = 1, \ldots, M\}$, the sum of *true* demand shocks. Since $\mathbf{T}(\cdot)$ is a known smooth function, it follows that there exists a positive constant L_A such that

$$
\|\hat{\boldsymbol{\phi}} - \tilde{\boldsymbol{\phi}}\|_2 \leq \|\check{\boldsymbol{\phi}} - \tilde{\boldsymbol{\phi}}\|_2 \leq L_A \frac{1}{NM}\sum_{t=1}^{M}\left|\sum_{i=1}^{N}\hat{\varepsilon}_{i,t} - \sum_{i=1}^{N}\varepsilon_{i,t}\right|, \qquad \text{(J.17)}
$$

where L_A depends only on $\boldsymbol{H}_A(\cdot)$, $\mathbf{T}(\cdot)$, and the neighborhoods V and W. The N on the denominator comes from the term $\nabla(\nabla A)^{-1}(\mathbf{z})$ since it takes the form of an inverse of a covariance matrix of the sufficient statistics $\mathbf{T}$ in terms of $\sum_{i=1}^{N} \varepsilon_{i,t}$, which is at least in the order of N. By (J.11), (J.17), and Proposition 3, we have

$$
\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\|\hat{\boldsymbol{\phi}} - \tilde{\boldsymbol{\phi}}\|_2^2 \leq L_A^2 R^2 C_{\boldsymbol{\theta}} \frac{\bar{s}\log(dNM)}{N \wedge M}, \qquad \text{(J.18)}
$$

where $C_{\boldsymbol{\theta}} = K_6 + K_7\theta_{\max}$ and $\theta_{\max} = \max\{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 : \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta\}$. By Proposition 3 of Keskin et al. (2021), we further have $\mathbb{E}_{\boldsymbol{\phi}}\|\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}\|_2^2 \leq C_{\boldsymbol{\phi}} \log M/M$ for some positive constant $C_{\boldsymbol{\phi}}$. It follows that

$$
\mathbb{E}^{\pi}_{\mathbf{X},\boldsymbol{\theta}}\|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}\|_2^2 \leq 4L_{\boldsymbol{\phi}} \frac{\bar{s}\log(dNM)}{N \wedge M}, \qquad \text{(J.19)}
$$

where $L_{\boldsymbol{\phi}} = \max\{L_A^2 R^2 C_{\boldsymbol{\theta}}, C_{\boldsymbol{\phi}}\}$.

For any $t = M+1, \ldots, T$, since $\tilde{\mathbf{d}}_t^{\pi} = G(\hat{\boldsymbol{\theta}}_t, \hat{\boldsymbol{\phi}}, \varkappa, \mathbf{x}_t)$, we deduce from (J.7) and (J.8) that there exists a positive constant L_G such that

$$\|\tilde{\mathbf{d}}_t^{\pi} - \tilde{\mathbf{d}}_t^{*}\|_2^2 \le L_G \left(\sum_{i=1}^{N} \|\hat{\boldsymbol{\theta}}_{\mathcal{C}(\mathbf{x}_{i,t})} - \boldsymbol{\theta}_{\mathcal{C}(\mathbf{x}_{i,t})}\|_2^2 + \|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}\|_2^2 \right). \tag{J.20}$$

Hence, it follows from (J.10), Proposition 3, and (J.19) that

$$\begin{aligned}
\Delta_{\boldsymbol{\theta}}^{\pi}(N;T) &= \frac{1}{NT}\mathbb{E}_{\mathbf{X},\boldsymbol{\theta}}^{\pi}\left\{ \sum_{t=1}^{M} [\mathscr{R}^*(\boldsymbol{\theta}_t, \boldsymbol{\phi}, \varkappa, \mathbf{X}_t) - \mathscr{R}(\tilde{\mathbf{d}}_t^{\pi}; \boldsymbol{\theta}_t, \boldsymbol{\phi}, \varkappa, \mathbf{X}_t)] \right\} \\
&\quad + \frac{1}{NT}\mathbb{E}_{\mathbf{X},\boldsymbol{\theta}}^{\pi}\left\{ \sum_{t=M+1}^{T} [\mathscr{R}^*(\boldsymbol{\theta}_t, \boldsymbol{\phi}, \varkappa, \mathbf{X}_t) - \mathscr{R}(\tilde{\mathbf{d}}_t^{\pi}; \boldsymbol{\theta}_t, \boldsymbol{\phi}, \varkappa, \mathbf{X}_t)] \mathbb{I}\{\hat{\mathcal{C}}(\mathbf{X}_{i,t}) \neq \mathcal{C}(\mathbf{X}_{i,t})\} \right\} \\
&\quad + \frac{1}{NT}\mathbb{E}_{\mathbf{X},\boldsymbol{\theta}}^{\pi}\left\{ \sum_{t=M+1}^{T} [\mathscr{R}^*(\boldsymbol{\theta}_t, \boldsymbol{\phi}, \varkappa, \mathbf{X}_t) - \mathscr{R}(\tilde{\mathbf{d}}_t^{\pi}; \boldsymbol{\theta}_t, \boldsymbol{\phi}, \varkappa, \mathbf{X}_t)] \mathbb{I}\{\hat{\mathcal{C}}(\mathbf{X}_{i,t}) = \mathcal{C}(\mathbf{X}_{i,t})\} \right\} \\
&\le \frac{M}{T} R_{rg} + 3R_{\mathrm{rg}} \bar{\mathcal{K}}(K_1 + K_2)\sqrt{\frac{\log(NT)}{N \wedge T}} + \bar{\mathscr{R}}'' L_G(K_6 + K_7\theta_{\max} + 4L_{\phi}) \frac{\bar{s}\log(dNM)}{N \wedge M} \\
&\le \frac{K_{12}\bar{\mathcal{K}}\bar{s}\log(dNT)}{\sqrt{N \wedge T}},
\end{aligned} \tag{J.21}$$

where $K_{12} = R_{\mathrm{rg}}(2\kappa + 3K_1 + 3K_2) + \bar{\mathscr{R}}'' L_G(K_6 + K_7\theta_{\max} + 4L_{\phi})$. Q.E.D.

Appendix K: Extensions to More General Cost Functions

We discuss how to extend our analysis to incorporate piecewise linear cost functions with multiple kinks and general differentiable cost functions, and characterize conditions under which the same theoretical result in Theorem 4 still holds. As in §7.2, the function $g(\cdot)$ takes the form of the identity function. For piecewise linear cost functions with K kinks $\tilde{q}_1, \ldots, \tilde{q}_K$ and parameters $\{(a_k, b_k) : k = 1, \ldots, K+1, a_k \ge 0\ \forall k, b_1 \ge 0\}$, the first-order condition of maximizing the single period profit $\mathscr{R}(\tilde{\mathbf{d}}_t; \boldsymbol{\theta}, \boldsymbol{\phi}, \varkappa, \mathbf{x}_t)$ yields the following system of non-linear equations

$$\beta_{\mathcal{C}(\mathbf{x}_{j,t})}^{-1}(2\tilde{d}_{j,t} - \gamma_{\mathcal{C}(\mathbf{x}_{j,t})} - \boldsymbol{\alpha}_{\mathcal{C}(\mathbf{x}_{j,t})}^{\intercal}\mathbf{x}_{j,t}) - a_{K+1} - \sum_{k=1}^{K}(a_k - a_{k+1})F\left(\tilde{q}_k - \sum_{i=1}^{N}\tilde{d}_{i,t}; \boldsymbol{\phi}\right) = 0, \quad \forall j = 1, \ldots, N. \tag{K.1}$$

Then, the condition for the existence and uniqueness of the full-information solution to the above system of non-linear equations becomes $\bar{f}N|\beta_{\min}|\sum_{k=1}^{K}(a_k - a_{k+1}) < 2$. It is worth noting that the results of Proposition 6 and hence Theorem 4 extend easily to piecewise linear convex cost functions because $a_1 < a_2 < \cdots < a_{K+1}$. However, for piecewise concave functions, the condition becomes more difficult to satisfy when there are more kinks.

For a general twice differentiable cost function $C(\cdot)$, the first-order condition of profit maximization leads to a system of non-linear equations given by

$$\beta_{\mathcal{C}(\mathbf{x}_{j,t})}^{-1}(2\tilde{d}_{j,t} - \gamma_{\mathcal{C}(\mathbf{x}_{j,t})} - \boldsymbol{\alpha}_{\mathcal{C}(\mathbf{x}_{j,t})}^{\intercal}\mathbf{x}_{j,t}) - \int_{-\infty}^{\infty} C'\left(\sum_{i=1}^{N}\tilde{d}_{i,t} + e\right) f(e; \boldsymbol{\phi})\,\mathrm{d}e = 0, \quad \forall j = 1, \ldots, N, \tag{K.2}$$

where $C'(\cdot)$ denotes the derivative of $C(\cdot)$. If there is a constant $\bar{C}' > 0$ such that $|C'(q)| \leq \bar{C}'$ for all q, then there exists a solution to the unconstrained system of non-linear equations. Since the second-order derivatives of $\mathscr{R}$ with respect to $\tilde{\mathbf{d}}_t$ are

$$\frac{\partial^2 \mathscr{R}}{\partial \tilde{d}_{j,t} \partial \tilde{d}_{k,t}}(\tilde{\mathbf{d}}_t; \boldsymbol{\theta}, \boldsymbol{\phi}, \varkappa, \mathbf{x}_t) = 2\beta^{-1}_{\mathcal{C}(\mathbf{x}_{j,t})} \mathbb{I}\{j = k\} - \int_{-\infty}^{\infty} C'' \left(\sum_{i=1}^{N} \tilde{d}_{i,t} + e \right) f(e; \boldsymbol{\phi}) \, \mathrm{d}e, \quad \forall j, k = 1, \ldots, N \,, \tag{K.3}$$

where $C''(\cdot)$ denotes the second derivative of $C(\cdot)$, the condition for uniqueness is given by $-\int_{-\infty}^{\infty} C'' \left(\sum_{i=1}^{N} \tilde{d}_{i,t} + e \right) f(e; \boldsymbol{\phi}) \, \mathrm{d}e \cdot N |\beta_{\min}| < 2$. Note that the same result in Theorem 4 still holds for twice differentiable convex functions since $C''(q) > 0$ for all q. For twice differentiable concave functions, the result in Theorem 4 will also follow if $\bar{C}'' N |\beta_{\min}| < 2$, where $\bar{C}'' = \sup\{C''(q)\}$.