Data-driven Dynamic Pricing and Ordering with Perishable Inventory in a Changing Environment

N. Bora Keskin* Yuexing Li* Jing-Sheng Song*

*Duke University, Fuqua School of Business, e-mail: {bora.keskin,yuexing.li,jingsheng.song}@duke.edu

We consider a retailer that sells a perishable product, making joint pricing and inventory ordering decisions over a finite time horizon of T periods with lost sales. Exploring a real-life data set from a leading supermarket chain, we identify several distinctive challenges faced by such a retailer that have not been jointly studied in the literature: the retailer does not have perfect information on (1) the demand-price relationship, (2) the demand noise distribution, (3) the inventory perishability rate, and (4) how the demand-price relationship changes over time. Furthermore, the demand noise distribution is nonparametric for some products but parametric for others. To tackle these challenges, we design two types of data-driven pricing and ordering (DDPO) policies for the cases of nonparametric and parametric noise distributions. Measuring performance by regret, i.e., the profit loss caused by not knowing (1)-(4), we prove that the T-period regret of our DDPO policies are in the order of $T^{2/3}(\log T)^{1/2}$ and $T^{1/2}\log T$ in the cases of nonparametric and parametric noise distributions, respectively. These are the best achievable growth rates of regret in these settings (up to logarithmic terms). Implementing our policies in the context of the aforementioned real-life data set, we show that our approach significantly outperforms the historical decisions made by the supermarket chain. Moreover, we characterize parameter regimes that quantify the relative significance of the changing environment and product perishability. Finally, we extend our model to allow for age-dependent perishability and demand censoring, and modify our policies to address these issues.

Key words: Dynamic pricing, inventory control, perishable inventory, non-stationary environment, data-driven analysis, estimation, exploration-exploitation.

History: First version: May 31, 2019. This version: December 30, 2020

1. Introduction

1.1. Overview and Practical Motivations

Recent technological advances such as digital marketing, electronic shelf labeling, and item-level RFID sensing have enabled dynamic pricing strategies in many industries. One example is grocery retailing. Amazon, including its online grocery store AmazonFresh, changes product prices about every 10 minutes on average, which is 50 times more often than Walmart (Business Insider 2018). E.Leclerc, a French hypermarket chain, uses electronic shelf tags in about 200 stores, making over 5,000 price changes per week (RW3 2016). The benefits of dynamic pricing, especially for perishable inventory such as fresh produce, cannot be overstated. From a sales perspective, dynamic pricing stimulates demand to better match supply, thereby increasing profitability. It also helps reduce grocery waste and the consequent effort of waste management.

Despite these benefits, it remains extremely challenging to optimize pricing and ordering decisions for perishable products, especially when there is no perfect information on either the demand-price relationship or the product perishability rates. To make informed decisions, a retailer must first learn about the demand-price relationship and perishability rates using historical data. Even more taxingly, the product demand features such as the potential market size and price sensitivity can shift over time in an unpredictable fashion

due to external factors (see below for examples). This makes the information collected before a demand shift obsolete after the shift. Indeed, this is what we observe when analyzing a real-life data set from a leading supermarket chain in China as detailed below.

1.1.1. Observations from a grocery data set. This data set consists of sales and inventory ordering information of fresh vegetables and fruits across multiple stores in 2013. Figure 1a illustrates the evolution of the demand and price of ginger in a certain store, labeled as "store A" to preserve anonymity. Visually inspecting potential shifts in the demand-price relationship, we estimate the price-sensitivity of demand using the data locally before and after a visualized change-point (see the vertical dashed line in Figure 1a for the visualized change-point). The price-sensitivity estimates are statistically significant with $p \leq 0.05$, and their 95% confidence intervals are non-overlapping. This is strong evidence of temporal shifts in the demand-price relationship. We conduct a more detailed analysis by calibrating the demand shifts with multiple change-points based on maximum quasi-likelihood estimation (MQLE)—see §4.2.1 and Table 2 for details. Figure 1b displays the variability in the perished proportions of order-up-to levels, illustrating the stochastic nature of inventory perishability in this context.

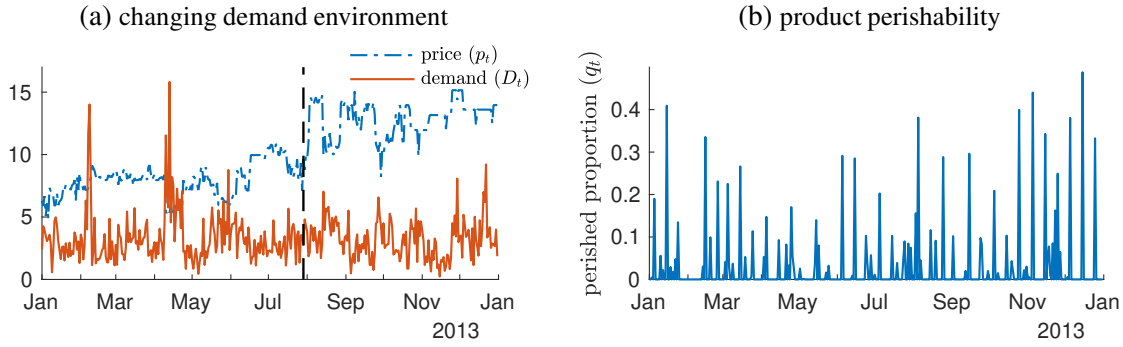

Figure 1 Demand, price, and inventory perishability data for ginger

Figure 2 shows the empirical cumulative distribution functions of the demand shocks for ginger and papaya, as well as the best normal fits to these empirical distributions. While the demand shocks for ginger appear to be normally distributed, the demand shocks for papaya do not. Thus, when designing decision tools, the retailer needs to distinguish between nonparametric and parametric demand noise distributions.

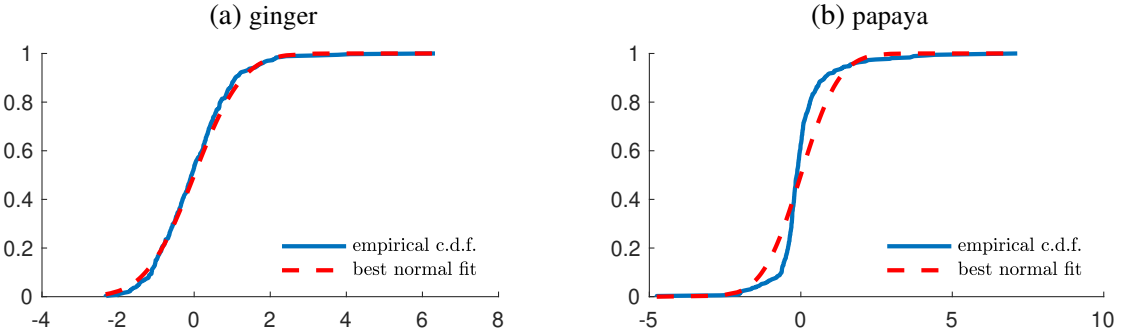

Figure 2 Empirical cumulative distribution function (c.d.f.) and best normal fit for demand noise

There are numerous examples of disruptive factors that give rise to unpredictable shifts in demand as well as perishability rates, such as those described below.

1.1.2. Pandemics. The COVID-19 pandemic resulted in dramatic shifts in consumer behavior, and gave rise to substantial uncertainty in how the consumer behavior will evolve after the pandemic (Briedis et al. 2020). This affected the demand-price relationship in almost all sectors: while many sectors faced sharp drops in demand, the demand for groceries "has risen to levels no one could have anticipated in early 2020, putting upward pressure on prices" (Abdelnour et al. 2020). According to a director at King Arthur Flour, "the demand ... is simply unprecedented and is outpacing the inventory," doubling the usual demand for flour during the busiest holiday months (CNN 2020). This indicates a sharp shift in the demand-price relationship, whose duration and magnitude cannot be inferred simply from past information. More importantly, the demand-price relationship is likely to keep changing after the pandemic (Briedis et al. 2020). A further layer of uncertainty arises from the varying degrees to which consumers purchase products in preparation for and in the aftermath of such disruptive events (Wong and Schuchard 2011). Therefore, retailers need to constantly track shifts in consumer behavior and adapt their plans accordingly (Becdach et al. 2020).

1.1.3. Weather events. Another disruptive factor is the weather. Despite ample production, the demand for watermelons and pineapples in India increased significantly along with a steady price hike when the temperature reached its record high in 2016, resulting in a prolonged and extreme summer (Fresh Plaza 2016). On the supply side, weather changes have a significant impact on food perishability, quality, and safety. Extreme heat and heavy rainfalls usually lead to accelerated food spoilage and bacterial contamination, increasing the risk of food poisoning. Ireland has reported skyrocketing E. coli infection cases (96 cases within 10 days) due to contaminated food as a heat wave swept the country in the summer of 2018 (The Irish Times 2018). Weather changes also threaten global food security by affecting major food suppliers. New malignant strains of wheat rust have spread over Africa, Asia, and Europe due to weather changes in recent years, devastating crops in those areas (Columbia University Earth Institute 2018). As a result, procurement has to rely on other sources of food supply whose standards may not be well-established, further increasing the unpredictability of food quality.

1.1.4. Rapid technological advances. Finally, upheavals brought about by technological changes play a vital role in the transformation of the retail industry. In recent years, technological advances have spawned various web-based retailers such as Amazon, Alibaba, and eBay, which largely intensified competition over all dimensions, including but not limited to price, product freshness, and food services. Besides the upsurging e-commerce, technology has been revolutionizing other aspects of retailing as well. On the demand side, augmented reality (AR) and virtual reality (VR) have been introduced into the shopping process to enhance consumer shopping experience and evaluation of the product, while on the supply side, Amazon and Walmart have attempted to deliver products using drones (HuffPost 2017). As noted by Pesaran et al. (2006),

such technological changes often lead to structural breaks in economic time series, further challenging the validity of forecasting based on past data.

To help retailers meet the above challenges, in this paper, we explicitly model the changing demand environment and product perishability in a single-product, periodic-review, lost-sales inventory system with a finite time horizon of T periods. We develop data-driven dynamic inventory and pricing policies to balance the trade-off between learning about an unknown environment to increase future profits and earning immediate profits. While we use fresh groceries as an illustration in the above examples, our model and methods generally apply to any perishable product.

1.2. Main Contributions and Qualitative Insights

Our work makes three main contributions to the literature on dynamic pricing and inventory control with demand learning.

1.2.1. Formulating a model motivated by observations on real-life data. Through an in-depth investigation of real-life data from a large supermarket chain, we construct a fairly general theoretical framework that entails dynamic pricing and ordering decisions with unknown demand-price relationship and inventory perishability rate in a changing environment. We design two versions of *data-driven pricing and ordering (DDPO) policies*, one for nonparametric noise distributions and another for exponential-family noise distributions. These policies prescribe how the retailer should make price experiments from time to time, and then collect and judiciously make use of demand and inventory perishability information to maximize profits. This data-driven approach provides evidence that real-life data can help modelers construct more realistic models, and thereby make more effective decisions. Our general approach of data-driven modeling and analysis may inspire and serve as a stepping stone for future research in other application areas, such as inventory planning in the presence of supply disruption risks or supply management in humanitarian operations.

1.2.2. Theoretical analysis: deriving rate-optimal regret bounds. We measure the policy performance by *regret*, i.e., the expected T-period profit loss relative to the full-information policy that optimizes pricing and ordering decisions with the perfect knowledge of the underlying demand environment and product perishability rates. We prove that the regret of the DDPO policies are $O(T^{2/3}(\log T)^{1/2})$ and $O(T^{1/2}\log T)$ for the nonparametric and parametric cases, respectively.[1] This shows that both versions of our policies are rate-optimal in the sense that they achieve the smallest possible growth rates of regret (up to logarithmic terms) in their respective settings. This theoretical analysis also reveals a significant difference between the best achievable regret performances for the nonparametric and parametric formulations of demand noise,

[1] Here and later, we say that the T-period regret of a policy is $O(g(T))$ if it is at most in the order of $g(T)$. In general, for all real-valued functions $f(\cdot)$ and $g(\cdot)$ defined on integers, we write $f(T)=O(g(T))$ if there exist $M\in(0,\infty)$ and $T_0\in\{1,2,\ldots\}$ such that $|f(T)|\leq Mg(T)$ for all $T\geq T_0$.

thereby quantifying the impact of demand noise distribution on profit performance. The same kind of analysis has the potential to be applied/adapted to a broader range of inventory models with or without pricing decisions.

1.2.3. Data-driven case study: managerial insights for practice. We use the aforementioned real-life data to test the performance of our policies in a realistically calibrated setting. We demonstrate that our policies outperform the historical decisions of the supermarket in real life, establishing the practical value of our approach. Furthermore, our analysis sheds light on the value of accounting for inventory perishability and changing environments in pricing and inventory decisions. Ignoring either feature leads to substantial profit loss, and we characterize parameter regimes under which ignoring one modeling feature causes more loss than ignoring the other. These insights inform practitioners about what kinds of information and data are the most valuable and how to use them. The case study also shows how to design data-driven pricing and inventory policies in practice.

1.3. Related Literature

There are a plethora of studies on dynamic pricing and learning (see Shin and Zeevi 2017, Ferreira et al. 2018, Keskin and Birge 2019, den Boer and Keskin 2019, 2020, Ban and Keskin 2020 for recent advances). We refer readers to den Boer (2015) and Chen and Chen (2015, §4) for broad reviews, and to Phillips (2005) for practical applications. In this paper, we consider a substantially more general formulation that entails both (i) changing environments and (ii) inventory ordering decisions with product perishability to address the learning-and-earning trade-off.

Focusing on dynamic pricing with demand learning in a changing environment, Besbes and Zeevi (2011) consider a single change-point such that the demand models before and after the change are known with certainty, and design a policy that has $O(N^{1/2}\log N)$ regret, where N is the number of customers that arrive sequentially. Keskin and Zeevi (2017) analyze more general demand environments that exhibit more frequent change-points, and develop a policy that achieves $O(T^{1/2}\log T)$ regret in the presence of abrupt changes. Unlike these studies, our work analyzes joint pricing and inventory ordering decisions. The introduction of inventory decisions greatly complicates the analysis because learning the demand noise distribution, which is inessential in the absence of inventory decisions, plays a major role in our setting. We design policies to address this difficulty and prove that the resulting regret is $O(T^{1/2}\log T)$ when the demand noise distribution adopts a parametric form, thereby extending this research stream to inventory management contexts.

Various authors consider joint dynamic pricing and inventory decisions for perishable products, assuming full knowledge of demand information; see, e.g., Li et al. (2012), Chen et al. (2014), and the references therein. In contrast, our work enables decision makers who have no knowledge about the demand-price relationship to learn such information on the fly while maximizing profits. Two more recent works are closer

to our setting, although neither of them studies a changing demand environment as we do; see Chen et al. (2019) for the backorder setting and Chen et al. (2020) for the lost-sales setting. The objective function in Chen et al. (2019) is jointly concave in mean demand and the order-up-to level, which enables them to design a policy with $O(T^{1/2})$ regret, but this is no longer the case in our lost-sales setting. Chen et al. (2020) design a nonparametric policy using spline approximation to tackle censored demand in the lost-sales setting and show that their policy has $O(T^{1/2+\varepsilon})$ regret, where $\varepsilon > 0$. This result is based on the assumption that the objective function is strictly concave in price (see their Assumption 1(v) in §2.5). However, with general noise distributions, the objective function may lose differentiability and the problem becomes more challenging. We show that the regret of the nonparametric version of our policy is $O(T^{2/3}(\log T)^{1/2})$, attaining the best achievable growth rate of regret, which is of order $T^{2/3}$ when the noise distribution takes a general nonparametric form. In addition, the changing environment in our setting leads to detection errors that further complicate the analysis. In our policies, we design a detection test to capture the changing demand-price relationship without further increasing the growth rate of regret.

Traditional perishable inventory models with exogenous demand (i.e., without consideration of pricing) can be classified into two categories: (i) all perishable items have the same fixed lifetime, after which the products are disposed of, such as packaged and processed food products; (ii) perishable items have random lifetimes with exponential decay, such as fruits and vegetables. Nahmias (1982) and Karaesmen et al. (2011) provide comprehensive reviews of inventory models of the first category, while Goyal and Giri (2001) and Bakker et al. (2012) summarize the inventory models that fall into the second category. We adopt the latter approach and formulate inventory perishability as a random proportion of the order-up-to level in each period, whereas all of the aforementioned joint pricing and inventory models assume a fixed lifetime, although Chen et al. (2014) discuss extensions to random lifetimes within the fixed lifetime framework.

Table 1 summarizes the major differences and the position of our work in the existing literature.

Table 1 Position of Our Work in the Existing Literature

	Dynamic Pricing	Demand Learning	Changing Environment	Inventory Decision	Product Perishability	Real-life Data
Our Work	✓	✓	✓	✓	✓	✓
Chen et al. (2014)	✓			✓	✓	
Keskin and Zeevi (2017)	✓	✓	✓			
Chen et al. (2019)	✓	✓		✓		

Organization of the paper. The rest of the paper is organized as follows. We describe the model in §2 and define our policies in §3. We present the main theoretical results and a case study in §4, and provide the detailed analysis in §5. We study several extensions in §6 and conclude the paper in §7. All proofs are in Appendix A, and additional numerical results for robustness checks are in Appendix B.

2. Problem Formulation

2.1. Basic Model Elements

Consider a retailer that sells a single perishable product over T periods with lost sales. In each period $t = 1, 2, \ldots, T$, the following sequence of events occurs:

(a) At the beginning of the period, the retailer observes the on-hand inventory x_t.

(b) The retailer chooses a price $p_t \in \mathscr{P} = [p_{\min}, p_{\max}]$ and an order-up-to level (i.e., an inventory position) $y_t \in \mathscr{Y} = [y_{\min}, y_{\max}]$, where $0 < p_{\min} < p_{\max} < \infty$, and $0 < y_{\min} < y_{\max} < \infty$.

(c) The replenishment lead time is zero. (In the context of the real-life data set introduced in §1.1.1, the orders are placed at the end of each day and arrive by the next morning, i.e., overnight delivery. Since no decisions are made during the lead time, it can be treated as 0.) By the time replenishment orders arrive at the store, a proportion q_t of the inventory position y_t perishes. Consumers are able to identify perished products and thus do not purchase them. This is equivalent to discarding those products before the demand is realized.

(d) The demand D_t is realized and satisfied to the maximum extent by the remaining on-hand inventory. We assume that unsatisfied demand is lost but observable to the retailer. We consider demand censoring as an extension in §6.2.

(e) The end-of-period inventory is updated as $x_{t+1} = [(1 - q_t)y_t - D_t]^+$.

Regarding the assumptions on q_t, note that existing studies usually model random perishability either by assuming that the product lifetime follows a certain distribution (see Kalpakam and Sapna 1994 for exponentially distributed lifetime), or by modeling the inventory process using a differential equation (see Liao 2007 for exponential decay with constant failure rate). In our setting, it is more appropriate to model product perishability in a way compatible with the periodic review system. Thus, we start by assuming that the product lifetime follows a geometric distribution with parameter q, the discrete counterpart of the exponential distribution. We allow for further randomness by assuming that the parameter q follows a beta distribution, so the perished inventories depend on the realizations of q_t and the order-up-to level y_t in period t, and thus can vary over time. However, this assumption implicitly implies that the perishability is independent of product age. We consider an extension to age-dependent perishability in §6.1.

In response to the selling price p_t, the demand in period t is realized as follows:

$$D_t = g(\alpha_t + \beta_t p_t) + \varepsilon_t \quad \text{for } t = 1, 2, \ldots \tag{1}$$

where $\alpha_t \in \mathbb{R}$ and $\beta_t \in (-\infty, 0)$ are demand model parameters that are unknown to the retailer, $g : \mathbb{R} \to \mathbb{R}$ is a twice continuously differentiable and increasing function that is known to the retailer, and ε_t is an unobservable demand shock (a.k.a. demand noise) in period t. Letting $\mathbf{X}_t = (1, p_t)^\intercal$ and $\boldsymbol{\theta}_t = (\alpha_t, \beta_t)^\intercal$, we express (1) in the following form:

$$D_t = g(\mathbf{X}_t^\intercal \boldsymbol{\theta}_t) + \varepsilon_t \quad \text{for } t = 1, 2, \ldots \tag{2}$$

The demand parameter vector $\boldsymbol{\theta}_t$ can vary over time: the sequence $\boldsymbol{\theta} = \{\boldsymbol{\theta}_t,\, t = 1,\ldots,T\}$ assumes $\mathscr{C}$ change-points, taking different values in a compact rectangle $\Theta \subset \mathbb{R} \times (-\infty, 0)$. Denoting the j^{th} change-point by t_j^*, we let $1 = t_0^* < t_1^* < \cdots < t_{\mathscr{C}}^* < t_{\mathscr{C}+1}^* = T+1$, and $t_j^* = \inf\{t > t_{j-1}^* : \boldsymbol{\theta}_t \neq \boldsymbol{\theta}_{t_{j-1}^*}\}$ for $j = 1,\ldots,\mathscr{C}+1$. Each change-point results in a minimum shift of $\delta_{\min} > 0$ in the value of $\boldsymbol{\theta}_t$, i.e., $\inf\{\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_s\| : \boldsymbol{\theta}_t \neq \boldsymbol{\theta}_s, 1 \leq s \leq t \leq T\} \geq \delta_{\min}$. The retailer knows neither the number of the change-points, nor the time the change-points occur, nor the values of $\boldsymbol{\theta}_t$ before and after a change.

The demand noise terms $\{\varepsilon_t,\, t = 1,2,\ldots\}$ are i.i.d. with a common cumulative distribution function F_ε such that $\mathbb{E}_\varepsilon[\varepsilon_t] = 0$ and there exist positive constants σ_0, ϖ_0, and ϱ_0 satisfying $\mathbb{E}_\varepsilon[\varepsilon_t^2] \leq \sigma_0^2$ and $\mathbb{E}_\varepsilon[\exp(\varpi\varepsilon_t)] \leq \exp(\frac{1}{2}\varrho_0\sigma_0^2\varpi^2)$ for $t \in \{1,\ldots,T\}$ and $\varpi \in \mathbb{R}$ with $|\varpi| \leq \varpi_0$. [2] The realizations of perished proportions $\{q_t,\, t = 1,2,\ldots\}$ are i.i.d. following a beta distribution with parameter vector $\boldsymbol{\xi} = (\lambda, \nu)$ chosen from a compact rectangle $\Xi \subset (0,\infty) \times (0,\infty)$. Moreover, q_t is independent of ε_t for all t. [3] The relevant cost parameters are h, b, w, and c, which are the per-period unit holding, lost-sales penalty, disposal, and ordering costs, respectively. They are all known to the retailer. We suppress the time-dependency of these cost parameters for simplicity but all of our analysis still holds for time-varying costs. Throughout the sequel, we assume that $|h - c| < b + p$ for all $p \in \mathscr{P}$. This is a mild assumption because, in practice, the unit lost-sales penalty cost usually exceeds the unit holding cost (i.e., $b > h$), and the selling price must exceed the unit ordering cost (i.e., $p > c$) so that the retailer can make a profit.

2.2. Admissible Policies and Performance Metric

If the retailer knew F_ε, $\boldsymbol{\theta}$, and $\boldsymbol{\xi}$ in advance, the retailer's optimal expected profit over T periods would be given by the following optimization problem:

$$\begin{aligned} V(\mathbf{x}) &= \max_{\substack{(p_t,y_t)\in\mathscr{P}\times\mathscr{Y}\\ y_t\geq x_t}} \left\{ \sum_{t=1}^{T} \mathbb{E}_{\varepsilon,q|\boldsymbol{\xi}} \Big[p_t \min\{D_t, (1-q_t)y_t\} - h[(1-q_t)y_t - D_t]^+ \right. \\ &\qquad \left. - wq_t y_t - b[D_t - (1-q_t)y_t]^+ - c(y_t - x_t) \Big] + cx_{T+1} \right\} \\ &= \max_{\substack{(p_t,y_t)\in\mathscr{P}\times\mathscr{Y}\\ y_t\geq x_t}} \left\{ cx_1 + \sum_{t=1}^{T} Q(p_t, y_t; \boldsymbol{\theta}_t, \boldsymbol{\xi}) \right\} \\ &\text{subject to} \quad x_{t+1} = [(1-q_t)y_t - g(\mathbf{X}_t^\intercal \boldsymbol{\theta}_t) - \varepsilon_t]^+ \ \text{ for } t = 1,\ldots,T, \end{aligned} \tag{3}$$

where $\mathbf{x} = (x_1,\ldots,x_{T+1})$, and $Q(p_t, y_t; \boldsymbol{\theta}_t, \boldsymbol{\xi})$ is the expected single-period profit, which equals the expected single-period revenue minus total per-period costs, i.e.,

$$Q(p_t, y_t; \boldsymbol{\theta}_t, \boldsymbol{\xi}) = p_t g(\mathbf{X}_t^\intercal \boldsymbol{\theta}_t) - H(y_t; p_t, \boldsymbol{\theta}_t, \boldsymbol{\xi}), \tag{4}$$

[2] Here and later, given an i.i.d. sequence of random variables $Z = \{Z_t,\, t = 1,2,\ldots\}$, we write $\mathbb{P}_Z$ to denote the probability measure governing Z, and $\mathbb{E}_Z$ to denote the corresponding expectation operator.

[3] We note that, while the demand parameter vector $\boldsymbol{\theta}_t$ is allowed to vary over time, the perishability parameter vector $\boldsymbol{\xi}$ does not change. This is solely for the purpose of exposition. Learning the perishability model is relatively easier than learning the demand model because perishability observations are not affected by the retailer's decisions. Accordingly, our analysis can be easily extended to accommodate time-varying $\boldsymbol{\xi}$.

with its second term, the total per-period cost as a function of y_t, given by

$$\begin{aligned} H(y_t;p_t,\boldsymbol{\theta}_t,\boldsymbol{\xi}) = (h-c)\mathbb{E}_{\varepsilon,q|\boldsymbol{\xi}}[(1-q_t)y_t - g(\mathbf{X}_t^\top\boldsymbol{\theta}_t) - \varepsilon_t]^+ \\ + (b+p_t)\mathbb{E}_{\varepsilon,q|\boldsymbol{\xi}}[g(\mathbf{X}_t^\top\boldsymbol{\theta}_t) + \varepsilon_t - (1-q_t)y_t]^+ + \big(w\mathbb{E}_{q|\boldsymbol{\xi}}[q_t] + c\big)y_t. \end{aligned} \tag{5}$$

For more details about the above derivation, see Heyman and Sobel (1982).

Suppose that an optimal solution to the problem (3) exists in the interior of $\mathscr{P}^T \times \mathscr{Y}^T$, and let $\{(p_t^*, y_t^*) : t = 1,\dots,T\}$ be one such optimal solution. Let π^* be the corresponding optimal policy that maps the state $\mathbf{x}$ to this sequence of optimal controls. We term this policy as the *full-information anticipatory* (FIA) policy. As the name implies, under this policy, the retailer has full information on all model parameters and can act in anticipation of the change-points in $\boldsymbol{\theta} = \{\boldsymbol{\theta}_t,\, t = 1,\dots,T\}$. We use the FIA policy as the benchmark to measure performance.

Because the retailer does not know F_ε, $\boldsymbol{\theta}$, or $\boldsymbol{\xi}$ in advance, the retailer utilizes the accumulated information to learn the underlying demand model as well as the inventory perishability. Let $\mathbf{I}_t$ be the information vector containing history of observed demands, perished proportions, and pricing and inventory decisions up to period $t = 1, 2, \dots$, i.e., $\mathbf{I}_t = (D_1, q_1, p_1, y_1, \dots, D_t, q_t, p_t, y_t)$, with $\mathbf{I}_0 = \varnothing$. An admissible policy is a sequence of functions $\pi = \{\pi_t : t = 1,\dots,T\}$ such that $\pi_t : \mathbf{I}_{t-1} \to \mathscr{P} \times \mathscr{Y}$ for all $t = 1,\dots,T$, with π_1 being a constant function. (The aforementioned FIA policy is not an admissible policy as it relies on the knowledge of F_ε, $\boldsymbol{\theta}$, and $\boldsymbol{\xi}$, which are unknown to the retailer.)

Given a demand noise distribution F_ε, a sequence of demand parameter vectors $\boldsymbol{\theta}$, a perishability parameter vector $\boldsymbol{\xi}$, and an admissible policy π, we define a probability measure $\mathbb{P}^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}$ on the space of demand sequences $\mathbf{D} = (D_1,\dots,D_T)$ and perished proportion sequences $\mathbf{q} = (q_1,\dots,q_T)$ as follows:

$$\mathbb{P}^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}(D_1 \in \mathrm{d}\tilde{D}_1,\dots,D_T \in \mathrm{d}\tilde{D}_T, q_1 \in \mathrm{d}\tilde{q}_1,\dots,q_T \in \mathrm{d}\tilde{q}_T) = \prod_{t=1}^T \mathbb{P}_\varepsilon(\alpha_t + \beta_t p_t + \varepsilon_t \in \mathrm{d}\tilde{D}_t)\mathbb{P}_q(q_t \in \mathrm{d}\tilde{q}_t) \tag{6}$$

for all $(\tilde{D}_1,\dots,\tilde{D}_T) \in \mathbb{R}^T$ and $(\tilde{q}_1,\dots,\tilde{q}_T) \in [0,1]^T$. We also let $\mathbb{E}^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}$ be the expectation operator associated with $\mathbb{P}^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}$. We measure the performance of an admissible policy π based on its T-period *regret*, given by the expected T-period profit loss under π relative to the FIA policy, i.e.,

$$\Delta^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}(T) = \mathbb{E}^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{t=1}^T \big(Q(p_t^*, y_t^*; \boldsymbol{\theta}_t, \boldsymbol{\xi}) - Q(p_t, y_t; \boldsymbol{\theta}_t, \boldsymbol{\xi})\big)\right]. \tag{7}$$

Here and later, we use the terms *regret* and *profit loss* interchangeably. We tend to use the former when theoretical treatment is of primary interest, and the latter to present managerial insights based on real-life data. We are interested in admissible policies that make the retailer's regret as small as possible.

Motivated by our observations on the real-life data set introduced in §1.1.1, we consider the following two settings:

- **Setting N:** The demand noise distribution F_ε does not necessarily bear a parametric form.
- **Setting E:** The demand noise distribution F_ε is known to belong to the exponential family of distributions. That is, the density of the demand noise terms has the form $f_\varepsilon(\varepsilon;\boldsymbol{\varphi}) = B(\varepsilon)\exp[\boldsymbol{\varphi}^\top\mathbf{T}(\varepsilon) -$

$A(\boldsymbol{\varphi})]$, where $A(\cdot)$, $B(\cdot)$, and $\mathbf{T}(\cdot)$ are known functions, and $\boldsymbol{\varphi}$ is an unknown parameter vector chosen from a compact set Φ. This family contains several commonly used distributions such as normal, exponential, gamma, beta, and Poisson, among others.

Here, the letter N stands for *nonparametric* and letter E stands for *exponential family*.

As alluded to earlier and explained in detail below, the analysis of Settings N and E enables us to identify how the uncertainty regarding the demand noise distribution affects the profit performance.

3. Data-driven Dynamic Pricing and Ordering (DDPO) Policy

In this section we explore plausible and practical pricing and ordering policies for the retailer. Since the retailer does not know F_ε, $\boldsymbol{\theta}$, or $\boldsymbol{\xi}$, it is extremely challenging to solve the problem (3) using traditional techniques such as dynamic programming. Even if such a solution were possible by re-writing the optimization problem, it would easily run into the curse of dimensionality. To overcome this intractability, we design a *data-driven dynamic pricing and ordering* (DDPO) policy that balances the learning-and-earning trade-off on the aforementioned uncertainties by maximizing the single-period profit function (4) with the most recent inference on the model parameters and demand noise distribution. This can be viewed as a dimensionality reduction technique that summarizes the accumulated information at any given time.

We design two versions of the aforementioned DDPO policy: The DDPO-N policy is for Setting N while the DDPO-E policy is for Setting E. Because Setting E is a special case of Setting N, and the two versions of the policy differ only in a few places in their specifications, we describe the DDPO policy in a unified framework and indicate the differences of the two versions when necessary. All the results pertaining to the DDPO-N (DDPO-E) policy are expressed in Setting N (Setting E).

The DDPO policy depends on five parameters and is formally denoted as $\mathscr{D}(\eta, \kappa, \omega_1, \omega_2, \upsilon)$, where $\eta > 0$, $\kappa > 0$, ω_1 and ω_2 are two distinct test prices in $\mathscr{P}$, and υ is a test order-up-to level in $\mathscr{Y}$. This policy divides the time horizon into cycles of n periods, labeled by $\tau = 0, 1, \ldots, \lfloor T/n \rfloor$. In each cycle, the first $2m$ periods are used for price experimentation and change-point detection, while the remaining periods are dedicated to profit maximization. For DDPO-N, we choose $m \equiv m_{\mathrm{N}} = \lceil \kappa T^{1/3} \rceil$ and $n \equiv n_{\mathrm{N}} = \lceil \kappa T^{2/3} \rceil$, whereas for DDPO-E, $m \equiv m_{\mathrm{E}} = \lceil \kappa \log T \rceil$ and $n \equiv n_{\mathrm{E}} = \lceil \kappa T^{1/2} \rceil$. For instance, if each period corresponds to a day and the time horizon T is 365 days, then a cycle may last about a month and $2m$ periods may span about a week. As detailed in our analysis below, the different choices of m and n reflect the distinct rates of information accumulation under the different assumptions on the demand noise distribution in Settings N and E. Let

$$\mathscr{X}_{i\tau} = \{t = \tau n + (i-1)m + s : s = 1, 2, \ldots, m\} \text{ for } i = 1, 2 \text{ and } \tau = 0, 1, \ldots, \lfloor T/n \rfloor. \tag{8}$$

Then, the set of periods dedicated to experimentation is $\mathscr{X} = \mathscr{X}_1 \cup \mathscr{X}_2$, where $\mathscr{X}_i = \cup_\tau \mathscr{X}_{i\tau}$ for $i = 1, 2$. For any cycle τ, let χ_τ be the indicator of a detected mean demand change (of magnitude at least η), with $\chi_0 = 1$. Denoting the *latest detection cycle* by $L(\tau) = \max\{\tau' \le \tau : \chi_{\tau'} = 1\}$, we have

$$\chi_{\tau+1} = \begin{cases} 1 \text{ if } \sup_{i,\tau'} \{|\bar{D}_{i\tau} - \bar{D}_{i\tau'}| : i = 1, 2, L(\tau) \le \tau' < \tau\} > \eta, \\ 0 \text{ otherwise,} \end{cases} \tag{9}$$

where $\bar{D}_{i\tau} = m^{-1}\sum_{t\in\mathscr{X}_{i\tau}} D_t$. Based on this, the DDPO policy discards all past demand information immediately after a change in the mean demand is detected. Note that a change in the demand parameter vector $\boldsymbol{\theta}_t$ is directly translated to a change in the mean demand because the DDPO policy uses the same pair of prices, ω_1 and ω_2, for detection in every cycle.

Given the above construction, we now describe the execution of $\mathscr{D}(\eta,\kappa,\omega_1,\omega_2,\upsilon)$ in period t in the following three steps:

1. **Inference on model parameters and demand noise distribution.** If $t\in\mathscr{X}_i$ for $i=1,2$, set $p_t=\omega_i$ and $y_t=\upsilon$. Otherwise, note that the cycle of period t is $\tau=\lceil t/n\rceil-1$, and let $\check{\boldsymbol{\theta}}_t$ be the (unconstrained) maximum quasi-likelihood estimator based on the information collected from the beginning of cycle $L(\tau)$ up to period t. Specifically, $\check{\boldsymbol{\theta}}_t$ solves the following equation in terms of $\boldsymbol{\vartheta}$:
$$\sum_{s=nL(\tau)+1}^{t}\mathbb{I}\{s\in\mathscr{X}\}(D_s-g(\mathbf{X}_s^{\intercal}\boldsymbol{\vartheta}))\mathbf{X}_s=\mathbf{0}. \tag{10}$$
Let $\hat{\boldsymbol{\theta}}_t$ be the projection of $\check{\boldsymbol{\theta}}_t$ onto the parameter space Θ, i.e., $\hat{\boldsymbol{\theta}}_t=\operatorname{argmin}_{\tilde{\boldsymbol{\theta}}\in\Theta}||\tilde{\boldsymbol{\theta}}-\check{\boldsymbol{\theta}}_t||$. Note that for any pair of periods $t,t'\notin\mathscr{X}$ in the same cycle τ, we have $\hat{\boldsymbol{\theta}}_t=\hat{\boldsymbol{\theta}}_{t'}$ by (10). The residual vector $\mathbf{e}$ is also recorded as a proxy for the demand noise: for $s\in\{nL(\tau)+1,\ldots,t\}\cap\mathscr{X}$, the s^{th} component of $\mathbf{e}$ is $e_s=D_s-g(\mathbf{X}_s^{\intercal}\hat{\boldsymbol{\theta}}_t)$.

   For DDPO-N, the retailer directly uses the recorded residuals to construct an estimator $\hat{F}_{\mathbf{e}}$ of F_ε as
$$\hat{F}_{\mathbf{e}}(v)=\frac{1}{2M_t}\sum_{s=nL(\tau)+1}^{t}\mathbb{I}\{s\in\mathscr{X}\}\mathbb{I}\{e_s\le v\}, \tag{11}$$
where $M_t=m(\lceil t/n\rceil-L(\tau))$ is the effective sample size in period t. For DDPO-E, however, the retailer knows that the demand noise distribution falls into the exponential family. The only unknown is the parameter vector $\boldsymbol{\varphi}$ chosen from a compact set Φ. Hence, based on the residual vector $\mathbf{e}$, the retailer uses maximum likelihood estimation (MLE), followed by projection onto Φ, to obtain an estimator $\hat{\boldsymbol{\varphi}}_t$ of $\boldsymbol{\varphi}$, in a way similar to the computation of $\hat{\boldsymbol{\theta}}_t$.

   The perishability parameter $\boldsymbol{\xi}$ is also estimated via MLE. To be more precise, the (unconstrained) maximum likelihood estimator $\check{\boldsymbol{\xi}}_t$ of $\boldsymbol{\xi}=(\lambda,\nu)$ in the beta distribution solves the following set of equations in terms of λ and ν:
$$\sum_{s=1}^{t}\mathbb{I}\{s\in\mathscr{X}\}\log q_s=\sum_{s=1}^{t}\mathbb{I}\{s\in\mathscr{X}\}[\digamma(\lambda)-\digamma(\lambda+\nu)], \tag{12}$$
$$\sum_{s=1}^{t}\mathbb{I}\{s\in\mathscr{X}\}\log(1-q_s)=\sum_{s=1}^{t}\mathbb{I}\{s\in\mathscr{X}\}[\digamma(\nu)-\digamma(\lambda+\nu)], \tag{13}$$
where $\digamma(z)=\frac{\partial}{\partial z}\log\Gamma(z)$ is the digamma function. The estimator $\hat{\boldsymbol{\xi}}_t$ of $\boldsymbol{\xi}$ is found by projecting $\check{\boldsymbol{\xi}}_t$ onto the compact set Ξ.

2. **Maximization of a proxy profit function.** For any $t \notin \mathscr{X}$ in cycle $\tau = \lceil t/n \rceil - 1$, the decisions of the DDPO-N policy are found by maximizing a proxy profit function as follows:

$$\max_{(p_t,y_t)\in\mathscr{P}_d\times\mathscr{Y}} \hat{Q}_t(p_t,y_t;\hat{\boldsymbol{\theta}}_t,\hat{\boldsymbol{\xi}}_t,\mathbf{e}) = \max_{p_t\in\mathscr{P}_d} \Big\{ \underbrace{p_t g(\mathbf{X}_t^\intercal \hat{\boldsymbol{\theta}}_t) - \min_{y_t\in\mathscr{Y}} \hat{H}_t(y_t;p_t,\hat{\boldsymbol{\theta}}_t,\hat{\boldsymbol{\xi}}_t,\mathbf{e})}_{\equiv \hat{G}_t(p_t;\hat{\boldsymbol{\theta}}_t,\hat{\boldsymbol{\xi}}_t,\mathbf{e})} \Big\}, \tag{14}$$

where

$$\begin{aligned} \hat{H}_t(y_t;p_t,\hat{\boldsymbol{\theta}}_t,\hat{\boldsymbol{\xi}}_t,\mathbf{e}) = \frac{1}{2M_t} \sum_{\substack{s=nL(\tau)+1 \\ s\in\mathscr{X}}}^{t} \mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t} \Big[& (h-c)[(1-q_t)y_t - g(\mathbf{X}_t^\intercal\hat{\boldsymbol{\theta}}_t) - e_s]^+ \\ & + (b+p_t)[g(\mathbf{X}_t^\intercal\hat{\boldsymbol{\theta}}_t) + e_s - (1-q_t)y_t]^+ \Big] + \big(w\mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[q_t] + c\big)y_t\,, \end{aligned} \tag{15}$$

$\mathscr{P}_d = \big\{p_{\min} + i\rho\sqrt{(\log M_t)/M_t} : i = 0,1,2,\ldots, \big\lfloor \frac{1}{\rho}(p_{\max}-p_{\min})\sqrt{M_t/(\log M_t)}\,\big\rfloor\big\}$, and ρ is a positive constant. We note that the maximization with respect to the pricing decision is over a sufficiently sparse grid $\mathscr{P}_d$ with step size $\iota_t = \rho\sqrt{(\log M_t)/M_t}$. This choice of ι_t strikes a balance between the estimation of F_ε and the optimization of p_t (see Proposition 6).

The decision of the DDPO-E policy is found by solving the following problem:

$$\max_{(p_t,y_t)\in\mathscr{P}\times\mathscr{Y}} Q(p_t,y_t;\hat{\boldsymbol{\theta}}_t,\hat{\boldsymbol{\xi}}_t,\hat{\boldsymbol{\varphi}}_t) = \max_{p_t\in\mathscr{P}} \Big\{ \underbrace{p_t g(\mathbf{X}_t^\intercal\hat{\boldsymbol{\theta}}_t) - \min_{y_t\in\mathscr{Y}} H(y_t;p_t,\hat{\boldsymbol{\theta}}_t,\hat{\boldsymbol{\xi}}_t,\hat{\boldsymbol{\varphi}}_t)}_{\equiv G^u(p_t;\hat{\boldsymbol{\theta}}_t,\hat{\boldsymbol{\xi}}_t,\hat{\boldsymbol{\varphi}}_t)} \Big\}, \tag{16}$$

where

$$\begin{aligned} H(y_t;p_t,\hat{\boldsymbol{\theta}}_t,\hat{\boldsymbol{\xi}}_t,\hat{\boldsymbol{\varphi}}_t) = \mathbb{E}_{\varepsilon|\hat{\boldsymbol{\varphi}}_t,q|\hat{\boldsymbol{\xi}}_t} \Big[& (h-c)[(1-q_t)y_t - g(\mathbf{X}_t^\intercal\hat{\boldsymbol{\theta}}_t) - \varepsilon_t]^+ \\ & + (b+p_t)[g(\mathbf{X}_t^\intercal\hat{\boldsymbol{\theta}}_t) + \varepsilon_t - (1-q_t)y_t]^+ \Big] + \big(w\mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[q_t] + c\big)y_t. \end{aligned} \tag{17}$$

Let $(\hat{p}_t^u, \hat{y}_t^u)$ be the generic optimal solution to the profit maximization problems above, namely (14) for DDPO-N and (16) for DDPO-E. Here, the superscript u indicates unconstrained solutions where the inventory constraint $y_t \geq x_t$ is not considered. Set $\hat{p}_t = \hat{p}_t^u$ and $\hat{y}_t = \max\{\hat{y}_t^u, x_t\}$.

3. **Change-point detection.** Update the sequence $\boldsymbol{\chi}$ of claimed detection indicators according to (9). Discard all historical information if $\chi_{\tau+1} = 1$, and keep accumulating information if $\chi_{\tau+1} = 0$ while proceeding to the next cycle.

The DDPO policy loops back and forth between the three steps with τ incremented by 1 in each iteration until the total number of periods T is reached.

4. Main Results and Case Study

In this section, we present our main theoretical results and a case study based on our real-life data set. We defer the detailed theoretical analysis to §5.

4.1. Main Theoretical Results

THEOREM 1. *(a) For the DDPO-N policy, there exists a positive constant K_1 such that*

$$\Delta^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}(T) \leq K_1 T^{2/3}(\log T)^{1/2} \text{ for } T = 3,4,\ldots \tag{18}$$

(b) For the DDPO-E policy, there exists a positive constant K_2 such that

$$\Delta^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}(T) \leq K_2 T^{1/2}\log T \text{ for } T = 3,4,\ldots \tag{19}$$

Theorem 1 states that the T-period regret is $O(T^{2/3}(\log T)^{1/2})$ for DDPO-N and $O(T^{1/2}\log T)$ for DDPO-E. It is worth emphasizing that this difference stems from the general nonparametric specification on the demand noise distribution. In Setting N, the profit function $Q(p,y)$ is Lipschitz but non-differentiable at some points. Thus, the relevant loss function determining regret is the mean absolute deviation of the demand parameter estimates. This is of the form $\mathbb{E}\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t\|$, and leads to a regret of order $T^{2/3}$. A matching lower bound related to Setting N is in Theorem 4.2 in Kleinberg (2005), which states that for any admissible policy there exists a problem instance such that the policy's regret is at least of order T^{β}, where $\beta = \frac{\alpha+1}{2\alpha+1}$ and α is the exponent of a uniformly locally Lipschitz condition on the objective function (Kleinberg 2005, §2). In Setting N, the objective function $Q(p,y)$ is Lipschitz with exponent $\alpha = 1$; hence, the corresponding lower bound is of order $T^{\beta} = T^{2/3}$. Note that this lower bound is obtained in a one-dimensional decision space. By contrast, our problem is more challenging with a two-dimensional decision space and an inventory constraint, yet our DDPO-N policy achieves a regret of order $T^{2/3}$ (up to logarithmic terms). In Setting E, the noise distribution has a continuous density. So, $Q(p,y)$ is differentiable in both variables and thus the relevant loss function determining regret is the mean squared error of the demand parameter estimates. This is of the form $\mathbb{E}\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t\|^2$, and leads to a regret of order $T^{1/2}$. Matching lower bounds in the same order exist in Besbes and Zeevi (2011) and Keskin and Zeevi (2014), who consider a single change-point setting and a static environment, respectively. In this regard, both of our DDPO policies are *rate-optimal* in the sense that they achieve the smallest possible growth rates of regret (up to logarithmic terms). The performance gap between DDPO-N and DDPO-E is a result of increasing difficulty in the underlying problem rather than the inefficiency or sub-optimality of the DDPO-N policy.

In a recent study, den Boer and Keskin (2020) analyze a pricing problem where the demand-price relationship exhibit discontinuities at unknown price points. The demand shifts that we study may also be viewed as a form of discontinuity, but they occur in the time space instead of the price space. While den Boer and Keskin (2020) also extend their analysis to include change-point detection, they only consider a single change-point. These differences lead to distinct policy design and proof arguments in our work. Another salient difference lies in the demand noise distribution. In den Boer and Keskin (2020), the exact distribution of the noise term is unimportant since it does not interact with the pricing decision, whereas in our setting, the noise term directly interacts with both pricing and inventory decisions, so its exact distribution plays a key role. This imposes additional technical challenges and requires different analysis techniques. In addition, our work allows for non-canonical link functions with much more general noise distributions.

4.2. Case Study on Real-Life Data

In this subsection, we conduct a case study on our policies using the real-life data described in §1.1.1. The data set contains information on orders, inventory, sales, price, and costs across 97 stores with 943 types of

perishable vegetables and fruits in each store on average. According to the supermarket chain manager, the orders are placed at the end of each day and delivered overnight by the next morning with no fixed cost for all supermarket stores. In each store, the pricing and ordering decisions are updated by the store manager on a daily basis, and largely depend on the past experiences of the store manager.

To validate our theoretical analysis and generate additional managerial insights, we analyze a set of products from various stores with a complete list of sales and inventory information throughout the year. For illustration purposes, we first create a test bed by analyzing the ginger data in store A.

4.2.1. Model calibration. We calibrate our model allowing for $\mathscr{C} = 3$ change-points in the demand environment over 365 days.[4] That is, the planning horizon is partitioned into $\mathscr{C} + 1 = 4$ non-overlapping time intervals such that the demand parameters α_t and β_t remain fixed within each interval. To obtain unique parameter estimates and to ensure sufficient sample sizes for estimation in all intervals, we let each interval contain at least 60 days. The ginger data in store A exhibit no stock-outs throughout the year, so the demand is equal to the observed sales. For every possible sequence of change-points, we use MQLE within each interval to obtain demand parameter estimates as well as the corresponding maximal value of quasi-likelihood for that interval. We then add up the 4 quasi-likelihood values to find the total quasi-likelihood for that sequence of change-points. After that, we find the sequence of change-points that give the highest total quasi-likelihood with the corresponding parameter estimates as our calibration result. While fitting the model (1) to the data in each of these intervals, we use various choices for the function $g(\cdot)$, and the exponential demand model, i.e., $g(\cdot) = \exp(\cdot)$, yields the best fit. The above procedure generates the most likely values for the change-points and demand parameters shown in Table 2. All parameter estimates in this table are statistically significant with p-values less than 0.002.

Table 2 Calibrated Demand Parameters for Ginger in Store A*

Dates	Time Intervals	α_t	β_t
Jan 1–Mar 2	$t \in \{1, \ldots, 61\}$	2.2657 (0.2989)	−0.1747 (0.0052)
Mar 3–May 15	$t \in \{62, \ldots, 135\}$	3.9114 (0.1385)	−0.3593 (0.0030)
May 16–Jul 29	$t \in \{136, \ldots, 210\}$	2.4087 (0.2329)	−0.1455 (0.0029)
Jul 30–Dec 31	$t \in \{211, \ldots, 365\}$	2.8284 (0.1371)	−0.1367 (0.0009)

*Values in parentheses indicate standard errors of estimated parameters.

In practice, demand can also be affected by holidays, and we account for this effect in our estimation. We quantify the impact of holidays on demand by using indicator variables for holidays such as Chinese Spring Festival and Christmas. Because our goal is to create a test bed isolated from holiday effects (instead of examining and describing these effects), we drop the indicator variables for holidays and their coefficients

[4] We conduct a robustness study in Appendix B.1 to investigate how the number of change-points affects performance.

from the calibrated demand model. We emphasize that, unlike the unobservable changes in α_t and β_t, the retailer knows when holidays occur. Thus, the inclusion of such indicator variables would not further complicate the model and the theoretical analysis—the same performance guarantees hold in the presence of these indicator variables.

After fitting the demand model, we use the residuals obtained from MQLE to calibrate the demand noise distribution. A Kolmogorov-Smirnov test fails to reject the null hypothesis that the demand noise distribution is normal. Noting that the standard deviation of residuals is 1.3506, we use the normal distribution $\mathcal{N}(0, 1.3506^2)$ as the underlying demand noise distribution. (This normality assumption on demand noise is only used in our case study for the ginger data in store A. Our theoretical results still apply regardless of the assumptions on the demand noise distribution.)

We next deduce the perished quantity from the data as follows: on each day, we add the quantity of new orders to the beginning level of inventory and subtract the sales and the end-of-day inventory. The sequence of perished proportions $\{q_t\}$ is then calculated by dividing the perished quantity by the order-up-to level, i.e., the sum of the beginning inventory and the new orders. The sequence $\{q_t\}$ is fitted to a beta distribution using a constrained MLE to calibrate the perishability parameter $\boldsymbol{\xi} = (\lambda, \nu)$. Since it is computationally challenging to numerically approximate integrals with respect to beta distributions whose parameters are close to 0, we choose 0.4 as the lower bound on λ and ν. The calibrated perishability parameter using this method is $\boldsymbol{\xi} = (0.4, 43.99)$.[5]

The cost parameters are obtained by computing the weighted average costs based on the data, in order to ensure comparability over different planning horizons. Specifically, the unit wholesale price c is the weighted average ordering cost with order quantities used as weights. According to the supermarket chain manager, the holding costs and disposal costs jointly account for approximately 20% of the wholesale price on an annual basis. Hence, we let the unit holding and disposal costs h and w be 10% of the unit wholesale price c, divided by 365 to reflect the daily opportunity cost of excess inventory. We let the unit lost-sales penalty b be the difference of weighted average sales price p and the unit wholesale price c, divided by 365 to match the scale of h and w. Note that this assumption on the unit lost-sales penalty essentially means that if demand is unmet then the supermarket would expect to lose an additional profit margin due to the loss of customer goodwill.[6] Based on this method, the unit cost parameters are given by $c = 8.6215$, $h = 0.0024$, $b = 0.0059$, $w = 0.0024$. We stress that, although the actual profit is sensitive to cost, the general theory does not depend on specific choice of these cost parameters.

[5] In Appendix B.2, we conduct a robustness study to examine the effect of $\boldsymbol{\xi}$ on performance.

[6] This penalty could be lower in some practical settings, and we consider lower values of b in Appendix B.3.

4.2.2. Policy performance. To study the performance of our policies in the setting in §4.2.1, we first compute the FIA policy, the performance benchmark for regret. Figure 3 illustrates the decisions of the FIA policy over a 365-day time horizon, i.e., how this policy dynamically sets the order-up-to levels and prices as functions of the on-hand inventory. Note that it is optimal for the FIA policy to gradually adjust the order-up-to level and price before a change-point because this policy is anticipatory and can take action in response to upcoming change-points.

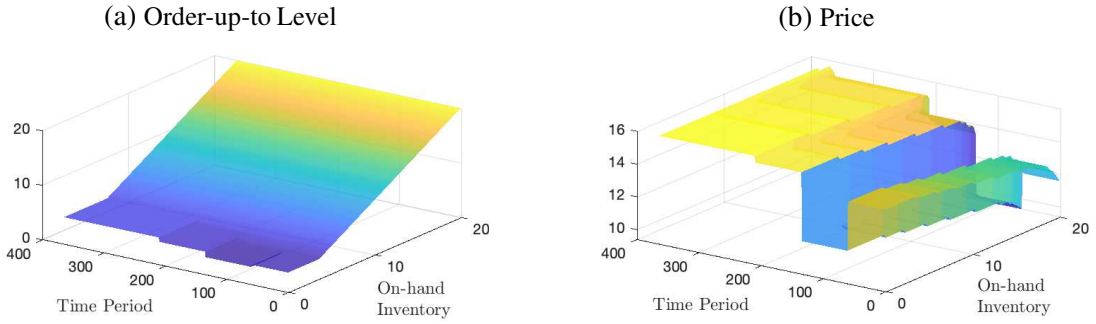

Figure 3 Decisions of the Full-Information Anticipatory Policy over 365 Periods

To examine how the regret of our DDPO policy grows as the number of sales opportunities increases, we consider a sequence of problems indexed by $T = 1, 2, \ldots$, where the change-points in the T^{th} problem are scaled up in proportion to T, given by $\{61, 135, 210\} \times T/365$ rounded to the nearest integers. The number of change-points stays unchanged. The parameter vector $(\eta, \kappa, \omega_1, \omega_2, \upsilon)$ is $(2, 1, 5, 15, 2)$ for DDPO-N and $(2, 4, 5, 15, 2)$ for DDPO-E. These are judiciously selected such that the two test prices are sufficiently apart from each other to estimate the demand parameters in any given subinterval, while enabling the detection of the shifts in the mean demand with reasonable choices of η and κ.

We simulate $S = 2000$ sample paths for the demand noise and perished proportion sequences. For each sample path and $T \in \{2000, 4000, \ldots, 20000\}$, we calculate the regret of DDPO (i.e., the profit loss of DDPO relative to FIA) over T periods. The regret $\Delta^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}(T)$ is thus approximated by

$$\widehat{\Delta}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}(T) = \frac{1}{S}\sum_{s=1}^{S}\sum_{t=1}^{T}\left[Q(p_t^{*(s)}, y_t^{*(s)}; \boldsymbol{\theta}_t, \boldsymbol{\xi}) - Q(\hat{p}_t^{(s)}, \hat{y}_t^{(s)}; \boldsymbol{\theta}_t, \boldsymbol{\xi})\right], \tag{20}$$

where $\{(p_t^{*(s)}, y_t^{*(s)}) : t = 1, \ldots, T\}$ and $\{(\hat{p}_t^{(s)}, \hat{y}_t^{(s)}) : t = 1, \ldots, T\}$ are the decision sequences of the FIA and DDPO policies, respectively, on the s^{th} sample path. Figure 4 shows the regret growth under DDPO-N and DDPO-E. A non-linear fit reveals that the T-period regret is $O(T^{0.66}(\log T)^{1/2})$ under DDPO-N, and $O(T^{0.51}\log T)$ under DDPO-E. This confirms our theoretical results that the T-period regret is $O(T^{2/3}(\log T)^{1/2})$ under DDPO-N, and $O(T^{1/2}\log T)$ under DDPO-E.[7]

[7] DDPO-N and DDPO-E use only the data from the experimentation periods. In a robustness study, we examine modified versions of these policies that use the data from all periods—see Appendix B.4.

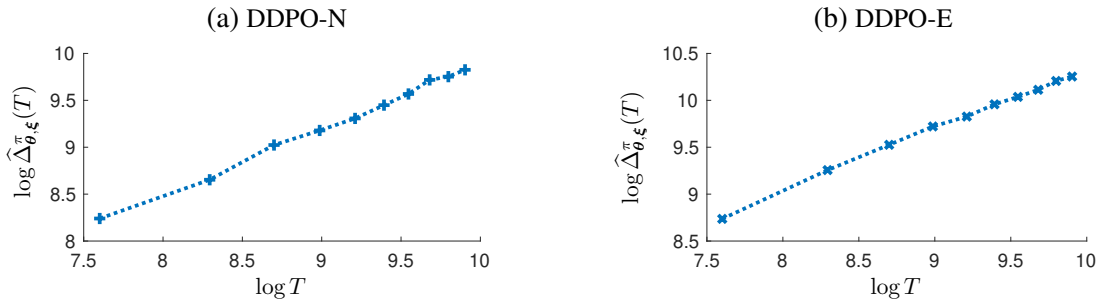

Figure 4 Growth Rate of Regret

4.2.3. DDPO versus the company policy. We compare our policies with the historical decisions of the supermarket in terms of regret, i.e., the profit loss relative to the FIA policy, based on the demand residuals and product perishability rates realized in the data set. The cost parameters are the same as in §4.2.1. (We conduct robustness checks on the values of h, b, and w in Appendix B.3.) Over 365 days, we apply both versions of our DDPO policy to 33 product-store pairs with complete sales and inventory data—ranging over ginger, papaya, dragon fruit, potato, tomato, and red onion in 25 different stores. Compared to the supermarket's decisions across the 33 product-store pairs, DDPO-N and DDPO-E have 86.6% and 80.5% less regret on average with standard deviations 10.2% and 16.0%, respectively. For the distributions of these regret savings, see Appendix B.5. This comparison provides strong evidence that implementing our DDPO policies could generate significantly higher profits.

4.2.4. Changing environment versus product perishability. To measure the impact of the changing demand environment and product perishability, we compare the profit of our DDPO policy with three variants over 365 days, based on the sample paths realized in the aforementioned 33 product-store pairs. Table 3 summarizes the average annual profits with corresponding ranges across these 33 product-store pairs for DDPO-N and DDPO-E. In this table, the case that has "Yes" for both the changing environment and the perishability corresponds to accounting for both modeling features, i.e., using the DDPO policy in the base setting. A "Yes" for the perishability and a "No" for the changing environment indicate a modified policy that takes only the perishability into account (ignoring the changes in the demand parameters by setting $\eta = \infty$). A "Yes" for the changing environment and a "No" for the perishability indicate another modified policy that takes only the changing environment into account (ignoring the perishability of products by setting $q_t = 0$ for all t). The case with "No" for both the changing environment and the perishability corresponds a policy that ignores both modeling features. The cost parameters are the same as in §4.2.1. We observe that ignoring changing environment results in a profit loss of 14.1% for DDPO-N and 10.2% for DDPO-E. Ignoring perishability leads to a profit loss of 3.4% for DDPO-N and 1.8% for DDPO-E. The annual gross profit of a comparable Chinese supermarket chain was about 4 billion RMB in 2013; this is approximately 612.6 million U.S. dollars.[8] For such a supermarket, the results in Table 3 translate to an

[8]The figure in U.S. dollars is based on the exchange rate of 6.53 RMB/USD.

annual profit loss of over 62 million U.S. dollars due to ignoring changing environment and an annual profit loss of over 11 million U.S. dollars due to ignoring product perishability.

Table 3 Impact of Accounting for Changing Environment and Inventory Perishability on Profits*

(a) Average Annual Profit of DDPO-N (RMB)

Account for Changing Environment	Account for Inventory Perishability: Yes	Account for Inventory Perishability: No
Yes	14980 [733, 59297]	14470 [726, 58817]
No	12873 [733, 44948]	12372 [726, 44555]

(b) Average Annual Profit of DDPO-E (RMB)

Account for Changing Environment	Account for Inventory Perishability: Yes	Account for Inventory Perishability: No
Yes	13351 [560, 47298]	13113 [537, 46877]
No	11986 [560, 40655]	11707 [537, 40468]

*Values in square brackets indicate ranges over 33 product-store pairs.

To further demonstrate the relative significance of the changing environment and perishability, we compare the regret caused by ignoring only the changing environment with the regret caused by ignoring only the perishability over 365 days, using the 33 product-store pairs described earlier. For each product-store pair, the demand parameter sequence $\boldsymbol{\theta}$ is calibrated using the exponential demand model with 3 change-points that are at least 60 days apart, and the product perishability parameter $\boldsymbol{\xi}$ is calibrated using the perished proportions $\{q_t\}$ induced from the real-life data as in §4.2.1. The regret is then calculated based on 100 independent samples generated from the calibrated noise and perishability distributions.

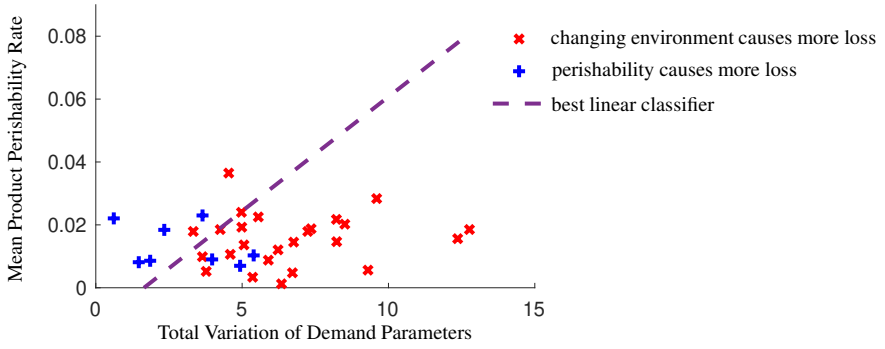

Figure 5 Changing Demand Environment versus Product Perishability

Figure 5 illustrates the comparison results for DDPO-E. Each point in the figure corresponds to a product-store pair. The horizontal axis shows the amount of change in the demand environment, measured by the total variation of the calibrated demand parameter sequence $\boldsymbol{\theta}$. The vertical axis shows the mean of the perished proportion distribution with the calibrated parameter $\boldsymbol{\xi}$. Each red cross (✖) is a product-store pair for which ignoring the changing environment is more costly, whereas each blue plus (✚) is a product-store pair for which ignoring the perishability is more costly. To find the best linear classification of the red cross and blue plus signs, we use a support vector machine classifier (see the dashed line). This classifier can be viewed as an "indifference curve" regarding the losses caused by the changing environment and product

perishability. The slope of the classifier is 0.0071; hence, on the indifference curve, one unit of increase in the total variation of demand parameters is as important as an increase of 0.0071 in the mean product perishability rate. This provides a practical guideline in managing the two challenges in a more cost efficient way. If the demand of a certain product in a store exhibits increasing volatility compared to the mean perishability rate (i.e., the product-store pair is to the right of the dashed line in Figure 5), then the store manager should consider advanced data-driven analytics to closely monitor and detect potential demand changes. Otherwise, the store manager should primarily focus on the inventory management, fine-tuning the order quantity and frequency.

4.2.5. Choice of the detection threshold parameter. This subsection studies the effect of the detection threshold parameter η on regret. If η is small, the DDPO policy claims detections too frequently, and if η is large, it can fail to detect the changes in demand parameters. Using the ginger data in store A, we first calculate the regret of the DDPO policy over $T = 2000$ periods, for the choices of $\eta \in \{0, 0.5, \ldots, 10\}$ based on 500 independent sample paths. Figure 6 shows the resulting regret as a function of η for both versions of the DDPO policy (with slight abuse of notation, we denote the sample-average regret in (20) as $\widehat{\Delta}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}(T;\eta)$ to express its dependence on η explicitly). Choosing $\eta = 2$ yields the smallest regret for both policies, and this is exactly the value of η we choose in the base setting in §4.2.2. Choosing other values of η leads to considerably larger regret, which is less sensitive to the value of η outside the range $[1.5, 2.5]$. Choosing η too large incurs slightly larger regret than choosing η too small.

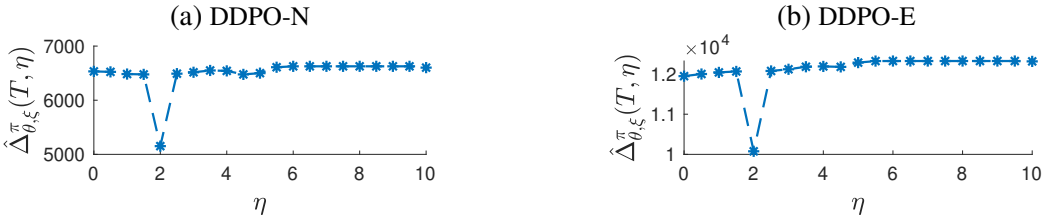

Figure 6 Effect of the Detection Threshold on Regret

We then examine the growth rate of regret over $T \in \{2000, 4000, \ldots, 20000\}$, for the choices of $\eta \in \{0, 0.5, \ldots, 9.5, 10\}$. The resulting regret is $O(T^{0.67}(\log T)^{1/2})$ under DDPO-N and $O(T^{0.51}\log T)$ under DDPO-E, and is stable for $\eta \in \{0, 0.5, \ldots, 4\}$. There is a slight increase in the growth rate as η exceeds 4, making the regret $O(T^{0.71}(\log T)^{1/2})$ under DDPO-N and $O(T^{0.55}\log T)$ under DDPO-E, which is also stable for $\eta \in \{4.5, 5, \ldots, 10\}$. This indicates that for the growth rate of regret, failing to detect potential change-points is more harmful than declaring detections too frequently.

5. Theoretical Analysis

In this section, we present a thorough theoretical analysis on the performance of the DDPO policy. Because our performance benchmark, the FIA policy, requires solving a dynamic program that is less tractable, we facilitate our analysis by introducing a connecting problem where a myopic policy endowed with substantial informational advantages is optimal—we call this policy the *full-information non-anticipatory* (FINA)

policy. We first bound the regret of FINA relative to FIA. We then create several other connecting problems between FINA and DDPO, and provide bounds on the regret contributed by each connecting step to ultimately obtain the theoretical performance guarantees. Figure 7 illustrates this process. Here, π denotes a generic policy whose input parameters are given in parentheses. Recall that the underlying problem parameters are $\boldsymbol{\theta}$, $\boldsymbol{\xi}$ and F_ε. In addition to these, we let $\boldsymbol{\tau}^*$ denote the collection of cycles containing the change-points in the underlying model—note that $\boldsymbol{\tau}^*$ can be deduced from $\boldsymbol{\theta}$, but we use it as an explicit parameter to emphasize the impact of change-point detection. As mentioned in §3, $\boldsymbol{\chi}$ is the vector of claimed detection cycles and $\mathbf{e}$ is the vector of residuals obtained from MQLE under the DDPO policy. All other notations featuring a hat in Figure 7 are estimates of the corresponding underlying parameters. Finally, the dotted rectangle highlights the source of the performance gap between DDPO-N and DDPO-E in Settings N and E, respectively.

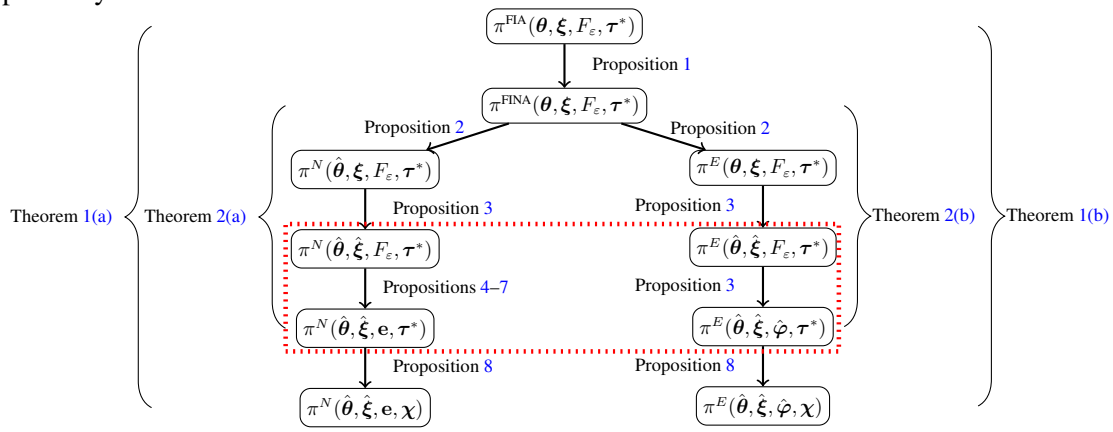

Figure 7 Diagram of Theoretical Results

5.1. Full-Information Non-Anticipatory (FINA) Policy

The FINA policy is obtained by maximizing the single-period profit $Q(p_t, y_t; \boldsymbol{\theta}_t, \boldsymbol{\xi})$ in (4) with respect to p_t and y_t in each period t. Considering the full-information scenario where the retailer knows $\boldsymbol{\theta}_t$, $\boldsymbol{\xi}$ and F_ε, we can write the maximization of $Q(p_t, y_t; \boldsymbol{\theta}_t, \boldsymbol{\xi})$ in an iterative manner by first choosing y_t to minimize the single-period cost $H(y_t; p_t, \boldsymbol{\theta}_t, \boldsymbol{\xi})$ given p_t, and then choosing a profit-maximizing p_t. Thus,

$$\max_{\substack{(p_t,y_t)\in\mathscr{P}\times\mathscr{Y}\\ y_t\geq x_t}} Q(p_t, y_t; \boldsymbol{\theta}_t, \boldsymbol{\xi}) = \max_{p_t\in\mathscr{P}} \underbrace{\left\{ p_t g(\mathbf{X}_t^\top \boldsymbol{\theta}_t) - \min_{\substack{y_t\in\mathscr{Y}\\ y_t\geq x_t}} H(y_t; p_t, \boldsymbol{\theta}_t, \boldsymbol{\xi}) \right\}}_{\equiv G(p_t;\boldsymbol{\theta}_t,\boldsymbol{\xi})}. \tag{21}$$

Denote by $(\check{p}_t, \check{y}_t)$ the optimal solution to (21) for period t. Implementing this solution yields the FINA policy because the retailer has full information (i.e., knows $\boldsymbol{\theta}_t$, $\boldsymbol{\xi}$ and F_ε) in each period t, but is non-anticipatory (i.e., cannot foresee the change-points before they occur). We denote the FINA policy by $\check{\pi}$, which maps the state $\mathbf{x}$ to $\{(\check{p}_t, \check{y}_t) : t = 1, \ldots, T\}$. We also denote by (p_t^u, y_t^u) the unconstrained optimizer of (21), and let $G^u(p_t; \boldsymbol{\theta}_t, \boldsymbol{\xi}) = Q(p_t, y_t^u; \boldsymbol{\theta}_t, \boldsymbol{\xi})$. This is a base-stock-list-price policy with $\check{p}_t = p_t^u$ and $\check{y}_t = \max\{y_t^u, x_t\}$, where the base stock level y_t^u is given by a newsvendor-type solution. We now provide an upper bound on the regret of FINA relative to FIA.

PROPOSITION 1. *There exists a positive constant $\tilde{N}$ such that*

$$\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{t=1}^{T}[Q(p_t^*,y_t^*;\boldsymbol{\theta}_t,\boldsymbol{\xi})-Q(\check{p}_t,\check{y}_t;\boldsymbol{\theta}_t,\boldsymbol{\xi})]\right]\leq 2\overline{Q'}\,\overline{d}(\mathscr{P}\times\mathscr{Y})(\tilde{N}+1)(\mathscr{C}+1) \tag{22}$$

for all $T=1,2,\ldots$, where $\overline{Q'}=\max\{\|\nabla Q(p,y)\|:p\in\mathscr{P},y\in\mathscr{Y}\}$ and $\overline{d}(\mathscr{P}\times\mathscr{Y})$ is the maximum distance between two points in $\mathscr{P}\times\mathscr{Y}$.

This proposition implies that the regret difference between FIA and FINA is $O(1)$, that is, bounded above by a constant independent of time horizon T. Recall that the only difference between them is that, under FIA, it is possible to anticipate the change-points so as to gradually adjust pricing and inventory decisions before each change-point. Under FINA, however, such adjustments in pricing and inventory decisions can occur only after the change-points. Therefore, the two policies can only differ within a finite number of periods around change-points in expectation. (Note that this argument can also be used for deriving a bound on regret due to missing the inventory target when the on-hand inventory level x_t is greater than the order-up-to level y_t in the DDPO policy.)

5.2. Regret due to Estimation Errors

We now proceed to assess the performance of the DDPO policy relative to the FINA policy. Because the DDPO-N policy makes no assumption on the demand noise distribution and hence is more general than the DDPO-E policy, we focus on the analysis of DDPO-N and highlight the main differences between the two versions at the end. As mentioned earlier, to derive an upper bound on regret incurred due to not knowing $\boldsymbol{\theta}$, $\boldsymbol{\xi}$ and F_ε, we examine a chain of connecting problems that differ from each other at only one point.

5.2.1. Estimation error for the demand parameter vector. To facilitate the exposition, we let the first and second claimed detection cycles following cycle τ_j^* be $\hat{\tau}_j^+=\inf\{\tau>\tau_j^*:\chi_\tau=1\}\wedge\tau_{j+1}^*$ and $\hat{\tau}_j^-=\inf\{\tau>\hat{\tau}_j^+:\chi_\tau=1\}\wedge\tau_{j+1}^*$, respectively, for $j=0,1,\ldots,\mathscr{C}$, with $\hat{\tau}_0^+=0$. The following proposition characterizes the squared estimation error between a cycle containing a true change-point and the second claimed detection cycle after it.

PROPOSITION 2. *For any period $t\notin\mathscr{X}$ in cycle τ with $\tau_j^*<\tau<\hat{\tau}_j^-$, there exist positive constants K_3 and K_4 such that*

$$\mathbb{P}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left\{\|\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t\|^2\geq\frac{K_3\log M_t}{M_t}\right\}\leq\frac{K_4}{M_t}\,. \tag{23}$$

Proposition 2 states that the convergence rate of the squared demand parameter estimation error is of order $(\log M_t)/M_t$, where $M_t=(\lceil t/n\rceil-L(\tau))m$ is the effective sample size in period t. In DDPO-N, m and n are of order $T^{1/3}$ and $T^{2/3}$, respectively. To establish this result, we construct a proof argument based on using an exponential supermartingale to derive an upper bound on the probability that the distance between the estimator $\hat{\boldsymbol{\theta}}_t$ and the true parameter $\boldsymbol{\theta}_t$ is large.

5.2.2. Estimation error for the product perishability parameter. Recall that the product perishability follows the beta distribution with parameter $\boldsymbol{\xi}$, which falls into the exponential family. To quantify the estimation error due to not knowing the underlying inventory perishability parameter vector $\boldsymbol{\xi}$, we derive a general result that holds for the exponential family of distributions.

For a generic random variable Z belonging to the exponential family of distributions, the density of Z is given by $f_Z(z;\boldsymbol{\phi}) = B(z)\exp[\boldsymbol{\phi}^\intercal \mathbf{T}(z) - A(\boldsymbol{\phi})]$, where $\boldsymbol{\phi}$ is a d-dimensional parameter vector located in a compact set Φ. The following proposition establishes the consistency as well as the convergence rate of MLE for the parameter vector of the exponential family of distributions.

PROPOSITION 3. *Assume that* $\sup_{\boldsymbol{\phi}\in\Phi} \mathbb{E}_{\boldsymbol{\phi}} \|\nabla_{\boldsymbol{\phi}} \log f_z(z;\boldsymbol{\phi})\|^\ell < \infty$ *for some* $\ell > d$. *Then for any* $\delta > 0$ *there exist positive constants* K_5 *and* K_6 *such that*

$$\mathbb{P}_{\boldsymbol{\phi}} \left\{ \|\hat{\boldsymbol{\phi}}_k - \boldsymbol{\phi}\|^2 \geq \frac{K_5 \log k}{k} \right\} \leq \frac{K_6}{k}, \tag{24}$$

where $\hat{\boldsymbol{\phi}}_k$ *is the projection onto* Φ *of the maximum likelihood estimator of* $\boldsymbol{\phi}$, *based on* k *i.i.d. observations from* $f_Z(z;\boldsymbol{\phi})$, *and* $\mathbb{P}_{\boldsymbol{\phi}}\{z_1 \in \mathrm{d}\tilde{z}_1, \ldots, z_k \in \mathrm{d}\tilde{z}_k\} = \prod_{i=1}^k f_Z(\tilde{z}_i;\boldsymbol{\phi})\mathrm{d}\tilde{z}_i$.

Proposition 3 states that the maximum likelihood estimator based on k i.i.d. observations converges to the underlying parameter vector in probability at rate $\sqrt{(\log k)/k}$. The assumption in the proposition is a regularity condition used to justify the exchange of limit operation and integration, which is satisfied by the exponential family of distributions. (Note that this result can also be used for quantifying the estimation error for the noise distribution parameter $\boldsymbol{\varphi}$ under DDPO-E.)

5.2.3. Estimation error for the demand noise distribution. To compare (4) and (14), we introduce the following connecting problem:

$$\max_{(p_t,y_t)\in\mathscr{P}\times\mathscr{Y}} \tilde{Q}_t(p_t,y_t;\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon}) = \max_{p_t\in\mathscr{P}} \Bigg\{ \underbrace{p_t g(\mathbf{X}_t^\intercal\boldsymbol{\theta}_t) - \min_{y_t\in\mathscr{Y}} \tilde{H}_t(y_t;p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon})}_{\equiv \tilde{G}_t(p_t;\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon})} \Bigg\}, \tag{25}$$

where

$$\begin{aligned} \tilde{H}_t(y_t;p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon}) = \frac{1}{2M_t} \sum_{\substack{s=nL(\tau)+1 \\ s\in\mathscr{X}}}^{t} \mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t} \Big[& (h-c)[(1-q_t)y_t - g(\mathbf{X}_t^\intercal\boldsymbol{\theta}_t) - \varepsilon_s]^+ \\ & + (b+p_t)[g(\mathbf{X}_t^\intercal\boldsymbol{\theta}_t) + \varepsilon_s - (1-q_t)y_t]^+ \Big] + \big(w\mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[q_t] + c\big)y_t. \end{aligned} \tag{26}$$

Note that the residuals $\{e_s\}$ and the parameter estimate vector $\hat{\boldsymbol{\theta}}_t$ in (15) are replaced by the unobservable noise terms $\{\varepsilon_s\}$ and the underlying parameter vector $\boldsymbol{\theta}_t$ in (26), respectively. The following result provides a theoretical bound on the difference between the single-period profit functions $G^u(p_t;\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)$ and $\tilde{G}_t(p_t;\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon})$.

PROPOSITION 4. *Given* $p_t \in \mathscr{P}$ *in period* $t \notin \mathscr{X}$, *there exist positive constants* K_7 *and* K_8 *such that*

$$\mathbb{P}_{\varepsilon} \left\{ \left| G^u(p_t;\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t) - \tilde{G}_t(p_t;\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon}) \right| \geq K_7 \sqrt{\frac{\log M_t}{M_t}} \right\} \leq \frac{K_8}{M_t}. \tag{27}$$

Because $G^u(p_t;\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)$ involves the expectation over the true demand noise distribution while $\tilde{G}_t(p_t;\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon})$ uses a sample average based on the unobservable demand noise terms $\boldsymbol{\varepsilon}$, Proposition 4 states that the estimation error due to not knowing the functional form of the underlying noise distribution decays at a rate proportional to $\sqrt{(\log M_t)/M_t}$.

The next result applies Proposition 2, leading to a probabilistic error bound on the single-period profit due to replacing the unobservable noise terms $\boldsymbol{\varepsilon}$ by the observed residuals $\mathbf{e}$.

PROPOSITION 5. *For any given $p_t \in \mathscr{P}$ in period $t \notin \mathscr{X}$ of cycle $\tau \in [\hat{\tau}_j^+, \hat{\tau}_j^-)$, there exist positive constants K_9 and K_{10} such that*

$$\mathbb{P}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left\{\left|\tilde{G}_t(p_t;\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon})-\hat{G}_t(p_t;\hat{\boldsymbol{\theta}}_t,\hat{\boldsymbol{\xi}}_t,\mathbf{e})\right| \geq K_9\sqrt{\frac{\log M_t}{M_t}}\right\} \leq \frac{K_{10}}{M_t}\,. \tag{28}$$

Combining the results of Proposition 4 and 5, we see that the convergence rate of the proxy profit function $\hat{G}_t(\cdot)$ to the true profit function $G^u(\cdot)$ at any price $p_t \in \mathscr{P}$ is in the order of $\sqrt{(\log M_t)/M_t}$ for the DDPO-N policy—this stands in contrast to the convergence rate of the squared estimation error of demand parameters, which is in the order of $(\log M_t)/M_t$. This result is a direct implication of the fact that $\hat{G}_t(\cdot)$ is a non-differentiable Lipschitz function in $\boldsymbol{\theta}_t$, which is a key factor leading to a slower convergence rate of approximation under the DDPO-N policy. On the other hand, the DDPO-E policy can take advantage of the differentiability of $G^u(\cdot)$ because the density of the noise distribution is known to the retailer in that case. Consequently, this is exactly the point where the performances of the two DDPO policies take a bifurcation (recall Figure 7).

5.2.4. Convergence of pricing and inventory decisions. Based on all the above results, we see that the two functions $G^u(\cdot)$ and $\hat{G}_t(\cdot)$ are "close" at each point with high probability. However, to establish the convergence of optimal controls, we need *uniform* closeness with high probability because otherwise the optimizers of $G^u(\cdot)$ and $\hat{G}_t(\cdot)$ can be far away with a non-diminishing probability. Recall that we choose a sparse price grid with step size ι_t to balance (i) the accuracy of estimating the noise distribution using sample average approximation and (ii) the accuracy of price optimization over the sparse grid. We show in the following result that the sparsity of our price grid does not slow down the convergence rate of approximation.

PROPOSITION 6. *For the DDPO-N policy, there exist positive constants K_{11} and K_{12} such that, for any period $t \notin \mathscr{X}$,*

$$\mathbb{P}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left\{|p_t^u-\hat{p}_t| \geq K_{11}\sqrt{\frac{\log M_t}{M_t}}\right\} \leq \frac{K_{12}}{\sqrt{M_t \log M_t}}. \tag{29}$$

Proposition 6 states that the absolute pricing error converges to zero at a rate of $\sqrt{(\log M_t)/M_t}$ for the DDPO-N policy—this is the essential factor in determining the regret of DDPO-N. As alluded to earlier, this convergence rate is a direct consequence of the step size choice $\iota_t = \rho\sqrt{(\log M_t)/M_t}$ for the price grid

$\mathscr{P}_d$, which matches the $O(\sqrt{(\log M_t)/M_t})$ rate inside the probability in (29) with the $O(1/\sqrt{M_t \log M_t})$ rate outside the probability in (29), up to a logarithmic term.

The next proposition characterizes the convergence of inventory decisions of DDPO-N. It shows that the inventory policy also converges at a rate of $O(\sqrt{(\log M_t)/M_t})$.

PROPOSITION 7. *For the DDPO-N policy, there exist positive constants K_{13} and K_{14} such that, for any period $t \notin \mathscr{X}$,*

$$\mathbb{P}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left\{|y^u_t(\hat{p}_t) - \hat{y}^u_t| \geq K_{13}\sqrt{\frac{\log M_t}{M_t}}\right\} \leq \frac{K_{14}}{M_t}, \tag{30}$$

where $y^u_t(\hat{p}_t)$ is the unconstrained optimum of $H(y_t; \hat{p}_t, \hat{\boldsymbol{\theta}}_t, \hat{\boldsymbol{\xi}}_t)$.

We now conclude this subsection by deriving an upper bound on the T-period regret due to estimation error under both versions of our DDPO policy. As discussed above, the regret due to estimation error under the DDPO-N policy is largely determined by the slower convergence rate of pricing decisions as well as the fact that the function $Q(p, y)$ is Lipschitz but non-differentiable. Hence, combining Propositions 6 and 7 yields the result for DDPO-N. On the contrary, the pricing and ordering decisions of the DDPO-E policy converge at essentially the same rate as that of the parameter estimates. Consequently, the regret due to estimation error under DDPO-E is given directly by Propositions 2 and 3. The following theorem summarizes these results.

THEOREM 2. *(a) For the DDPO-N policy, there exists a positive constant K_{15} such that*

$$\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{j=0}^{\mathscr{C}}\sum_{t=n\hat{\tau}_j^+ +1}^{n\hat{\tau}_j^-}\left[Q(p^u_t, y^u_t; \boldsymbol{\theta}_t, \boldsymbol{\xi}) - Q(\hat{p}_t, \hat{y}^u_t; \boldsymbol{\theta}_t, \boldsymbol{\xi})\right]\mathbb{I}\{t \notin \mathscr{X}\}\right] \leq K_{15}\, T^{2/3}(\log T)^{1/2} \text{ for } T = 3, 4, \ldots$$

(b) For the DDPO-E policy, there exists a positive constant K_{16} such that

$$\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{j=0}^{\mathscr{C}}\sum_{t=n\hat{\tau}_j^+ +1}^{n\hat{\tau}_j^-}\left[Q(p^u_t, y^u_t; \boldsymbol{\theta}_t, \boldsymbol{\xi}, \boldsymbol{\varphi}) - Q(\hat{p}_t, \hat{y}^u_t; \boldsymbol{\theta}_t, \boldsymbol{\xi}, \boldsymbol{\varphi})\right]\mathbb{I}\{t \notin \mathscr{X}\}\right] \leq K_{16}\, T^{1/2}\log T \text{ for } T = 3, 4, \ldots$$

5.3. Regret Due to Detection Errors

In this subsection, we examine the regret due to detection errors, which are composed of two parts: late detections of a true change-point and false alarms when no change occurs. The following proposition provides a theoretical upper bound on the cumulative impact of these detection errors.

PROPOSITION 8. *(a) (Delay of True Detections) For the DDPO-N policy, there exists a positive constant K_{17} such that*

$$\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{j=0}^{\mathscr{C}}\sum_{t=n\tau_j^* +1}^{n\hat{\tau}_j^+}\left[Q(p^u_t, y^u_t; \boldsymbol{\theta}_t, \boldsymbol{\xi}) - Q(\hat{p}_t, \hat{y}^u_t; \boldsymbol{\theta}_t, \boldsymbol{\xi})\right]\mathbb{I}\{t \notin \mathscr{X}\}\right] \leq K_{17}\, T^{2/3}(\log T)^{1/2} \text{ for } T = 3, 4, \ldots$$

For the DDPO-E policy, there exists a positive constant K_{18} such that

$$\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{j=0}^{\mathscr{C}}\sum_{t=n\tau_j^*+1}^{n\hat{\tau}_j^+}\left[Q(p_t^u,y_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi},\boldsymbol{\varphi})-Q(\hat{p}_t,\hat{y}_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi},\boldsymbol{\varphi})\right]\mathbb{I}\{t\notin\mathscr{X}\}\right]\leq K_{18}\,T^{1/2}\log T \;\; \textit{for } T=3,4,\ldots$$

(b) (Early False Alarms) For the DDPO-N policy, there exists a positive constant K_{19} such that

$$\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{j=0}^{\mathscr{C}}\sum_{t=\hat{\tau}_j^-+1}^{n\tau_{j+1}^*}\left[Q(p_t^u,y_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi})-Q(\hat{p}_t,\hat{y}_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi})\right]\mathbb{I}\{t\notin\mathscr{X}\}\right]\leq K_{19}\,T^{1/2} \;\; \textit{for } T=3,4,\ldots$$

For the DDPO-E policy, there exists a positive constant K_{20} such that

$$\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{j=0}^{\mathscr{C}}\sum_{t=\hat{\tau}_j^-+1}^{n\tau_{j+1}^*}\left[Q(p_t^u,y_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi},\boldsymbol{\varphi})-Q(\hat{p}_t,\hat{y}_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi},\boldsymbol{\varphi})\right]\mathbb{I}\{t\notin\mathscr{X}\}\right]\leq K_{20}\,T^{1/2} \;\; \textit{for } T=3,4,\ldots$$

It is possible to claim a detection several cycles later than a true change-point and also have multiple early false alarms before a true change-point. Proposition 8 shows that the regret due to such detection errors grows sub-linearly in T for both DDPO policies. The difference between the upper bounds in Proposition 8(a) results from the different convergence rates of decision variables: for the DDPO-N policy, the regret due to the delay of true detections is determined by the *absolute* convergence rate of pricing decisions, while for the DDPO-E policy, this regret is determined by the *squared* convergence of pricing and ordering decisions. In Proposition 8(b), the detection error due to early false alarms reduces to the difference of mean demand noises because there are no change-points between τ_j^* and τ_{j+1}^*. As a result, this component of the regret does not depend on the convergence of decision variables or parameter estimates. We also note that the constants in the preceding upper bounds depend polynomially on the number of change-points, $\mathscr{C}$. Thus, the growth rate of regret does not significantly change as long as $\mathscr{C}$ does not increase with T faster than logarithmically.

Combining Proposition 1, Theorem 2, and Proposition 8 yields the main results in Theorem 1.

6. Extensions

In this section, we discuss two extensions to our base model: (1) age-dependent product perishability and (2) demand censoring.

6.1. Age-dependent Product Perishability

In this subsection, we set up a model that accommodates age-dependent product perishability. To demonstrate the regret performance of the DDPO policy, we formulate the full-information problem for this model, and calibrate the model using the real-life data set described in §1.1.1. Because solving a dynamic program with a large state space is computationally challenging due to the curse of dimensionality, we consider the case where there are three possible states of the product: freshest (age 1), moderately fresh (age 2), and perished (age 3). The state transitions follow a Markov chain: products of age 1 can become age 2 with

probability q_1 or stay in age 1 with probability $1-q_1$ in the next period; similarly, products of age 2 can become perished (age 3) with probability q_2 or stay in age 2 with probability $1-q_2$ in the next period. Perished products are disposed of at the end of each period, so the overall state of the system is determined by the quantities of age-1 and age-2 products. Note that age 3 is an absorbing state and that age-1 products cannot become age-3 products within a single period.

In each period $t=1,2,\ldots,T$, events occur in the following sequence:

1. The retailer observes the on-hand inventory levels $x_{t,1}$ and $x_{t,2}$ of age-1 and age-2 products, respectively.
2. The retailer chooses a price $p_t \in \mathscr{P} = [p_{\min}, p_{\max}]$ and an order-up-to level $y_t \in \mathscr{Y} = [y_{\min}, y_{\max}]$, where $0 < p_{\min} < p_{\max} < \infty$ and $0 < y_{\min} < y_{\max} < \infty$.
3. The retailer receives the replenishment order, which consists of $y_t - (x_{t,1}+x_{t,2})$ units of freshest (age-1) products. By the time the products are put on shelves for sale, a proportion $q_{t,1}$ of the age-1 products deteriorate to age-2, and a proportion $q_{t,2}$ of the age-2 products become perished. We assume that $\{q_{t,1} : t=1,2,\ldots,T\}$ and $\{q_{t,2} : t=1,2,\ldots,T\}$ are independent sequences of i.i.d. random variables following the beta distributions with parameters $\boldsymbol{\xi}_1 = (\lambda_1, \nu_1)$ and $\boldsymbol{\xi}_2 = (\lambda_2, \nu_2)$, respectively.
4. The demand D_t is realized and satisfied to the maximum extent by the remaining on-hand inventory. We assume that consumers always purchase products of younger ages until they are depleted and do not purchase perished goods. As a result, it is appropriate to use the Last-In-First-Out (LIFO) inventory issuing policy.
5. The end-of-period inventory is updated as follows:

$$x_{t+1,1} = [(y_t - x_{t,2})(1-q_{t,1}) - D_t]^+, \tag{31}$$

$$\begin{aligned} x_{t+1,2} &= [(1-q_{t,2})x_{t,2} + q_{t,1}(y_t - x_{t,2}) - (D_t - (y_t - x_{t,2})(1-q_{t,1}))^+]^+ \\ &= [(1-q_{t,2})x_{t,2} + q_{t,1}(y_t - x_{t,2}) - (D_t - (y_t - x_{t,2})(1-q_{t,1})) - (D_t - (y_t - x_{t,2})(1-q_{t,1}))^-]^+ \\ &= [(1-q_{t,2})x_{t,2} + (y_t - x_{t,2}) - D_t - ((y_t - x_{t,2})(1-q_{t,1}) - D_t)^+]^+ \\ &= [y_t - q_{t,2}x_{t,2} - D_t - x_{t+1,1}]^+. \end{aligned} \tag{32}$$

We now describe our calibration method. As we do not observe the products of age 2 or the corresponding deteriorated quantities from age 1 to age 2 in each period, we suppose that the sequence $\{q_{t,1} : t = 1,2,\ldots,T\}$ follows a given beta distribution with parameter $\boldsymbol{\xi}_1$, and conduct robustness studies with various different values of $\boldsymbol{\xi}_1$ (see Table 5 for details). Table 4 shows an example illustrating the calibration method and system dynamics. In this table, O_t denotes the order quantity, $(E_{t,1}, E_{t,2})$ the deteriorated quantities of age-1 and age-2 products, and $(S_{t,1}, S_{t,2})$ the quantities of age-1 and age-2 products that are available for sale in period t. The numbers in red and bold font are either directly observed or calculated using information from the data set. The numbers in blue and italic font are deduced based on the system dynamics, the Markov transition probabilities, and the value of $\boldsymbol{\xi}_1$ using (31) and (32).

In period $t=1$, $q_{t,1}$ is simulated from the beta distribution with the chosen parameter $\boldsymbol{\xi}_1$. Then, the deteriorated quantity of age-1 products is $E_{t,1} = (x_{t,1} + O_t)q_{t,1} = 2$, which becomes part of the age-2

Table 4 Calibration of Age-dependent Perishability Rate

t	$(x_{t,1},x_{t,2})$	O_t	$q_{t,1}$	$(E_{t,1},E_{t,2})$	$(S_{t,1},S_{t,2})$	D_t	$(x_{t+1,1},x_{t+1,2})$	$q_{t,2}$
1	$(0,0)$	**50**	*0.04*	$(2,\mathbf{0})$	$(48,2)$	**20**	$(28,2)$	*0*
2	$(28,2)$	**12**	*0.01*	$(0.4,\mathbf{0.1})$	$(39.6,2.3)$	**40**	$(0,1.9)$	*0.05*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

products available for sale in this period. The perished quantity $E_{t,2}=0$ is calculated using the real-life data set as in the base setting, where products of ages 1 and 2 are collapsed into a single state. As a result, the quantities that are available for sale are $S_{t,1}=(x_{t,1}+O_t)(1-q_{t,1})=48$ for age-1 products, and $S_{t,2}=x_{t,2}-E_{t,2}+E_{t,1}=2$ for age-2 products. Because $E_{t,2}=0$, the perished proportion $q_{t,2}$ is simply 0. Satisfying the demand and proceeding to period $t=2$ in a similar fashion, we deduce that $S_{t,1}=(x_{t,1}+O_t)(1-q_{t,1})=39.6$ and $S_{t,2}=x_{t,2}-E_{t,2}+E_{t,1}=2.3$. Because the 40 units of demand exceed the available age-1 products ($S_{t,1}=39.6<D_t=40$) and inventories are issued according to LIFO, we have $x_{t+1,1}=(S_{t,1}-D_t)^+=0$ and $x_{t+1,2}=(S_{t,2}-(D_t-S_{t,1})^+)^+=1.9$. The perished proportion is then calculated as $q_{t,2}=E_{t,2}/x_{t,2}=0.05$. We repeat this process until the end of the time horizon and fit $\{q_{t,2}\}$ to a beta distribution in order to calibrate $\boldsymbol{\xi}_2$.

By expressing D_t using the demand model (1), we formulate the age-dependent full information problem as the following dynamic program: for $t\in\{1,2,\ldots,T\}$ and $x_{t,1},x_{t,2}\in[0,y_{\max}]$,

$$\begin{aligned} V_t(x_{t,1},x_{t,2})=\max_{\substack{p_t\in\mathscr{P},y_t\in\mathscr{Y}\\ y_t\geq x_{t,1}+x_{t,2}}}\Big\{ & p_t g(\alpha_t+\beta_t p_t)-h\mathbb{E}_{q_2,\varepsilon}[y_t-q_{t,2}x_{t,2}-g(\alpha_t+\beta_t p_t)-\varepsilon_t]^+ \\ & -(b+p_t)\mathbb{E}_{q_2,\varepsilon}[g(\alpha_t+\beta_t p_t)+\varepsilon_t-y_t+q_{t,2}x_{t,2}]^+ \\ & +\mathbb{E}_{q_1,q_2,\varepsilon}[V_{t+1}(x_{t+1,1},x_{t+1,2})]-cy_t\Big\}+c(x_{t,1}+x_{t,2})-w\mathbb{E}_{q_2}[q_{t,2}]x_{t,2} \end{aligned} \tag{33}$$

$$\begin{aligned} \text{subject to}\quad & x_{t+1,1}=[(y_t-x_{t,2})(1-q_{t,1})-g(\alpha_t+\beta_t p_t)-\varepsilon_t]^+ \\ & x_{t+1,2}=[y_t-q_{t,2}x_{t,2}-g(\alpha_t+\beta_t p_t)-\varepsilon_t-x_{t+1,1}]^+, \end{aligned}$$

with $V_{T+1}(x_{T+1,1},x_{T+1,2})=c(x_{T+1,1}+x_{T+1,2})$, where c, h, b, and w are the unit ordering, holding, lost-sales penalty, and disposal costs, respectively. The solution to (33) is termed as FIA-Age.

To implement the DDPO policy in this setting, we make the following modifications. In each period t in the learning phase of a cycle, we observe $q_{t,1}$, $q_{t,2}$, and the demand response to the test prices, then conduct the detection test, and use all the information starting from the latest detection cycle to estimate $\boldsymbol{\theta}$, $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2$, and F_ε. Based on these estimates, we determine the price and order-up-to levels for the earning phase of the same cycle by solving a dynamic program akin to (33), except with a time horizon spanning the upcoming earning phase. We denote this modified policy as DDPO-Age. The reason for using the current cycle as a rolling horizon is the changing environment: if the dynamic program were solved until the end of period T based on the current estimates, this would lead to model misspecification when the demand parameters change. The rolling horizon helps the retailer guard against the dynamic nature of this model misspecification.

We measure the performance of the DDPO-Age policy by its regret (i.e., the profit loss relative to FIA-Age), using ginger in store A as an example. Recall that, based on the real-life data set where products of ages 1 and 2 are combined into a single state, the calibrated perishability parameter is $\boldsymbol{\xi} = (0.4, 43.99)$. Informed by this, we consider three settings for $\boldsymbol{\xi}_1$: $(0.4, 10)$, $(0.4, 20)$, and $(0.4, 30)$. The corresponding calibrated values of $\boldsymbol{\xi}_2$ are $(0.4, 21.46)$, $(0.4, 17.98)$, and $(0.4, 16.09)$. These three parameter sets represent the cases where the perishability rate from age-1 to age-2 products is higher, about the same, and lower than that from age-2 to age-3 products, respectively. When sampling $\{q_{t,1} : t = 1, \ldots, T\}$ from the beta distribution with parameter $\boldsymbol{\xi}_1$, we use the acceptance-rejection method to ensure that $x_{t,2} \geq E_{t,2}$ and thus $q_{t,2} \in [0, 1]$ for all t. As shown in §4.2.1, the demand noise of ginger in store A can be fitted to a normal distribution, so we consider the parametric version of the DDPO-Age policy with parameters $(\eta, \kappa, \omega_1, \omega_2, \upsilon) = (2, 2, 5, 12, 2)$. The other parameters are the same as in §4.2.1.

Table 5 shows the regret of DDPO-Age and DDPO over 365 days, based on 30 sample paths. First, we observe that the regret is not very sensitive to the perishability parameters $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$, displaying only a slight increase as age-1 products become less perishable and age-2 products become more perishable. Second, the percentage regret reduction by taking the age-dependent perishability into account (using DDPO-Age instead of DDPO) is in the range 11.1–11.7%. This indicates that keeping track of product freshness can offer substantial value to the retailer.

Table 5 Regret with Age-dependent Perishability over 365 Periods

$\boldsymbol{\xi}_1$	$\boldsymbol{\xi}_2$	Regret of DDPO-Age	Regret of DDPO
$(0.4, 10)$	$(0.4, 21.46)$	1657.9	1878.4
$(0.4, 20)$	$(0.4, 17.98)$	1710.7	1926.7
$(0.4, 30)$	$(0.4, 16.09)$	1726.9	1942.6

As a result, the tools we developed here can help grocery retailers assess the costs and benefits of tracking product ages. In the context of our real-life data set, the supermarket chain does not record the ages of fresh fruits and vegetables, but as digital technologies advance, tracking product ages is expected to become more convenient and inexpensive. For example, internet-enabled sensors and the blockchain technology seamlessly provide retailers with first-hand information on the freshness of perishable products, as well as their status in transit to store (Kearney 2020). These technologies have the potential to facilitate the implementation of policies like DDPO-Age, giving rise to more profitable, transparent, and responsive supply chains of fresh produce.

6.2. Demand Censoring

This subsection investigates the extension of our base model to demand censoring, using the Kaplan-Meier estimator for generalized linear models (Yu et al. 2009). As an illustrative example, we use the tomato data in store A, because there are days when tomatoes are out of stock in this store. We first calibrate the demand parameter sequence $\boldsymbol{\theta}$ using the exponential demand model with $\mathscr{C} = 3$ change-points that are at

least 60 days apart. (As mentioned above, we employ the Kaplan-Meier estimator to account for censored demand in this calibration.) Using multiple Kolmogorov-Smirnov tests, we reject the null hypotheses that the demand noise follows the normal, gamma or exponential distributions, so we treat the noise distribution as nonparametric (Setting N) in this case. The calibrated perishability parameter $\boldsymbol{\xi}$ is $(0.4, 17)$, and the unit cost parameters are $c = 4.9011$, $h = 0.0013$, $b = 0.0026$, $w = 0.0013$, obtained in the same way as in §4.2.1.

In the presence of censored demand, we consider two variants of our DDPO-N policy. The first one uses the Kaplan-Meier estimator to account for demand censoring, while the second ignores demand censoring by treating sales as demand. The demand quantities used for change-point detection are demand estimates that account for censoring based on the Kaplan-Meier estimator. The parameter vector $(\eta, \kappa, \omega_1, \omega_2, \upsilon)$ is $(4, 1, 8, 12, 20)$. We estimate the growth rate of regret over $T \in \{2000, 4000, \ldots, 20000\}$ based on 500 sample paths. Table 6 shows the growth rates of the T-period regret under the above variants of DDPO-N. We observe that ignoring demand censoring when it is present leads to a slight increase in the growth rate of regret. We also observe that the regret of the DDPO-N policy that uses the Kaplan-Meier estimator grows at essentially the same rate as our theoretical upper bound on the regret of DDPO-N in the absence of censoring. This suggests that the Kaplan-Meier estimator may be able to mitigate the potential negative impacts of demand censoring.

Table 6 Effect of Demand Censoring on the Growth Rate of Regret

Censored Demand	Account for Censoring	T-period Regret of DDPO-N
No	No	$O(T^{0.65}(\log T)^{1/2})$
Yes	Yes	$O(T^{0.66}(\log T)^{1/2})$
Yes	No	$O(T^{0.69}(\log T)^{1/2})$

We now examine the effect of the test order-up-to level υ on the regret of DDPO-N in the presence of demand censoring over 365 days. To ensure that each cycle has sufficient data points for demand estimation and to prevent multiple change-points from being included in the same cycle, we set the cycle-length parameter κ as 0.5. All other parameters remain unchanged. Table 7 displays the regret and fill rates under various values of υ based on 500 sample paths. Ignoring demand censoring while it is present leads to a percentage regret increase in the range of $0.1-3.8\%$.

Table 7 also quantifies the effect of demand censoring on the retailer's profits. As υ decreases from 20 to 5, the percentage regret increase due to the presence of censored demand rises dramatically from 1.1% to 37.3% (even if censoring is properly accounted for with the Kaplan-Meier estimator). This negative effect of demand censoring can be alleviated for retailers who can estimate their lost sales accurately, such as online retailers that track customers' browsing data, as well as any physical stores equipped with facial recognition technology to record unsatisfied customers. Our numerical findings suggest that technologies that mitigate demand censoring offer substantial value.

Table 7 Effect of Test Order-up-to Level on Regret over 365 Periods

υ	Censored Demand	Account for Censoring	Regret of DDPO-N	Fill Rate of DDPO-N	% Regret Increase Due to Censored Demand	% Regret Increase Due to Ignoring Censoring
20	No	No	1934.1	94.32%	1.1%	0.1%
	Yes	Yes	1955.0	93.65%		
	Yes	No	1956.5	93.55%		
15	No	No	2015.3	92.20%	3.2%	0.2%
	Yes	Yes	2080.4	90.60%		
	Yes	No	2084.7	90.16%		
10	No	No	2341.4	86.84%	14.2%	0.7%
	Yes	Yes	2673.4	82.17%		
	Yes	No	2691.8	80.91%		
5	No	No	3205.2	78.87%	37.3%	3.8%
	Yes	Yes	4401.8	65.12%		
	Yes	No	4567.7	64.83%		

The fill rates in Table 7 provide managerial guidelines for grocery retailers to hedge against demand censoring: targeting a higher fill rate by setting a higher test order-up-to level can effectively eliminate the effect of demand censoring and considerably increase profitability.

Regarding the methods used in this subsection, we note that Yu et al. (2009) present theoretical results on the asymptotic performance of the Kaplan-Meier estimator. Our regret analysis, on the other hand, is built on convergence results that are applicable to any finite sample. Despite this difference, Yu et al. (2009) provide helpful intuition on finite-sample performance guarantees for the Kaplan-Meier estimator in the context of data-driven dynamic pricing and ordering with censored demand. The theoretic analysis of this extension is beyond the scope of this paper and we leave it to future work.

7. Concluding Remarks

Motivated by observations on a real-life data set, we study joint pricing and ordering decisions for a multi-period perishable lost-sales inventory system with several new features. That is, the demand-price relationship, demand noise distribution, and product perishability rate are all unknown to the retailer. In addition, the demand-price relationship is non-stationary with unknown change-points. Thus, the knowledge gained based on past information becomes obsolete after a shift in the demand environment.

We design two data-driven pricing and ordering policies—DDPO-N and DDPO-E—for nonparametric and exponential-family demand noise distributions, respectively. We establish theoretical upper bounds on the T-period regret of these policies, and show that they are both rate-optimal in their respective settings (up to logarithmic terms). Through a case study on the real-life data set, we demonstrate that our policies significantly outperform the historical decisions of the supermarket. Finally, we extend our base model by allowing for age-dependent product perishability and demand censoring.

Our paper advances the state-of-the-art inventory decision models in several dimensions. First, it develops a data-driven joint pricing and ordering model for perishable products, whereas the previous literature only

studied such decisions with known demand information. Second, our model takes into account a dynamic demand-price relationship with unknown change-points; the previous studies have essentially focused on Markovian environments with known transition probabilities. Third, we observe in our real-life data set that the perishability rate follows a non-stationary distribution, a feature that has not been studied in the literature, but our approach can be adapted to accommodate it. We believe the methodology developed here opens the door to many other applications in the field of inventory management and beyond.

References

Abdelnour, A., Babbitz, T., and Moss, S. (2020). Pricing in a pandemic: Navigating the COVID-19 crisis. *McKinsey & Company*, May 1.

Bakker, M., Riezebos, J., and Teunter, R. H. (2012). Review of inventory systems with deterioration since 2001. *European Journal of Operational Research*, 221(2):275–284.

Ban, G.-Y. and Keskin, N. B. (2020). Personalized dynamic pricing with machine learning: High dimensional features and heterogeneous elasticity. *Management Science*, forthcoming. Available at SSRN: https://ssrn.com/abstract=2972985.

Becdach, C., Brown, B., Halbardier, F., Henstorf, B., and Murphy, R. (2020). Rapidly forecasting demand and adapting commercial plans in a pandemic. *McKinsey & Company*, April 21.

Besbes, O. and Zeevi, A. (2011). On the minimax complexity of pricing in a changing environment. *Operations Research*, 59(1):66–79.

Borovkov, A. A. (1998). *Mathematical Statistics*. Gordon and Breach Science Publishers, Amsterdam.

Briedis, H., Kronschnabl, A., Rodriguez, A., and Ungerman, K. (2020). Adapting to the next normal in retail: The customer experience imperative. *McKinsey & Company*, May 14.

Business Insider (2018). Amazon changes prices on its products about every 10 minutes—here's how and why they do it. Accessed: December 4, 2018.

Chen, B., Chao, X., and Ahn, H.-S. (2019). Coordinating pricing and inventory replenishment with nonparametric demand learning. *Operations Research*, 67(4):1035–1052.

Chen, B., Chao, X., and Shi, C. (2020). Nonparametric learning algorithms for joint pricing and inventory control with lost-sales and censored demand. Available at SSRN: https://ssrn.com/abstract=2836057.

Chen, M. and Chen, Z.-L. (2015). Recent developments in dynamic pricing research: multiple products, competition, and limited demand information. *Production and Operations Management*, 24(5):704–731.

Chen, X., Pang, Z., and Pan, L. (2014). Coordinating inventory control and pricing strategies for perishable products. *Operations Research*, 62(2):284–300.

CNN (2020). Forget pork. Here's why you can't buy flour. Accessed: May 2, 2020.

Columbia University Earth Institute (2018). How climate change will alter our food. Accessed: August 27, 2018.

den Boer, A. V. (2015). Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in Operations Research and Management Science*, 20(1):1–18.

den Boer, A. V. and Keskin, N. B. (2019). Dynamic pricing with demand learning and reference effects. Available at SSRN: https://ssrn.com/abstract=3092745.

den Boer, A. V. and Keskin, N. B. (2020). Discontinuous demand functions: Estimation and pricing. *Management Science*, forthcoming. https://doi.org/10.1287/mnsc.2019.3446.

den Boer, A. V. and Zwart, B. (2014). Mean square convergence rates for maximum quasi-likelihood estimators. *Stochastic Systems*, 4(2):375–403.

Ferreira, K. J., Simchi-Levi, D., and Wang, H. (2018). Online network revenue management using Thompson sampling. *Operations Research*, 66(6):1586–1602.

Fresh Plaza (2016). India: Hot weather increases summer fruit demand. Accessed: July 24, 2018.

Goyal, S. and Giri, B. C. (2001). Recent trends in modeling of deteriorating inventory. *European Journal of Operational Research*, 134(1):1–16.

Heyman, D. P. and Sobel, M. J. (1982). *Stochastic Models in Operations Research: Stochastic Optimization*, volume 2. Courier Corporation.

HuffPost (2017). Retail tech is changing how we shop. Accessed: March 3, 2019.

Kalpakam, S. and Sapna, K. (1994). Continuous review (s, S) inventory system with random lifetimes and positive leadtimes. *Operations Research Letters*, 16(2):115–119.

Karaesmen, I. Z., Scheller-Wolf, A., and Deniz, B. (2011). Managing perishable and aging inventories: Review and future research directions. In *Planning production and inventories in the extended enterprise*, pages 393–436. Springer.

Kearney (2020). A strawberry's journey through the digital supply chain. Accessed: May 10, 2020.

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. and Zeevi, A. (2014). Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research*, 62(5):1142–1167.

Keskin, N. B. and Zeevi, A. (2017). Chasing demand: Learning and earning in a changing environment. *Mathematics of Operations Research*, 42(2):277–307.

Kleinberg, R. D. (2005). Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704.

Li, Y., Cheang, B., and Lim, A. (2012). Grocery perishables management. *Production and Operations Management*, 21(3):504–517.

Liao, J.-J. (2007). On an EPQ model for deteriorating items under permissible delay in payments. *Applied Mathematical Modelling*, 31(3):393–403.

Nahmias, S. (1982). Perishable inventory theory: A review. *Operations Research*, 30(4):680–708.

Pesaran, M. H., Pettenuzzo, D., and Timmermann, A. (2006). Forecasting time series subject to multiple structural breaks. *The Review of Economic Studies*, 73(4):1057–1084.

Phillips, R. L. (2005). *Pricing and Revenue Optimization*. Stanford University Press.

RW3 (2016). Is today's consumer ready for dynamic pricing while grocery shopping? Accessed: December 4, 2018.

Shin, D. and Zeevi, A. (2017). Dynamic pricing and learning with online product reviews. Working paper, Columbia University.

The Irish Times (2018). Increase in cases of E. coli infections due to hot weather. Accessed: August 27, 2018.

Wong, J. and Schuchard, R. (2011). Adapting to climate change: A guide for the consumer products industry. Accessed: August 27, 2018.

Yu, L., Yu, R., and Liu, L. (2009). Quasi-likelihood for right-censored data in the generalized linear model. *Communications in Statistics–Theory and Methods*, 38(13):2187–2200.

Appendix A: Proofs of Theoretical Results

Let us first state and prove a key lemma that we use in some of our proofs.

LEMMA A.1. *For $p_t \in \mathscr{P}$, $\boldsymbol{\theta}_t \in \Theta$, $\boldsymbol{\xi} \in \Xi$, the unconstrained optimal order-up-to level in period t is given by*

$$y_t^u(p_t;\boldsymbol{\theta}_t,\boldsymbol{\xi}) = \inf\{y_t \in \mathscr{Y} : H'_+(y_t;p_t,\boldsymbol{\theta}_t,\boldsymbol{\xi}) \geq 0\} = \sup\{y_t \in \mathscr{Y} : H'_-(y_t;p_t,\boldsymbol{\theta}_t,\boldsymbol{\xi}) < 0\}, \tag{A.1}$$

where $H'_+(y_t;p_t,\boldsymbol{\theta}_t,\boldsymbol{\xi}) = (h-c+b+p_t)\mathbb{E}_{q|\boldsymbol{\xi}}[(1-q_t)F_\varepsilon(z_t(q_t))] - (b+p_t)(1-\mathbb{E}_{q|\boldsymbol{\xi}}[q_t]) + (w\mathbb{E}_{q|\boldsymbol{\xi}}[q_t]+c)$ *and* $H'_-(y_t;p_t,\boldsymbol{\theta}_t,\boldsymbol{\xi}) = (h-c+b+p_t)\mathbb{E}_{q|\boldsymbol{\xi}}[(1-q_t)F_\varepsilon(z_t(q_t)^-)] - (b+p_t)(1-\mathbb{E}_{q|\boldsymbol{\xi}}[q_t]) + (w\mathbb{E}_{q|\boldsymbol{\xi}}[q_t]+c)$ *are the right and left derivatives of $H(\cdot\,;p_t,\boldsymbol{\theta}_t,\boldsymbol{\xi})$ at y_t, with $z_t(q_t) = (1-q_t)y_t - g(\mathbf{X}_t^\intercal\boldsymbol{\theta}_t)$ and $F_\varepsilon(x^-)$ denotes the left limit of F_ε at x.*

Proof of Lemma A.1. For any given $p_t \in \mathscr{P}$, $\boldsymbol{\theta}_t \in \Theta$, $\boldsymbol{\xi} \in \Xi$ in period t, the single period total cost function (5) can be re-written as

$$H(y_t;p_t,\boldsymbol{\theta}_t,\boldsymbol{\xi}) = (h-c+b+p_t)\mathbb{E}_{\varepsilon,q|\boldsymbol{\xi}}[z_t(q_t)-\varepsilon_t]^+ - (b+p_t)\mathbb{E}_{q|\boldsymbol{\xi}}[z_t(q_t)] + (w\mathbb{E}_{q|\boldsymbol{\xi}}[q_t]+c)y_t,$$

where $z_t(q_t) = (1-q_t)y_t - g(\mathbf{X}_t^\intercal\boldsymbol{\theta}_t)$. If the demand noise ε_t is a continuous random variable, then by the Leibniz integral rule,

$$\frac{\partial H(y_t;p_t,\boldsymbol{\theta}_t,\boldsymbol{\xi})}{\partial y} = (h-c+b+p_t)\mathbb{E}_{q|\boldsymbol{\xi}}[(1-q_t)F_\varepsilon(z_t)] - (b+p_t)(1-\mathbb{E}_{q|\boldsymbol{\xi}}[q_t]) + (w\mathbb{E}_{q|\boldsymbol{\xi}}[q_t]+c),$$

where exchange of differentiation and the expectation follows from the dominated convergence theorem. Letting this expression be equal to 0 gives (A.1).

If the demand noise ε_t is a discrete random variable, then the total cost function is differentiable with respect to y_t except at the "kinks" induced by the values ε_t can take. Specifically, let $\{\epsilon_k\}_{k=1}^\infty$ be the set of possible values taken by ε_t, ordered from smallest to largest, with corresponding probabilities $\{o_k\}_{k=1}^\infty$. Thus, the right derivative of $H(\cdot;p_t,\boldsymbol{\theta}_t,\boldsymbol{\xi})$ at y_t is

$$H'_+(y_t;p_t,\boldsymbol{\theta}_t,\boldsymbol{\xi}) = (h-c+b+p_t)\mathbb{E}_{q|\boldsymbol{\xi}}\left[\sum_{k=1}^{\ell(z_t(q_t))} o_k(1-q_t)\right] - (b+p_t)(1-\mathbb{E}_{q|\boldsymbol{\xi}}[q_t]) + (w\mathbb{E}_{q|\boldsymbol{\xi}}[q_t]+c)\,,$$

where $\ell(z_t(q_t)) = \sup\{k : \epsilon_k \leq z_t(q_t)\}$. Alternatively, the left derivative of H at y_t is given by

$$H'_-(y_t;p_t,\boldsymbol{\theta}_t,\boldsymbol{\xi}) = (h-c+b+p_t)\mathbb{E}_{q|\boldsymbol{\xi}}\left[\sum_{k=1}^{\ell(z_t(q_t))-1} o_k(1-q_t)\right] - (b+p_t)(1-\mathbb{E}_{q|\boldsymbol{\xi}}[q_t]) + (w\mathbb{E}_{q|\boldsymbol{\xi}}[q_t]+c)\,.$$

Note that $\sum_{k=1}^{\ell(z_t(q_t))} o_k = F_\varepsilon(z_t(q_t))$ and $\sum_{k=1}^{\ell(z_t(q_t))-1} o_k = F_\varepsilon(z_t(q_t)^-)$, where $F_\varepsilon(x^-)$ is the left limit of $F_\varepsilon(\cdot)$ at x. Clearly, it is optimal to increase y_t until $H'_+(y_t;p_t,\boldsymbol{\theta}_t,\boldsymbol{\xi}) \geq 0$. This completes the proof. Q.E.D.

Proof of Proposition 1. For $j = 0,1,\ldots,\mathscr{C}$, denote the number of periods between the j^{th} and $(j+1)^{\text{st}}$ change-points by $\mathscr{L}_j = t^*_{j+1} - t^*_j$. Let $\{(\check{p}_t,\check{y}_t) : t = 1,\ldots,T\}$ be the full-information non-anticipatory policy while $\{(p_t^u,y_t^u) : t = 1,\ldots,T\}$ be the one that ignores the constraint $y_t \geq x_t$. Suppose that $x_{t_j^*} =$

$\check{y}_{t_j^*} > y^u_{t_j^*}$ at the j^{th} change-point t_j^*, where $x_{t_j^*}$ is the initial on-hand inventory level at t_j^*. Let $\tilde{\tau}_j = \inf\{t > t_j^* : \check{y}_t = y_t^u\} \wedge t_{j+1}^*$ be the first period when the inventory decisions coincide, and thus, $\check{p}_t = p_t^u$ before the next change-point. By definition, $\check{y}_t > y_t^u \geq y_{\min} > 0$ and $\check{y}_t = x_t$ for all t satisfying $t_{j^*} \leq t < \tilde{\tau}_j$. Therefore, $x_{t+1} = x_t - (q_t x_t + D_t)$ for all t such that $t_{j^*} \leq t < \tilde{\tau}_j$. As a result, $(\check{p}_t, \check{y}_t) = (p_t^u, y_t^u)$ for all t satisfying $\tilde{\tau}_j \leq t \leq t_{j+1}^*$.

Note that for any positive integer M, if $\sum_{t=t_j^*}^{t_j^*+M-1} D_t \geq x_{t_j^*} - y^u_{t_j^*+M}$, then $\tilde{\tau}_j - t_j^* \leq M$. For any $\delta > 0$, construct a stochastic process $\{\tilde{Z}_t : t = 0, 1, \ldots\}$ with $\tilde{Z}_0 = 1$ and

$$\tilde{Z}_t = \exp\left\{\frac{1}{\zeta}\left(-\delta \sum_{s=1}^{t} \varepsilon_s - \frac{1}{2}\delta^2 t\right)\right\},$$

for $t = 1, 2, \ldots$, where $\zeta = \frac{\delta}{\varpi_0} \vee (\varrho_0 \sigma_0^2)$. Clearly, $\tilde{Z}_t$ is integrable for all t. Let $\mathcal{F}_t^{\varepsilon} = \sigma(\varepsilon_1, \ldots, \varepsilon_t)$. Then,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}^{\pi}[\tilde{Z}_t | \mathcal{F}_{t-1}] &= \mathbb{E}_{\boldsymbol{\theta}}^{\pi}\left[\exp\left\{\frac{1}{\zeta}\left(-\delta\sum_{s=1}^{t-1}\varepsilon_s - \frac{1}{2}\delta^2(t-1)\right)\right\}\exp\left\{\frac{1}{\zeta}\left(-\delta\varepsilon_t - \frac{1}{2}\delta^2\right)\right\}\middle| \mathcal{F}_{t-1}\right] \\
&= \tilde{Z}_{t-1}\mathbb{E}_{\boldsymbol{\theta}}^{\pi}\left[\exp\left\{\frac{1}{\zeta}\left(-\delta\varepsilon_t - \frac{1}{2}\delta^2\right)\right\}\right] \\
&\leq \tilde{Z}_{t-1}\exp\left\{\frac{1}{2}\varrho_0\sigma_0^2\frac{\delta^2}{\zeta^2} - \frac{\delta^2}{2\zeta}\right\} \leq \tilde{Z}_{t-1}.
\end{aligned}$$

Therefore, $\{(\tilde{Z}_t, \mathcal{F}_t) : t = 0, 1, \ldots\}$ is a supermartingale with the initial value of 1. As a result, by the Markov inequality,

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\theta}}^{\pi}\left\{\sum_{s=1}^{t}\varepsilon_s \leq -\delta t\right\} = \mathbb{P}_{\boldsymbol{\theta}}^{\pi}\left\{\sum_{s=1}^{t}\varepsilon_s + \frac{1}{2}\delta t \leq -\frac{1}{2}\delta t\right\} &= \mathbb{P}_{\boldsymbol{\theta}}^{\pi}\left\{\tilde{Z}_t \geq \exp\left(\frac{\delta^2 t}{2\zeta}\right)\right\} \\
&\leq \exp\left(-\frac{\delta^2 t}{2\zeta}\right) \\
&= \exp\left(\frac{-\varpi_0\delta t}{2} \vee \frac{-\delta^2 t}{2\varrho_0\sigma_0^2}\right). \qquad \text{(A.2)}
\end{aligned}$$

For any $a > 0$, there exists a positive integer N_1 such that $N\mu_{\min} - \sqrt{2a\varrho_0\sigma_0^2 N \log N} \geq y_{\max} - y_{\min}$ for all $N \geq N_1$. Moreover, there exists a positive integer N_2 such that $2a \log N \leq \varpi_0^2 \varrho_0 \sigma_0^2 N$ for all $N \geq N_2$. Hence, for $j = 0, 1, \ldots, \mathscr{C}$, there exists a positive integer $\tilde{N}_j$ such that $\tilde{N}_j^a \geq \mathscr{L}_j$ with $\tilde{N}_j \geq N_1 \vee N_2$. If $\tilde{N}_j \geq \mathscr{L}_j$, then we deduce from the mean value theorem and the Cauchy-Schwarz inequality that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}^{\pi}\left[\sum_{t=t_j^*}^{t_{j+1}^*-1}\left[Q(\check{p}_t, \check{y}_t; \boldsymbol{\theta}_t, \boldsymbol{\xi}) - Q(p_t^u, y_t^u; \boldsymbol{\theta}_t, \boldsymbol{\xi})\right]\right] &\leq \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\xi}}^{\pi}\left[\sum_{t=t_j^*}^{t_{j+1}^*-1}\|\nabla Q(\bar{p}, \bar{y})\| \cdot \|(\check{p}_t, \check{y}_t) - (p_t^u, y_t^u)\|\right] \\
&\leq \overline{Q'}\,\overline{d}(\mathscr{P} \times \mathscr{Y})\tilde{N}_j,
\end{aligned}$$

where $\overline{Q'} = \max\{\|\nabla Q(p, y)\| : p \in \mathscr{P}, y \in \mathscr{Y}\}$, $(\bar{p}, \bar{y})$ is on the line segment connecting (p_t, y_t) and (p_t^u, y_t^u) in $\mathscr{P} \times \mathscr{Y}$, and $\overline{d}(\mathscr{P} \times \mathscr{Y}) = \sqrt{(p_{\max} - p_{\min})^2 + (y_{\max} - y_{\min})^2}$. Having obtained the desired result in this case, let us now consider the case where $\tilde{N}_j < \mathscr{L}_j$. If we take $\delta = \sqrt{2a\varrho_0\sigma_0^2 \log \tilde{N}_j / \tilde{N}_j}$ in (A.2), we

then have

$$\mathbb{P}^{\pi}_{\boldsymbol{\theta}}\left\{\sum_{t=t_j^*}^{t_j^*+\tilde{N}_j-1}\varepsilon_t \le -\sqrt{2a\varrho_0\sigma_0^2\tilde{N}_j\log\tilde{N}_j}\right\} \le \exp\left(-a\log\tilde{N}_j\right) = \tilde{N}_j^{-a}, \tag{A.3}$$

because $2a\log\tilde{N}_j \le \varpi_0^2\varrho_0\sigma_0^2\tilde{N}_j$ implies that $-\varpi_0\sqrt{2a\varrho_0\sigma_0^2\tilde{N}_j\log\tilde{N}_j} \le -a\log\tilde{N}_j$. Let

$$\mathscr{B}_j = \left\{\sum_{t=t_j^*}^{t_j^*+\tilde{N}_j-1}\varepsilon_t \ge -\sqrt{2a\varrho_0\sigma_0^2\tilde{N}_j\log\tilde{N}_j}\right\}.$$

Note that on $\mathscr{B}_j$, we have

$$\begin{aligned}\sum_{t=t_j^*}^{t_j^*+\tilde{N}_j-1} D_t \ge \sum_{t=t_j^*}^{t_j^*+\tilde{N}_j-1} g(\mathbf{X}_t^{\intercal}\boldsymbol{\theta}_t) - \sqrt{2a\varrho_0\sigma_0^2\tilde{N}_j\log\tilde{N}_j} &\ge \tilde{N}_j\mu_{\min} - \sqrt{2a\varrho_0\sigma_0^2\tilde{N}_j\log\tilde{N}_j} \ge y_{\max} - y_{\min}\\ &\ge x_{t_j^*} - y^u_{t_j^*+\tilde{N}_j},\end{aligned}$$

because $\tilde{N}_j \ge N_1$. Therefore, $t_j^* + \tilde{N}_j \ge \tilde{\tau}_j$ and $(\check{p}_t, \check{y}_t) = (p_t^u, y_t^u)$ for all t such that $t_j^* + \tilde{N}_j \le t < t_{j+1}^*$. Consequently,

$$\begin{aligned}&\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{t=t_j^*}^{t_{j+1}^*-1}|Q(\check{p}_t,\check{y}_t;\boldsymbol{\theta}_t,\boldsymbol{\xi}) - Q(p_t^u,y_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi})|\right]\\
&\quad= \mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{t=t_j^*}^{t_{j+1}^*-1}|Q(\check{p}_t,\check{y}_t;\boldsymbol{\theta}_t,\boldsymbol{\xi}) - Q(p_t^u,y_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi})|\,\middle|\,\mathscr{B}_j\right]\mathbb{P}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}(\mathscr{B}_j)\\
&\qquad+ \mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{t=t_j^*}^{t_{j+1}^*-1}|Q(\check{p}_t,\check{y}_t;\boldsymbol{\theta}_t,\boldsymbol{\xi}) - Q(p_t^u,y_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi})|\,\middle|\,\mathscr{B}_j^c\right]\mathbb{P}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}(\mathscr{B}_j^c)\\
&\quad\le \overline{Q'}\,\overline{d}(\mathscr{P}\times\mathscr{Y})\tilde{N}_j + \overline{Q'}\,\overline{d}(\mathscr{P}\times\mathscr{Y})\mathscr{L}_j\tilde{N}_j^{-a}\\
&\quad\le \overline{Q'}\,\overline{d}(\mathscr{P}\times\mathscr{Y})(\tilde{N}_j+1).\end{aligned}$$

Let $\tilde{N} = \max\{\tilde{N}_j : j = 0,1,\ldots,\mathscr{C}\}$. Then,

$$\begin{aligned}\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{t=1}^{T}|Q(\check{p}_t,\check{y}_t;\boldsymbol{\theta}_t,\boldsymbol{\xi}) - Q(p_t^u,y_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi})|\right] &= \mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{j=0}^{\mathscr{C}}\sum_{t=t_j^*}^{t_{j+1}^*-1}|Q(\check{p}_t,\check{y}_t;\boldsymbol{\theta}_t,\boldsymbol{\xi}) - Q(p_t^u,y_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi})|\right]\\
&\le \overline{Q'}\,\overline{d}(\mathscr{P}\times\mathscr{Y})(\tilde{N}+1)(\mathscr{C}+1).\end{aligned}$$

On the other hand, under the full-information anticipatory policy $\{(p_t^*, y_t^*) : t = 1,\ldots,T\}$, it is possible to adjust the order-up-to level before change-points. Following the same procedure as above, we have

$$\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{t=1}^{T}|Q(p_t^*,y_t^*;\boldsymbol{\theta}_t,\boldsymbol{\xi}) - Q(p_t^u,y_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi})|\right] \le \overline{Q'}\,\overline{d}(\mathscr{P}\times\mathscr{Y})(\tilde{N}+1)(\mathscr{C}+1).$$

Combining these two inequalities leads to the result. Q.E.D.

Proof of Proposition 2. Fix a period $t \notin \mathscr{X}$ in cycle τ such that $\tau_j^* < \tau < \hat{\tau}_j^-$. By definition, $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t_j^*}$. Fix $\delta > 0$. Let $\mathcal{A}_\tau = \{||\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t||^2 > K_3\delta\}$ and $d_* = \max\{||\boldsymbol{\vartheta} - \tilde{\boldsymbol{\vartheta}}|| : \boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}} \in \Theta\}$. Because $\hat{\boldsymbol{\theta}}_t, \boldsymbol{\theta}_t \in \Theta$, we deduce

that $\mathbb{P}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}(\mathcal{A}_\tau)=0$ if $K_3\delta \geq d_*^2$. Thus, we consider the case where $K_3\delta < d_*^2$ for the rest of the proof. By Corollary 1 in den Boer and Zwart (2014), on the event $\mathcal{A}_\tau$, there exists $\boldsymbol{\vartheta}_t \in \mathbb{R}^2$ such that $||\boldsymbol{\vartheta}_t - \boldsymbol{\theta}_t||^2 = K_3\delta$ and

$$(\boldsymbol{\vartheta}_t - \boldsymbol{\theta}_t)^\intercal \sum_{s=nL(\tau)+1}^{t} \mathbb{I}\{s \in \mathscr{X}\}(D_s - g(\mathbf{X}_s^\intercal \boldsymbol{\vartheta}_t))\mathbf{X}_s > 0. \tag{A.4}$$

Note that the number of terms in the summation is $N_j^\tau = 2M_t$ and that there exists a one-to-one correspondence $\mathcal{L}_j(\cdot)$ that maps $\ell \in \{1,2,\ldots,N_j^\tau\}$ to $t \in \{1,\ldots,T\}$. Specifically, $\mathcal{L}_j(\ell) = nL(\tau) + (n-2m)\lfloor \ell/(2m) \rfloor + \ell$. Define $\Lambda_j = \{\ell : L(\tau_j^*) \leq \lceil \mathcal{L}_j(\ell)/n \rceil - 1 < \tau_j^*\}$. This set is non-empty only if $\tau_j^* < \tau < \hat{\tau}_j^+$, in which case $L(\tau) = L(\tau_j^*)$ and the true demand parameter in cycles from $L(\tau)$ to τ_j^* is different from $\boldsymbol{\theta}_{t_j^*}$. Recalling that $D_s = g(\mathbf{X}_s^\intercal \boldsymbol{\theta}_s) + \varepsilon_s$ for all s, we can rewrite (A.4) as

$$\begin{aligned}(\boldsymbol{\vartheta}_t - \boldsymbol{\theta}_t)^\intercal \sum_{\ell=1}^{N_j^\tau} \mathbf{X}_{\mathcal{L}_j(\ell)}\varepsilon_{\mathcal{L}_j(\ell)} > (\boldsymbol{\vartheta}_t - \boldsymbol{\theta}_t)^\intercal \Bigg\{ &\sum_{\ell=1}^{N_j^\tau} [g(\mathbf{X}_{\mathcal{L}_j(\ell)}^\intercal \boldsymbol{\vartheta}_t) - g(\mathbf{X}_{\mathcal{L}_j(\ell)}^\intercal \boldsymbol{\theta}_t)]\mathbf{X}_{\mathcal{L}_j(\ell)} \\ &+ \sum_{\ell \in \Lambda_j} [g(\mathbf{X}_{\mathcal{L}_j(\ell)}^\intercal \boldsymbol{\theta}_t) - g(\mathbf{X}_{\mathcal{L}_j(\ell)}^\intercal \boldsymbol{\theta}_{\mathcal{L}_j(\ell)})]\mathbf{X}_{\mathcal{L}_j(\ell)} \Bigg\}. \end{aligned} \tag{A.5}$$

Let $\mathscr{M}_j^{\mathcal{L}}(u) = \sum_{\ell=1}^{u} \mathbf{X}_{\mathcal{L}_j(\ell)}\varepsilon_{\mathcal{L}_j(\ell)}$, $\mathscr{F}_j^{\mathcal{L}}(u) = \sum_{\ell=1}^{u} \mathbf{X}_{\mathcal{L}_j(\ell)}\mathbf{X}_{\mathcal{L}_j(\ell)}^\intercal$ and $\mathscr{R}_j^{\mathcal{L}} = \sum_{\ell \in \Lambda_j} [g(\mathbf{X}_{\mathcal{L}_j(\ell)}^\intercal \boldsymbol{\theta}_t) - g(\mathbf{X}_{\mathcal{L}_j(\ell)}^\intercal \boldsymbol{\theta}_{\mathcal{L}_j(\ell)})]\mathbf{X}_{\mathcal{L}_j(\ell)}$. By the mean value theorem, (A.5) simplifies to the following:

$$\begin{aligned}(\boldsymbol{\vartheta}_t - \boldsymbol{\theta}_t)^\intercal \mathscr{M}_j^{\mathcal{L}}(N_j^\tau) &> (\boldsymbol{\vartheta}_t - \boldsymbol{\theta}_t)^\intercal \left\{ \sum_{\ell=1}^{N_j^\tau} g(\mathbf{X}_{\mathcal{L}_j(\ell)}^\intercal \bar{\boldsymbol{\theta}}_t)\mathbf{X}_{\mathcal{L}_j(\ell)}^\intercal (\boldsymbol{\vartheta}_t - \boldsymbol{\theta}_t)\mathbf{X}_{\mathcal{L}_j(\ell)} + \mathscr{R}_j^{\mathcal{L}} \right\} \\ &\geq \mu_{\min}(\boldsymbol{\vartheta}_t - \boldsymbol{\theta}_t)^\intercal \mathscr{F}_j^{\mathcal{L}}(N_j^\tau)(\boldsymbol{\vartheta}_t - \boldsymbol{\theta}_t) + (\boldsymbol{\vartheta}_t - \boldsymbol{\theta}_t)^\intercal \mathscr{R}_j^{\mathcal{L}}, \end{aligned} \tag{A.6}$$

where $\mu_{\min} = \min\{g(\mathbf{X}^\intercal \boldsymbol{\vartheta}) : p \in \mathscr{P}, \boldsymbol{\vartheta} \in \Theta\} > 0$ and $\bar{\boldsymbol{\theta}}_t$ is on the line segment connecting $\boldsymbol{\vartheta}_t$ and $\boldsymbol{\theta}_t$.

Denote the ball centered at $\boldsymbol{\theta}$ with radius r by $B_r(\boldsymbol{\theta}) = \{\boldsymbol{\vartheta} : ||\boldsymbol{\vartheta} - \boldsymbol{\theta}|| \leq r\}$ and its boundary by $\partial B_r(\boldsymbol{\theta}) = \{\boldsymbol{\vartheta} : ||\boldsymbol{\vartheta} - \boldsymbol{\theta}|| = r\}$. Let $r = (K_3\delta)^{1/2}$. For any $\boldsymbol{\vartheta} \in \partial B_r(\boldsymbol{\theta}_t)$, define $\{Z_{\boldsymbol{\vartheta},j}^{\mathcal{L}}(u) : u = 1,\ldots,N_j^\tau\}$ as $Z_{\boldsymbol{\vartheta},j}^{\mathcal{L}}(1) = 1$ and

$$Z_{j,\boldsymbol{\vartheta}}^{\mathcal{L}}(u) = \exp\left\{ \frac{1}{2}\psi\mu_{\min}(\boldsymbol{\vartheta} - \boldsymbol{\theta}_t)^\intercal \mathscr{M}_j^{\mathcal{L}}(u) - \frac{1}{4}\psi\mu_{\min}^2(\boldsymbol{\vartheta} - \boldsymbol{\theta}_t)^\intercal \mathscr{F}_j^{\mathcal{L}}(u)(\boldsymbol{\vartheta} - \boldsymbol{\theta}_t) \right\}, \tag{A.7}$$

for $u = 2,\ldots,N_j^\tau$, where $\psi = \frac{2}{\varrho_0\sigma_0^2} \wedge \frac{2\varpi_0}{\mu_{\min} d_*(1+p_{\max})}$. Let $\mathcal{F}_j^{\mathcal{L}}(u) = \sigma(\varepsilon_{\mathcal{L}_j(1)},\ldots,\varepsilon_{\mathcal{L}_j(u)})$. Clearly, $Z_{j,\boldsymbol{\vartheta}}^{\mathcal{L}}(u) \in \mathcal{F}_j^{\mathcal{L}}(u)$ for $u = 1,\ldots,N_j^\tau$. Furthermore, by the independence of the noise terms, we have

$$\begin{aligned}&\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}[Z_{j,\boldsymbol{\vartheta}}^{\mathcal{L}}(u) | \mathcal{F}_j^{\mathcal{L}}(u-1)] \\ &\quad= \exp\left\{ \frac{1}{2}\psi\mu_{\min}(\boldsymbol{\vartheta} - \boldsymbol{\theta}_t)^\intercal \mathscr{M}_j^{\mathcal{L}}(u-1) - \frac{1}{4}\psi\mu_{\min}^2(\boldsymbol{\vartheta} - \boldsymbol{\theta}_t)^\intercal \mathscr{F}_j^{\mathcal{L}}(u)(\boldsymbol{\vartheta} - \boldsymbol{\theta}_t) \right\} \\ &\qquad\cdot \mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\exp\left\{ \frac{1}{2}\psi\mu_{\min}(\boldsymbol{\vartheta} - \boldsymbol{\theta}_t)^\intercal \mathbf{X}_{\mathcal{L}_j(u)}\varepsilon_{\mathcal{L}_j(u)} \right\} \middle| \mathcal{F}_j^{\mathcal{L}}(u-1) \right] \\ &\quad= Z_{\boldsymbol{\vartheta},j}^{\mathcal{L}}(u-1)\exp\left\{ -\frac{1}{4}\psi\mu_{\min}^2[(\boldsymbol{\vartheta} - \boldsymbol{\theta}_t)^\intercal \mathbf{X}_{\mathcal{L}_j(u)}]^2 \right\} \\ &\qquad\cdot \mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\exp\left\{ \frac{1}{2}\psi\mu_{\min}(\boldsymbol{\vartheta} - \boldsymbol{\theta}_t)^\intercal \mathbf{X}_{\mathcal{L}_j(u)}\varepsilon_{\mathcal{L}_j(u)} \right\} \right]. \end{aligned} \tag{A.8}$$

Recall that the demand shocks ε_t are i.i.d. with a light-tailed distribution and a variance bounded by σ_0^2, i.e., there exist $\varpi_0, \varrho_0 > 0$ such that $\mathbb{E}_\varepsilon[\exp(\varpi\varepsilon_t)] \le \exp(\frac{1}{2}\varrho_0\sigma_0^2\varpi^2)$ for all ϖ with $|\varpi| \le \varpi_0$ and $t = 1,\ldots,T$. Since $|\frac{1}{2}\psi\mu_{\min}(\boldsymbol{\vartheta}-\boldsymbol{\theta}_t)^\intercal \mathbf{X}_{\mathcal{L}_j(u)}| \le \frac{1}{2}\psi\mu_{\min}||\boldsymbol{\vartheta}-\boldsymbol{\theta}_t||\,||\mathbf{X}_{\mathcal{L}_j(u)}|| < \frac{1}{2}\psi\mu_{\min}d_*(1+p_{\max}) \le \varpi_0$, we further conclude that

$$\begin{aligned}\mathbb{E}^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\exp\left\{\frac{1}{2}\psi\mu_{\min}(\boldsymbol{\vartheta}-\boldsymbol{\theta}_t)^\intercal \mathbf{X}_{\mathcal{L}_j(u)}\varepsilon_{\mathcal{L}_j(u)}\right\}\right] &\le \exp\left\{\frac{1}{8}\varrho_0\sigma_0^2\psi^2\mu_{\min}^2\left[(\boldsymbol{\vartheta}-\boldsymbol{\theta}_t)^\intercal \mathbf{X}_{\mathcal{L}_j(u)}\right]^2\right\}\\ &\le \exp\left\{\frac{1}{4}\psi\mu_{\min}^2\left[(\boldsymbol{\vartheta}-\boldsymbol{\theta}_t)^\intercal \mathbf{X}_{\mathcal{L}_j(u)}\right]^2\right\}. \qquad \text{(A.9)}\end{aligned}$$

Combining (A.8) and (A.9), we obtain $\mathbb{E}^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}[Z^{\mathcal{L}}_{j,\boldsymbol{\vartheta}}(u)|\mathcal{F}^{\mathcal{L}}_j(u-1)] \le Z^{\mathcal{L}}_{j,\boldsymbol{\vartheta}}(u-1)$, for $u = 2,\ldots,N^\tau_j$. Note that this automatically implies the integrability of $Z^{\mathcal{L}}_{j,\boldsymbol{\vartheta}}(u)$ for all $u = 1,\ldots,N^\tau_j$. Therefore, we conclude that $\{(Z^{\mathcal{L}}_{j,\boldsymbol{\vartheta}}(u), \mathcal{F}^{\mathcal{L}}_j(u)) : u = 1,\ldots,N^\tau_j\}$ is a supermartingale.

By Lemma 2 in Keskin and Zeevi (2014), the smallest eigenvalue of $\mathscr{F}^{\mathcal{L}}_j(N^\tau_j)$ is no less than

$$\gamma\sum_{\ell=1}^{N^\tau_j}(p_{\mathcal{L}_j(\ell)} - \bar{p}_{N^\tau_j})^2 = \frac{1}{4}\gamma N^\tau_j(\omega_1-\omega_2)^2 = \frac{1}{2}\gamma M_t(\omega_1-\omega_2)^2, \qquad \text{(A.10)}$$

since $p_t = \omega_1$ if $t \in \mathscr{X}_1$ and $p_t = \omega_2$ if $t \in \mathscr{X}_2$, where $\gamma = 2/(1+2p_{\max}-p_{\min})^2$ and $\bar{p}_u = u^{-1}\sum_{\ell=1}^u p_{\mathcal{L}_j(\ell)}$. Consequently, $(\boldsymbol{\vartheta}_t-\boldsymbol{\theta}_t)^\intercal\mathscr{F}^{\mathcal{L}}_j(N^\tau_j)(\boldsymbol{\vartheta}_t-\boldsymbol{\theta}_t) \ge \frac{1}{2}\gamma K_3 M_t(\omega_1-\omega_2)^2\delta$. Furthermore, given the claimed detection cycles, the number of terms in Λ_j (if any) is a fixed constant $|\Lambda_j|$ that does not depend on M_t. We thus have $|(\boldsymbol{\vartheta}_t-\boldsymbol{\theta}_t)^\intercal\mathscr{R}^{\mathcal{L}}_j| \le ||\boldsymbol{\vartheta}-\boldsymbol{\theta}_t||\,||\mathscr{R}^{\mathcal{L}}_j|| \le 2C_1(K_3\delta)^{1/2}\mu_{\max}(1+p_{\max})$ for some positive constant C_1. Hence, (A.6) and (A.7) imply that

$$\begin{aligned}\mathbb{P}^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}(\mathcal{A}_\tau) &\le \mathbb{P}^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}\left\{Z^{\mathcal{L}}_{j,\boldsymbol{\vartheta}_t}(N^\tau_j) > \exp\left[\frac{1}{4}\psi\mu_{\min}^2(\boldsymbol{\vartheta}_t-\boldsymbol{\theta}_t)^\intercal\mathscr{F}^{\mathcal{L}}_j(N^\tau_j)(\boldsymbol{\vartheta}_t-\boldsymbol{\theta}_t)\right.\right.\\ &\qquad\left.\left. + \frac{1}{2}\psi\mu_{\min}(\boldsymbol{\vartheta}_t-\boldsymbol{\theta}_t)^\intercal\mathscr{R}^{\mathcal{L}}_j\right] \text{ for some } \boldsymbol{\vartheta}_t \in \partial B_r(\boldsymbol{\theta}_t)\right\}\\ &\le \mathbb{P}^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}\left\{Z^{\mathcal{L}}_{j,\boldsymbol{\vartheta}_t}(N^\tau_j) > \exp\left[\rho_1 K_3 M_t\delta - \rho_2(K_3\delta)^{1/2}\right] \text{ for some } \boldsymbol{\vartheta}_t \in \partial B_r(\boldsymbol{\theta}_t)\right\}, \qquad \text{(A.11)}\end{aligned}$$

where $\rho_1 = \frac{1}{8}\psi\mu_{\min}^2\gamma(\omega_1-\omega_2)^2$ and $\rho_2 = C_1\mu_{\max}(1+p_{\max})\psi\mu_{\min}$. Taking $K_3 \ge 4\rho_2^2/\rho_1^2$ and $\delta = \log M_t/M_t$, we have $\rho_2(K_3\delta)^{1/2} \le \frac{1}{2}\rho_1 K_3 M_t\delta$. (A.11) thus reduces to the following:

$$\mathbb{P}^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}(\mathcal{A}_\tau) \le \mathbb{P}^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}\left\{Z^{\mathcal{L}}_{j,\boldsymbol{\vartheta}_t}(N^\tau_j) > e^{\frac{1}{2}\rho_1 K_3\log M_t} \text{ for some } \boldsymbol{\vartheta}_t \in \partial B_r(\boldsymbol{\theta}_t)\right\}. \qquad \text{(A.12)}$$

To make use of (A.12), we define a discretized counterpart of $\partial B_r(\boldsymbol{\theta}_t)$ as

$$\tilde{\partial}B_r(\boldsymbol{\theta}_t) = \left\{(r\cos\phi_\varsigma, r\sin\phi_\varsigma) : r = \left(K_3\frac{\log M_t}{M_t}\right)^{1/2}, \phi_\varsigma = \frac{2\pi\varsigma}{\lceil M_t r\rceil}, \varsigma = 1,\ldots,\lceil M_t r\rceil\right\}, \qquad \text{(A.13)}$$

and $f^{\mathcal{L}}_{j,\tau}(\boldsymbol{\vartheta}) = (\boldsymbol{\vartheta}-\boldsymbol{\theta}_t)^\intercal\mathscr{M}^{\mathcal{L}}_j(N^\tau_j) - \frac{1}{2}\mu_{\min}(\boldsymbol{\vartheta}-\boldsymbol{\theta}_t)^\intercal\mathscr{F}^{\mathcal{L}}_j(N^\tau_j)(\boldsymbol{\vartheta}-\boldsymbol{\theta}_t)$ for all $\boldsymbol{\vartheta} \in \mathbb{R}^2$. For all $\boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}} \in \partial B_r(\boldsymbol{\theta}_t)$ satisfying $||\boldsymbol{\vartheta}-\tilde{\boldsymbol{\vartheta}}|| \le 2\pi/M_t$, we have

$$\begin{aligned}f^{\mathcal{L}}_{j,\tau}(\boldsymbol{\vartheta}) - f^{\mathcal{L}}_{j,\tau}(\tilde{\boldsymbol{\vartheta}}) &\le (\boldsymbol{\vartheta}-\tilde{\boldsymbol{\vartheta}})^\intercal\left[\mathscr{M}^{\mathcal{L}}_j(N^\tau_j) - \mu_{\min}\mathscr{F}^{\mathcal{L}}_j(N^\tau_j)(\boldsymbol{\vartheta}-\tilde{\boldsymbol{\vartheta}})\right]\\ &\le ||\boldsymbol{\vartheta}-\tilde{\boldsymbol{\vartheta}}||\left[||\mathscr{M}^{\mathcal{L}}_j(N^\tau_j)|| + \mu_{\min}||\mathscr{F}^{\mathcal{L}}_j(N^\tau_j)(\boldsymbol{\vartheta}-\tilde{\boldsymbol{\vartheta}})||\right]\end{aligned}$$

$$\leq \frac{2\pi}{M_t}\left[||\mathscr{M}_j^{\mathcal{L}}(N_j^\tau)|| + \mu_{\min}(1+p_{\max}^2)N_j^\tau d_*\right], \tag{A.14}$$

where the first inequality follows by the concavity of $f_{j,\tau}^{\mathcal{L}}(\boldsymbol{\vartheta})$ in $\boldsymbol{\vartheta}$, the second inequality by the Cauchy-Schwarz inequality, and the last by the fact that the maximum eigenvalue of $\mathscr{F}_j^{\mathcal{L}}(N_j^\tau)$ is no greater than $\text{tr}(\mathscr{F}_j^{\mathcal{L}}(N_j^\tau)) \leq (1+p_{\max}^2)N_j^\tau$. Hence, on the event $\{||\mathscr{M}_j^{\mathcal{L}}(N_j^\tau)|| \leq N_j^\tau d_*\}$, we have $f_{j,\tau}^{\mathcal{L}}(\boldsymbol{\vartheta}) - f_{j,\tau}^{\mathcal{L}}(\tilde{\boldsymbol{\vartheta}}) \leq 4\pi(1+\mu_{\min}+\mu_{\min}p_{\max}^2)d_*$, and therefore $Z_{j,\tilde{\boldsymbol{\vartheta}}}^{\mathcal{L}}(N_j^\tau) \geq \underline{c} Z_{j,\boldsymbol{\vartheta}}^{\mathcal{L}}(N_j^\tau)$ for all $\boldsymbol{\vartheta}, \tilde{\boldsymbol{\vartheta}} \in \partial B_r(\boldsymbol{\theta}_t)$ such that $||\boldsymbol{\vartheta} - \tilde{\boldsymbol{\vartheta}}|| \leq 2\pi/M_t$, where $\underline{c} = \exp\{-2\pi\psi\mu_{\min}(1+\mu_{\min}+\mu_{\min}p_{\max}^2)d_*\}$. As a result, (A.12) implies that

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\xi}}^{\pi}(\mathcal{A}_\tau) &\leq \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\xi}}^{\pi}\left\{Z_{j,\boldsymbol{\vartheta}_t}^{\mathcal{L}}(N_j^\tau) > e^{\frac{1}{2}\rho_1 K_3 \log M_t} \text{ for some } \boldsymbol{\vartheta}_t \in \partial B_r(\boldsymbol{\theta}_t), ||\mathscr{M}_j^{\mathcal{L}}(N_j^\tau)|| \leq N_j^\tau d_*\right\} \\
&\qquad + \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\xi}}^{\pi}\left\{||\mathscr{M}_j^{\mathcal{L}}(N_j^\tau)|| > N_j^\tau d_*\right\} \\
&\leq \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\xi}}^{\pi}\left\{Z_{j,\boldsymbol{\vartheta}_t}^{\mathcal{L}}(N_j^\tau) > e^{\frac{1}{2}\rho_1 K_3 \log M_t} \text{ for some } \boldsymbol{\vartheta}_t \in \partial B_r(\boldsymbol{\theta}_t), ||\mathscr{M}_j^{\mathcal{L}}(N_j^\tau)|| \leq N_j^\tau d_*\right\} + 4e^{-\rho_3 M_t} \\
&\leq \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\xi}}^{\pi}\left\{Z_{j,\tilde{\boldsymbol{\vartheta}}_t}^{\mathcal{L}}(N_j^\tau) > \underline{c} e^{\frac{1}{2}\rho_1 K_3 \log M_t} \text{ for some } \tilde{\boldsymbol{\vartheta}}_t \in \tilde{\partial} B_r(\boldsymbol{\theta}_t)\right\} + 4e^{-\rho_3 M_t},
\end{aligned} \tag{A.15}$$

where $\rho_3 = \left[\left(\frac{1}{2\varrho_0\sigma_0^2}\right) \wedge \left(\frac{1}{\sqrt{2}}\varpi_0\right) \wedge \left(\frac{1}{\sqrt{2}}p_{\max}^2\varpi_0\right)\right] d_*(1 \wedge d_*)$. The second inequality follows from Lemma A.1 in den Boer and Keskin (2020), and the third one follows because there exists $\tilde{\boldsymbol{\vartheta}} \in \tilde{\partial} B_r(\boldsymbol{\theta}_t)$ such that $||\boldsymbol{\vartheta} - \tilde{\boldsymbol{\vartheta}}|| \leq 2\pi/M_t$ for all $\boldsymbol{\vartheta} \in \partial B_r(\boldsymbol{\theta}_t)$. Therefore, we further deduce that

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\xi}}^{\pi}(\mathcal{A}_\tau) &\leq \sum_{\tilde{\boldsymbol{\vartheta}}_t \in \tilde{\partial} B_r(\boldsymbol{\theta}_t)} \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\xi}}^{\pi}\left\{Z_{j,\tilde{\boldsymbol{\vartheta}}_t}^{\mathcal{L}}(N_j^\tau) > \underline{c} e^{\frac{1}{2}\rho_1 K_3 \log M_t}\right\} + 4e^{-\rho_3 M_t} \\
&\leq \sum_{\tilde{\boldsymbol{\vartheta}}_t \in \tilde{\partial} B_r(\boldsymbol{\theta}_t)} \underline{c}^{-1} e^{-\frac{1}{2}\rho_1 K_3 \log M_t} + 4e^{-\rho_3 M_t} \\
&\leq 2d_* M_t \underline{c}^{-1} e^{-\frac{1}{2}\rho_1 K_3 \log M_t} + 4e^{-\rho_3 M_t} \\
&= 2d_* \underline{c}^{-1} e^{-(\frac{1}{2}\rho_1 K_3 - 1)\log M_t} + 4e^{-\rho_3 M_t}.
\end{aligned} \tag{A.16}$$

The first inequality simply invokes the union bound, the second makes use of the Markov's inequality and the property of supermartingales, and the third follows because the cardinality of $\tilde{\partial} B_r(\boldsymbol{\theta}_t)$ is $\lceil M_t r \rceil \leq 2d_* M_t$. Note that there exists a positive constant C_2 such that $e^{-\rho_3 M_t} \leq C_2/M_t$. By choosing $K_3 = \frac{2}{\rho_1} \vee \frac{4\rho_2^2}{\rho_1^2}$, we conclude that $\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\xi}}^{\pi}(\mathcal{A}_\tau) \leq K_4/M_t$, where $K_4 = 2d_*\underline{c}^{-1} + 4C_2$. Q.E.D.

Proof of Proposition 3. Because the density $f_z(z;\boldsymbol{\phi}) = B(z)\exp[\boldsymbol{\phi}^\intercal \mathbf{T}(z) - A(\boldsymbol{\phi})]$ must integrate to 1, we have

$$A(\boldsymbol{\phi}) = \log\left(\int_{-\infty}^{\infty} B(z)\exp[\boldsymbol{\phi}^\intercal \mathbf{T}(z)]\mathrm{d}z\right). \tag{A.17}$$

Taking derivative with respect to $\boldsymbol{\phi}$ and using the Leibniz rule, we obtain the following:

$$\begin{aligned}
\nabla A(\boldsymbol{\phi}) &= e^{-A(\boldsymbol{\phi})} \frac{\partial}{\partial \boldsymbol{\phi}} \left(\int_{-\infty}^{\infty} B(z)\exp[\boldsymbol{\phi}^\intercal \mathbf{T}(z)]\mathrm{d}z\right) \\
&= \int_{-\infty}^{\infty} B(z) e^{-A(\boldsymbol{\phi})} \exp[\boldsymbol{\phi}^\intercal \mathbf{T}(z)]\mathbf{T}(z)\mathrm{d}z \\
&= \int_{-\infty}^{\infty} f_z(z;\boldsymbol{\phi})\mathbf{T}(z)\mathrm{d}z = \mathbb{E}_{\boldsymbol{\phi}}[\mathbf{T}(z)].
\end{aligned} \tag{A.18}$$

Again by the Leibniz rule, the Hessian matrix of $A(\boldsymbol{\phi})$ is

$$\begin{aligned}\boldsymbol{H}_A(\boldsymbol{\phi}) &= \int_{-\infty}^{\infty} f_z(z;\boldsymbol{\phi})\mathbf{T}(z)\frac{\partial}{\partial \boldsymbol{\phi}^{\intercal}}[\boldsymbol{\phi}^{\intercal}\mathbf{T}(z) - A(\boldsymbol{\phi})]\mathrm{d}z \\ &= \int_{-\infty}^{\infty} f_z(z;\boldsymbol{\phi})\left[\mathbf{T}(z) - \nabla A(\boldsymbol{\phi})\right]\left[\mathbf{T}(z) - \nabla A(\boldsymbol{\phi})\right]^{\intercal}\mathrm{d}z \\ &\quad + \left[\nabla A(\boldsymbol{\phi})\right]\int_{-\infty}^{\infty} f_z(z;\boldsymbol{\phi})\left[\mathbf{T}(z) - \nabla A(\boldsymbol{\phi})\right]^{\intercal}\mathrm{d}z \\ &= \mathrm{Var}_{\boldsymbol{\phi}}[\mathbf{T}(z)]. \end{aligned} \tag{A.19}$$

The last equality follows by (A.18) and the passage of differentiation under the integral sign follows from the dominated convergence theorem. Furthermore, note that the Fisher Information $\boldsymbol{I}(\boldsymbol{\phi})$ equals $\boldsymbol{H}_A(\boldsymbol{\phi})$ for the exponential family of distributions because

$$\begin{aligned}\boldsymbol{I}(\boldsymbol{\phi}) &\equiv \mathbb{E}_{\boldsymbol{\phi}}\left[\nabla_{\boldsymbol{\phi}} \log f_z(z;\boldsymbol{\phi}) \cdot \nabla_{\boldsymbol{\phi}} \log f_z(z;\boldsymbol{\phi})^{\intercal}\right] \\ &= \mathbb{E}_{\boldsymbol{\phi}}\left[(\mathbf{T}(z) - \nabla A(\boldsymbol{\phi}))(\mathbf{T}(z) - \nabla A(\boldsymbol{\phi}))^{\intercal}\right] \\ &= \mathrm{Var}_{\boldsymbol{\phi}}[\mathbf{T}(z)] = \boldsymbol{H}_A(\boldsymbol{\phi}). \end{aligned} \tag{A.20}$$

Since all components of $\mathbf{T}(z)$ are linearly independent of each other, $\mathrm{Var}_{\boldsymbol{\phi}}[\mathbf{T}(z)]$ is positive definite, and thus, $A(\cdot)$ is strictly convex. Additionally, the Frobenius norm of $\boldsymbol{I}(\boldsymbol{\phi})$ is

$$\|\boldsymbol{I}(\boldsymbol{\phi})\|_F = \sqrt{\mathrm{tr}(\boldsymbol{I}(\boldsymbol{\phi})^{\intercal}\boldsymbol{I}(\boldsymbol{\phi}))} = \sqrt{\sum_{i=1}^{d}\lambda_i^2(\boldsymbol{\phi})}\,,$$

where $\lambda_i(\boldsymbol{\phi})$ is the i^{th} largest eigenvalue of $\boldsymbol{I}(\boldsymbol{\phi})$. Note that $\lambda_1(\boldsymbol{\phi})$ is bounded for all $\boldsymbol{\phi} \in \Phi$ because it is a continuous function of $\boldsymbol{\phi}$. Hence, $\boldsymbol{I}(\boldsymbol{\phi})$ is bounded. With the additional regularity condition that $\sup_{\boldsymbol{\phi}\in\Phi}\mathbb{E}_{\boldsymbol{\phi}}\|\nabla_{\boldsymbol{\phi}} \log f_z(z;\boldsymbol{\phi})\|^{\ell} < \infty$ for some $\ell > d$, we deduce from Theorem 36.3 in Borovkov (1998) that for any $v > 0$ there exist positive constants C_3 and C_4 such that

$$\mathbb{P}_{\boldsymbol{\phi}}\left\{\sqrt{k}\|\check{\boldsymbol{\phi}}_k - \boldsymbol{\phi}\| \geq v\right\} \leq C_3 e^{-C_4 v^2}\,. \tag{A.21}$$

We complete the proof by taking $v = \sqrt{K_5 \log k}$, $K_5 = 1/C_4$, $K_6 = C_3$, and noting that $||\hat{\boldsymbol{\phi}}_k - \boldsymbol{\phi}|| \leq ||\check{\boldsymbol{\phi}}_k - \boldsymbol{\phi}||$. Q.E.D.

Proof of Proposition 4. We complete the proof in two steps.

Step 1: Proving that $H(y_t^u; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t)$ and $\tilde{H}_t(y_t^u; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t, \boldsymbol{\varepsilon})$ are close with high probability. For notational brevity, given $p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t$ in period $t \notin \mathscr{X}$, let $y_t^u = y_t^u(p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t)$ be the unconstrained optimizer of (5). Plugging y_t^u into $H(\cdot\,; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t)$ and $\tilde{H}_t(\cdot\,; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t, \boldsymbol{\varepsilon})$, respectively, we have

$$H(y_t^u; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t) = \mathbb{E}_{\varepsilon, q|\hat{\boldsymbol{\xi}}_t}\left\{(h-c)[z_t^u(q_t) - \varepsilon_t]^+ + (b+p_t)[\varepsilon_t - z_t^u(q_t]^+\right\} + (w\mathbb{E}_{q|\hat{\boldsymbol{\xi}}}[q_t] + c)y_t^u,$$

and

$$\tilde{H}_t(y_t^u; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t, \boldsymbol{\varepsilon}) = \frac{1}{2M_t}\sum_{\substack{s=nL(\tau)+1 \\ s\in\mathscr{X}}}^{t} \mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}\left[(h-c)[z_t^u(q_t) - \varepsilon_s]^+ + (b+p_t)[\varepsilon_s - z_t^u(q_t)]^+\right]$$

$$+ (w\mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[q_t] + c)y_t^u,$$

where $z_t^u(q_t) = z_t^u(q_t; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t) = (1-q_t)y_t^u(p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t) - g(\mathbf{X}_t^\top \boldsymbol{\theta}_t)$ for any given q_t. Given $p_t \in \mathscr{P}$ in period $t \notin \mathscr{X}$ and $s \in \{nL(\tau)+1, \ldots, t\} \cap \mathscr{X}$, define

$$\begin{aligned}\Omega_s^t(p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t) = (h-c)\left[\mathbb{E}_{\varepsilon, q|\hat{\boldsymbol{\xi}}_t}[z_t^u(q_t; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t) - \varepsilon_t]^+ - \mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[z_t^u(q_t; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t) - \varepsilon_s]^+\right] \\ + (b+p_t)\left[\mathbb{E}_{\varepsilon, q|\hat{\boldsymbol{\xi}}_t}[\varepsilon_t - z_t^u(q_t; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t)]^+ - \mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[\varepsilon_s - z_t^u(q_t; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t)]^+\right], \qquad \text{(A.22)}\end{aligned}$$

where $z_t^u(q_t; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t) = (1-q_t)y_t^u(p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t) - g(\mathbf{X}_t^\top \boldsymbol{\theta}_t)$. Note that $\mathbb{E}_\varepsilon[\Omega_s^t(p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t)] = 0$.

The following lemma provides several useful preliminary results to establish the bound on regret due to replacing the cumulative distribution function F_ε with its sample average approximation. Lemma A.2(a) provides an upper bound on the absolute moment of the noise terms, Lemma A.2(b) shows that $\Omega_s^t(p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t)$ has a light-tailed distribution, and Lemma A.2(c) provides an exponential bound on its tail probability.

LEMMA A.2. *(a) For any positive integer k and any period t,*

$$\mathbb{E}_\varepsilon[|\varepsilon_t|^k] \le k(2\varrho_0\sigma_0^2)^{k/2}\Gamma(k/2) + 2^{k+1}k\varpi_0^{-k}\Gamma(k), \qquad \text{(A.23)}$$

where σ_0, ϱ_0, and ϖ_0 are given in the characterization of the distribution of $\{\varepsilon_t\}$.

(b) Given $p_t \in \mathscr{P}$ in period $t \notin \mathscr{X}$ and $s = nL(\tau)+1, \ldots, t$ and $s \in \mathscr{X}$, there exists a positive constant C_5 such that

$$\mathbb{E}_\varepsilon[\exp(\varpi\Omega_s^t(p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t))] \le C_5 \exp\left(\frac{1}{2}\varrho_0\sigma_0^2(|h-c|+b+p_{\max})^2\varpi^2\right) \qquad \text{(A.24)}$$

for all ϖ with $|\varpi| \le \varpi_0/(|h-c|+b+p_{\max})$.

(c) Given $p_t \in \mathscr{P}$ in period $t \notin \mathscr{X}$ and $s = nL(\tau)+1, \ldots, t$ with $s \in \mathscr{X}$, there exists a positive constant C_6 such that

$$\mathbb{P}_\varepsilon\left\{\left|\sum_{\ell=1}^{u}\Omega_{\mathcal{J}_\tau(\ell)}^t(p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t)\right| \ge \delta u\right\} \le \frac{2}{C_5}\exp\left(-C_6(\delta \wedge \delta^2)u\right), \qquad \text{(A.25)}$$

for any $\delta > 0$ and any $u = 1, 2, \ldots, 2m(\tau - L(\tau)+1)$, where τ is the cycle to which period t belongs and $\mathcal{J}_\tau(\ell) = nL(\tau) + (n-2m)\lfloor(\ell-1)/m\rfloor + \ell$.

Proof of Lemma A.2. We prove each part separately.

(a) Because the noise term has a light-tailed distribution, there exist $\varpi_0, \varrho_0 > 0$ such that $\mathbb{E}_\varepsilon[\exp(\varpi\varepsilon_t)] \leq \exp(\frac{1}{2}\varrho_0\sigma_0^2\varpi^2)$ for all ϖ with $|\varpi| \leq \varpi_0$ and $t = 1, \ldots, T$. Thus, by Markov inequality, for any $\delta > 0$,

$$\mathbb{P}_\varepsilon\{\varepsilon_t > \delta\} \leq e^{-\varpi\delta}\mathbb{E}_\varepsilon[\exp(\varpi\varepsilon_t)] \leq \exp\left(\frac{1}{2}\varrho_0\sigma_0^2\varpi^2 - \varpi\delta\right),$$

for any $\varpi \in (0, \varpi_0]$. We then minimize the right-hand side in the preceding inequality to obtain the tightest upper bound. If $\delta \leq \varpi_0\varrho_0\sigma_0^2$, then the minimum of the right-hand side is $\exp\left(-\frac{\delta^2}{2\varrho_0\sigma_0^2}\right)$, achieved at $\varpi_0 = \frac{\delta}{\varrho_0\sigma_0^2}$; if $\delta > \varpi_0\varrho_0\sigma_0^2$, then the minimum is achieved at $\varpi = \varpi_0$ and $\mathbb{P}_\varepsilon\{\varepsilon_t > \delta\} \leq \exp\left(\frac{1}{2}\varrho_0\sigma_0^2\varpi_0^2 - \varpi_0\delta\right) \leq \exp\left(-\frac{1}{2}\varpi_0\delta\right)$. Similar analysis applies to the event $\{\varepsilon_t < -\delta\}$. We thus conclude that

$$\mathbb{P}_\varepsilon\{|\varepsilon_t| > \delta\} \leq \begin{cases} 2\exp\left(-\frac{1}{2}\delta\varpi_0\right) & \text{if } \delta > \varpi_0\varrho_0\sigma_0^2, \\ 2\exp\left(-\frac{\delta^2}{2\varrho_0\sigma_0^2}\right) & \text{if } \delta \leq \varpi_0\varrho_0\sigma_0^2. \end{cases}$$

Consequently, for any positive integer k, we have

$$\begin{aligned}
\mathbb{E}_\varepsilon[|\varepsilon_t|^k] &= \int_0^\infty \mathbb{P}_\varepsilon\{|\varepsilon_t|^k > \delta\}\,\mathrm{d}\delta \\
&\leq \int_0^{(\varpi_0\varrho_0\sigma_0^2)^k} 2\exp\left(-\frac{\delta^{2/k}}{2\varrho_0\sigma_0^2}\right)\mathrm{d}\delta + \int_{(\varpi_0^2\varrho_0\sigma_0^2)^k}^\infty 2\exp\left(-\frac{1}{2}\delta^{1/k}\varpi_0\right)\mathrm{d}\delta \\
&\leq k(2\varrho_0\sigma_0^2)^{k/2}\int_0^\infty e^{-u}u^{k/2-1}\,\mathrm{d}u + 2^{k+1}k(\varpi_0)^{-k}\int_0^\infty e^{-v}v^{k-1}\,\mathrm{d}v \\
&= k(2\varrho_0\sigma_0^2)^{k/2}\Gamma(k/2) + 2^{k+1}k\varpi_0^{-k}\Gamma(k),
\end{aligned}$$

where the second inequality follows by the expansion of the upper and lower limit of the integrals and the change of variables $u = \frac{\delta^{2/k}}{2\varrho_0\sigma_0^2}$ and $v = \frac{1}{2}\delta^{1/k}\varpi_0$.

(b) For any $s \in \mathscr{X}$ with $s = nL(\tau) + 1, \ldots, t$, there are constants C_7 and C_8 such that

$$\begin{aligned}
|\Omega_s^t(p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t)| &\leq |h-c|\left(\mathbb{E}_{\varepsilon,q|\hat{\boldsymbol{\xi}}_t}[z_t^u(q_t; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t) - \varepsilon_t]^+ + \mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[z_t^u(q_t; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t) - \varepsilon_s]^+\right) \\
&\quad + (b + p_t)\left(\mathbb{E}_{\varepsilon,q|\hat{\boldsymbol{\xi}}_t}[\varepsilon_t - z_t^u(q_t; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t)]^+ + \mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[\varepsilon_s - z_t^u(q_t; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t)]^+\right) \\
&\leq (|h-c| + b + p_t)\left(2\mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}|z_t^u(q_t; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t)| + \mathbb{E}_\varepsilon|\varepsilon_t| + |\varepsilon_s|\right) \\
&\leq C_7 + C_8|\varepsilon_s|,
\end{aligned}$$

where $C_7 = C_8[2(y_{\max} + \mu_{\max}) + (2\pi\varrho_0\sigma_0^2)^{1/2} + \frac{4}{\varpi_0}]$, $C_8 = |h-c| + b + p_{\max}$, and $\mu_{\max} = \max\{g(\mathbf{X}^\intercal\boldsymbol{\vartheta}) : p \in \mathscr{P}, \boldsymbol{\vartheta} \in \Theta\}$. In the above derivation, the first inequality follows from the triangle inequality, the second inequality follows because $a^+ \leq |a|$ for any real number a, and the last inequality makes use of Lemma A.2(a) and the fact that $|z_t^u(q_t; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t)| \leq y_{\max} + \mu_{\max}$. Pick any $\varpi \in [-\varpi_0, \varpi_0]$. Because $\mathbb{E}_\varepsilon[\exp(\varpi\varepsilon_s)] \leq \exp(\frac{1}{2}\varrho_0\sigma_0^2\varpi^2)$, we deduce that

$$\begin{aligned}
\mathbb{E}_\varepsilon[\exp(\varpi|\varepsilon_s|)] &= \mathbb{E}_\varepsilon[\exp(\varpi\varepsilon_s)\mathbb{I}\{\varepsilon_s \geq 0\}] + \mathbb{E}_\varepsilon[\exp(-\varpi\varepsilon_s)\mathbb{I}\{\varepsilon_s < 0\}] \\
&\leq \mathbb{E}_\varepsilon[\exp(\varpi\varepsilon_s)] + \mathbb{E}_\varepsilon[\exp(-\varpi\varepsilon_s)]
\end{aligned}$$

$$\leq 2\exp\left(\frac{1}{2}\varrho_0\sigma_0^2\varpi^2\right).$$

As a result, $\mathbb{E}_\varepsilon[\exp(\varpi\Omega_s^t(p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t))] \leq C_5\exp(\frac{1}{2}\varrho_0\sigma_0^2C_8\varpi^2)$ for all $|\varpi| \leq \varpi_0/C_8$, where $C_5 = 2\exp([2(y_{\max}+\mu_{\max})+(2\pi\varrho_0\sigma_0^2)^{1/2}]\varpi_0+4)$.

(c) Fix $\delta>0$. Given $p_t\in\mathscr{P}$ in period $t\notin\mathscr{X}$, which belongs to cycle $\tau=\lceil t/n\rceil-1$, we first note that there exists a one-to-one correspondence between $\{s\in\mathscr{X}: s=nL(\tau)+1,\ldots,t\}$ and $\{1,2,\ldots,2m(\tau-L(\tau)+1)\}$ represented by $\mathcal{J}_\tau(\cdot)$ satisfying $\mathcal{J}_\tau(\ell)=nL(\tau)+(n-2m)\lfloor(\ell-1)/m\rfloor+\ell$. Define a stochastic process $\{Z_\tau(u): u=0,1,\ldots,2m(\tau-L(\tau)+1)\}$ with $Z_\tau(0)=1$ and

$$Z_\tau(u)=C_5\exp\left\{\frac{1}{\zeta}\left(2\delta\sum_{\ell=1}^{u}\Omega^t_{\mathcal{J}_\tau(\ell)}(p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)-\delta^2u\right)\right\}, \tag{A.26}$$

with the filtration $\mathcal{F}_\tau(u)=\sigma(\varepsilon_{\mathcal{J}_\tau(1)},\ldots,\varepsilon_{\mathcal{J}_\tau(u)})$ for $u=1,2,\ldots,2m(\tau-L(\tau)+1)$, where $\zeta=\frac{2\delta C_8}{\varpi_0}\vee(2C_8^2\varrho_0\sigma_0^2)$. Note that $Z_\tau(u)$ is integrable since $\Omega^t_{\mathcal{J}_\tau(\ell)}(p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)$ are i.i.d. and $2\delta/\zeta\leq\varpi_0/C_8$. Additionally, we have

$$\begin{aligned}
\mathbb{E}_\varepsilon[Z_\tau(u)|\mathcal{F}_\tau(u-1)] &= C_5\exp\left(\frac{1}{\zeta}\left(2\delta\sum_{\ell=1}^{u-1}\Omega^t_{\mathcal{J}_\tau(\ell)}(p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)-\delta^2u\right)\right)\\
&\quad\cdot\mathbb{E}_\varepsilon\left[\exp\left(\frac{2}{\zeta}\delta\Omega^t_{\mathcal{J}_\tau(u)}(p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)\right)\middle|\mathcal{F}_\tau(u-1)\right]\\
&\leq C_5\exp\left(\frac{1}{\zeta}\left(2\delta\sum_{\ell=1}^{u-1}\Omega^t_{\mathcal{J}_\tau(\ell)}(p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)-\delta^2u\right)\right)\exp\left(\frac{2\delta^2}{\zeta^2}C_8^2\varrho_0\sigma_0^2\right)\\
&\leq C_5\exp\left(\frac{1}{\zeta}\left(2\delta\sum_{\ell=1}^{u-1}\Omega^t_{\mathcal{J}_\tau(\ell)}(p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)-\delta^2(u-1)\right)\right)=Z_\tau(u-1),
\end{aligned}$$

where first inequality follows by Lemma A.2(b) and the fact that $2\delta/\zeta\leq\varpi_0/C_8$, while the second because $2C_8^2\varrho_0\sigma_0^2\leq\zeta$. Thus, $\{(Z_\tau(u),\mathcal{F}_\tau(u)): u=0,1,\ldots,2m(\tau-L(\tau)+1)\}$ is a supermartingale. As a result, we deduce that

$$\begin{aligned}
\mathbb{P}_\varepsilon\left\{\sum_{\ell=1}^{u}\Omega^t_{\mathcal{J}_\tau(\ell)}(p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)\geq\delta u\right\} &= \mathbb{P}_\varepsilon\left\{2\delta\sum_{\ell=1}^{u}\Omega^t_{\mathcal{J}_\tau(\ell)}(p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)-\delta^2u\geq\delta^2u\right\}\\
&= \mathbb{P}_\varepsilon\left\{Z_\tau(u)\geq C_5\exp\left(\frac{1}{\zeta}\delta^2u\right)\right\}\\
&\leq \frac{1}{C_5}\exp\left(-\frac{1}{\zeta}\delta^2u\right)\\
&= \frac{1}{C_5}\exp\left(-u\left(\frac{\varpi_0\delta}{2C_8}\wedge\frac{\delta^2}{2C_8^2\varrho_0\sigma_0^2}\right)\right)\\
&\leq \frac{1}{C_5}\exp\left(-C_6(\delta\wedge\delta^2)u\right),
\end{aligned}$$

where $C_6=\frac{\varpi_0}{2C_8}\wedge\frac{1}{2C_8^2\varrho_0\sigma_0^2}$, and the first inequality follows by the Markov inequality and the fact that $\{Z_\tau(u),\mathcal{F}_\tau(u)\}$ is a supermartingale with $Z_\tau(0)=1$. By replacing δ with $-\delta$ in (A.26) and following

a similar argument, we obtain

$$\mathbb{P}_\varepsilon\left\{\sum_{\ell=1}^{u}\Omega^t_{\mathcal{J}_\tau(\ell)}(p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)\le-\delta u\right\}\le\frac{1}{C_5}\exp\left(-C_6(\delta\wedge\delta^2)u\right).$$

Combining the last two inequalities completes the proof. Q.E.D.

Having stated and proved Lemma A.2, we now note that

$$H(y^u_t;p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)-\tilde{H}_t(y^u_t;p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon})=\frac{1}{2M_t}\sum_{\substack{s=nL(\tau)+1\\ s\in\mathscr{X}}}^{t}\Omega^t_s(p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t).$$

Hence, by Lemma A.2(c), we have the following for any $\delta>0$:

$$\begin{aligned}\mathbb{P}_\varepsilon\left\{\left|H(y^u_t;p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)-\tilde{H}_t(y^u_t;p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon})\right|>\delta\right\}&=\mathbb{P}_\varepsilon\left\{\left|\sum_{\substack{s=nL(\tau)+1\\ s\in\mathscr{X}}}^{t}\Omega^t_s(p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)\right|>2M_t\delta\right\}\\&\le\frac{2}{C_5}\exp\left(-2C_6M_t(\delta\wedge\delta^2)\right).\end{aligned}\tag{A.27}$$

Step 2: Proving that $\tilde{H}_t(y^u_t;p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon})$ and $\tilde{H}_t(\tilde{y}_t;p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon})$ are close with high probability. Define the empirical distribution of ε_s for $s=nL(\tau)+1,\dots,t$ and $s\in\mathscr{X}$ by

$$\hat{F}^t_\varepsilon(v)=\frac{1}{2M_t}\sum_{s=nL(\tau)+1}^{t}\mathbb{I}\{s\in\mathscr{X}\}\mathbb{I}\{\varepsilon_s\le v\}.\tag{A.28}$$

Because $\mathbb{E}_\varepsilon[\hat{F}^t_\varepsilon(z^u_t(q_t))]=F_\varepsilon(z^u_t(q_t))$ for all q_t and $\mathbb{I}\{\varepsilon_s\le v\}\in[0,1]$ for all s and v, we deduce from Hoeffding's inequality that for any $\delta>0$,

$$\mathbb{P}_\varepsilon\left\{\left|\hat{F}^t_\varepsilon(z^u_t(q_t))-F_\varepsilon(z^u_t(q_t))\right|>\delta\right\}\le2\exp(-4M_t\delta^2).\tag{A.29}$$

Denote by $\tilde{y}^u_t=\tilde{y}^u_t(p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)$ and $\tilde{y}_t=\tilde{y}_t(p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)$ the unconstrained and constrained optimizers of (26), respectively. In addition, let $\tilde{z}^u_t(q_t)=(1-q_t)\tilde{y}^u_t-g(\mathbf{X}_t^\intercal\boldsymbol{\theta}_t)$ and $\tilde{z}_t(q_t)=(1-q_t)\tilde{y}_t-g(\mathbf{X}_t^\intercal\boldsymbol{\theta}_t)$ for any given realized q_t correspondingly. Hence,

$$\tilde{z}_t(q_t)=\min\left\{\max\left\{\tilde{z}^u_t(q_t),(1-q_t)y_{\min}-g(\mathbf{X}_t^\intercal\boldsymbol{\theta}_t)\right\},(1-q_t)y_{\max}-g(\mathbf{X}_t^\intercal\boldsymbol{\theta}_t)\right\}.$$

Let $\tilde{H}_t(y_t;p_t,\boldsymbol{\theta}_t,q_t,\boldsymbol{\varepsilon})$ be obtained by replacing the expectation over q in $\tilde{H}_t(y_t;p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon})$ with the realized q_t when the control (p_t,y_t) is applied and the demand parameter vector is $\boldsymbol{\theta}_t$. Suppose that $y^u_t\le\tilde{y}_t$. We then have

$$\begin{aligned}&\tilde{H}_t(y^u_t;p_t,\boldsymbol{\theta}_t,q_t,\boldsymbol{\varepsilon})\\&\quad=\frac{1}{2M_t}\sum_{s=nL(\tau)+1}^{t}\mathbb{I}\{s\in\mathscr{X}\}\Big[(h-c)(z^u_t(q_t)-\varepsilon_s)\mathbb{I}\{\varepsilon_s\le z^u_t(q_t)\}\\&\qquad\qquad+(b+p_t)(\varepsilon_s-z^u_t(q_t))\mathbb{I}\{z^u_t(q_t)<\varepsilon_s\le\tilde{z}_t(q_t)\}+(b+p_t)(\varepsilon_s-z^u_t(q_t))\mathbb{I}\{\tilde{z}_t(q_t)<\varepsilon_s\}\Big]\\&\qquad\qquad+\frac{wq_t+c}{1-q_t}(z^u_t(q_t)+g(\mathbf{X}_t^\intercal\boldsymbol{\theta}_t)),\end{aligned}\tag{A.30}$$

and

$$\begin{aligned}
&\tilde{H}_t(\tilde{y}_t; p_t, \boldsymbol{\theta}_t, q_t, \boldsymbol{\varepsilon}) \\
&\quad = \frac{1}{2M_t} \sum_{s=nL(\tau)+1}^{t} \mathbb{I}\{s \in \mathscr{X}\} \Big[(h-c)(\tilde{z}_t(q_t) - \varepsilon_s)\mathbb{I}\{\varepsilon_s \le z_t^u(q_t)\} \\
&\qquad\qquad + (h-c)(\tilde{z}_t(q_t) - \varepsilon_s)\mathbb{I}\{z_t^u(q_t) < \varepsilon_s \le \tilde{z}_t(q_t)\} \\
&\qquad\qquad + (b+p_t)(\varepsilon_s - \tilde{z}_t(q_t))\mathbb{I}\{\tilde{z}_t(q_t) < \varepsilon_s\} \Big] + \frac{wq_t + c}{1-q_t}(\tilde{z}_t(q_t) + g(\mathbf{X}_t^\top \boldsymbol{\theta}_t)) \\
&\quad \ge \frac{1}{2M_t} \sum_{s=nL(\tau)+1}^{t} \mathbb{I}\{s \in \mathscr{X}\} \Big[(h-c)(\tilde{z}_t(q_t) - \varepsilon_s)\mathbb{I}\{\varepsilon_s \le z_t^u(q_t)\} \\
&\qquad\qquad - (b+p_t)(\tilde{z}_t(q_t) - \varepsilon_s)\mathbb{I}\{z_t^u(q_t) < \varepsilon_s \le \tilde{z}_t(q_t)\} \\
&\qquad\qquad + (b+p_t)(\varepsilon_s - \tilde{z}_t(q_t))\mathbb{I}\{\tilde{z}_t(q_t) < \varepsilon_s\} \Big] + \frac{wq_t + c}{1-q_t}(\tilde{z}_t(q_t) + g(\mathbf{X}_t^\top \boldsymbol{\theta}_t)),
\end{aligned} \tag{A.31}$$

where the inequality follows from the assumption that $|h-c| < b+p$ for all $p \in \mathscr{P}$. Subtracting (A.31) from (A.30), we obtain the following:

$$\begin{aligned}
&\tilde{H}_t(y_t^u; p_t, \boldsymbol{\theta}_t, q_t, \boldsymbol{\varepsilon}) - \tilde{H}_t(\tilde{y}_t; p_t, \boldsymbol{\theta}_t, q_t, \boldsymbol{\varepsilon}) \\
&\quad \le (h-c)[z_t^u(q_t) - \tilde{z}_t(q_t)]\hat{F}_\varepsilon^t(z_t^u(q_t)) + (b+p_t)[\tilde{z}_t(q_t) - z_t^u(q_t)][1 - \hat{F}_\varepsilon^t(\tilde{z}_t(q_t))] \\
&\qquad + (b+p_t)[\tilde{z}_t(q_t) - z_t^u(q_t)][\hat{F}_\varepsilon^t(\tilde{z}_t(q_t)) - \hat{F}_\varepsilon^t(z_t^u(q_t))] + \frac{wq_t + c}{1-q_t}[z_t^u(q_t) - \tilde{z}_t(q_t)] \\
&\quad = [\tilde{z}_t(q_t) - z_t^u(q_t)] \left[-(h-c+b+p_t)\hat{F}_\varepsilon^t(z_t^u(q_t)) + (b+p_t) - \frac{wq_t + c}{1-q_t} \right] \\
&\quad = (\tilde{y}_t - y_t^u) \left[-(1-q_t)(h-c+b+p_t)\hat{F}_\varepsilon^t(z_t^u(q_t)) + (1-q_t)(b+p_t) - (wq_t + c) \right],
\end{aligned} \tag{A.32}$$

where the last equality follows because $\tilde{z}_t(q_t) - z_t^u(q_t) = (1-q_t)(\tilde{y}_t - y_t^u)$. Taking expectation with respect to q_t on both sides with the estimated perishability parameter $\hat{\boldsymbol{\xi}}_t$ and invoking Lemma A.1, we deduce that

$$\begin{aligned}
&\tilde{H}_t(y_t^u; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t, \boldsymbol{\varepsilon}) - \tilde{H}_t(\tilde{y}_t; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t, \boldsymbol{\varepsilon}) \\
&\quad \le (\tilde{y}_t - y_t^u)(h-c+b+p_t)\mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t} \left\{ (1-q_t)[F_\varepsilon(z_t^u(q_t)) - \hat{F}_\varepsilon^t(z_t^u(q_t))] \right\}.
\end{aligned}$$

Note that $\tilde{H}_t(y_t^u; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t, \boldsymbol{\varepsilon}) \ge \tilde{H}_t(\tilde{y}_t; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t, \boldsymbol{\varepsilon})$ because $\tilde{y}_t$ is the minimizer of $\tilde{H}_t(\cdot)$. A similar reasoning applies when $y_t^u > \tilde{y}_t$, yielding

$$\begin{aligned}
&\tilde{H}_t(\tilde{y}_t; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t, \boldsymbol{\varepsilon}) - \tilde{H}_t(y_t^u; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t, \boldsymbol{\varepsilon}) \\
&\quad \ge (\tilde{y}_t - y_t^u)(h-c+b+p_t)\mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t} \left\{ (1-q_t)[\hat{F}_\varepsilon^t(z_t^u(q_t)^-) - F_\varepsilon(z_t^u(q_t)^-)] \right\}.
\end{aligned}$$

Because (A.29) holds for all q_t, we thus conclude that for any $\delta > 0$,

$$\mathbb{P}_\varepsilon \left\{ \left| \tilde{H}_t(y_t^u; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t, \boldsymbol{\varepsilon}) - \tilde{H}_t(\tilde{y}_t; p_t, \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t, \boldsymbol{\varepsilon}) \right| \ge \delta \right\} \le 2\exp\left(-\frac{4M_t\delta^2}{C_8^2 \overline{d}(\mathscr{Y})^2} \right), \tag{A.33}$$

where $\overline{d}(\mathscr{Y}) = y_{\max} - y_{\min}$.

Combining (A.27) and (A.33), we deduce from the triangle inequality that for any $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}_\varepsilon\left\{\left|G^u(p_t;\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)-\tilde{G}_t(p_t;\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon})\right|\geq 2\delta\right\}\\
&\quad=\mathbb{P}_\varepsilon\left\{\left|H(y_t^u;p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)-\tilde{H}_t(\tilde{y}_t;p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon})\right|\geq 2\delta\right\}\\
&\quad\leq\mathbb{P}_\varepsilon\left\{\left|H(y_t^u;p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)-\tilde{H}_t(y_t^u;p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon})\right|\geq \delta\right\}\\
&\qquad+\mathbb{P}_\varepsilon\left\{\left|\tilde{H}_t(y_t^u;p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon})-\tilde{H}_t(\tilde{y}_t;p_t,\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon})\right|\geq \delta\right\}\\
&\quad\leq\frac{2}{C_5}\exp\left(-2C_6M_t(\delta\wedge\delta^2)\right)+2\exp\left(-\frac{4M_t\delta^2}{C_8^2\,\overline{d}(\mathscr{Y})^2}\right). \qquad \text{(A.34)}
\end{aligned}$$

Let $K_7=2(\frac{1}{\sqrt{2C_6}}\vee\frac{C_8\overline{d}(\mathscr{Y})}{2})$ and $K_8=2(1/C_5+1)$. Taking $\delta=\frac{1}{2}K_7\sqrt{\log M_t/M_t}$ completes the proof. Q.E.D.

Proof of Proposition 5. Given $p_t\in\mathscr{P}$, $\hat{\boldsymbol{\xi}}_t\in\Xi$, and the price-and-demand history $\mathcal{D}_t=\{(p_s,D_s):s=nL(\tau)+1,\ldots,t,s\in\mathscr{X}\}$, let $W_t(\boldsymbol{\vartheta}_t;p_t,\hat{\boldsymbol{\xi}}_t,\mathcal{D}_t)$ be a function of $\boldsymbol{\vartheta}_t$ defined as follows:

$$\begin{aligned}
&W_t(\boldsymbol{\vartheta}_t;p_t,\hat{\boldsymbol{\xi}}_t,\mathcal{D}_t)\\
&\quad=p_tg(\mathbf{X}_t^\top\boldsymbol{\vartheta}_t)-\min_{y_t\in\mathscr{Y}}\left\{\frac{1}{2M_t}\sum_{\substack{s=nL(\tau)+1\\ s\in\mathscr{X}}}^{t}\mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}\Big[(h-c)[(1-q_t)y_t-g(\mathbf{X}_t^\top\boldsymbol{\vartheta}_t)\right.\\
&\qquad\qquad-\mathscr{J}(\boldsymbol{\vartheta}_t;p_s,D_s)]^+\left.+(b+p_t)[g(\mathbf{X}_t^\top\boldsymbol{\vartheta}_t)+\mathscr{J}(\boldsymbol{\vartheta}_t;p_s,D_s)-(1-q_t)y_t]^+\Big]\right.\\
&\qquad\qquad\left.+(w\mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[q_t]+c)y_t\right\}. \qquad \text{(A.35)}
\end{aligned}$$

Note that the noise term $\mathscr{J}(\boldsymbol{\vartheta}_t;p_s,D_s)=D_s-g(\mathbf{X}_s^\top\boldsymbol{\vartheta}_t)$ is viewed as a function of $\boldsymbol{\vartheta}_t$ given (p_s,D_s) with $e_s=\mathscr{J}(\hat{\boldsymbol{\theta}}_t;p_s,D_s)=D_s-g(\mathbf{X}_s^\top\hat{\boldsymbol{\theta}}_t)$ and $\varepsilon_s=\mathscr{J}(\boldsymbol{\theta}_t;p_s,D_s)=D_s-g(\mathbf{X}_s^\top\boldsymbol{\theta}_t)$ because $\boldsymbol{\theta}_s=\boldsymbol{\theta}_t=\boldsymbol{\theta}^*_{t_j}$, for all $s=n\hat{\tau}_j^++1,\ldots,t$ satisfying $s\in\mathscr{X}$. Because the objective function of the above minimization problem is convex in y_t and Lipschitz in $\boldsymbol{\vartheta}_t$, the unconstrained optimizer $y_t^u(\boldsymbol{\vartheta}_t;p_t,\hat{\boldsymbol{\xi}}_t,\mathcal{D}_t)$ is also Lipschitz in $\boldsymbol{\vartheta}_t$ and so is $W_t(\boldsymbol{\vartheta}_t;p_t,\hat{\boldsymbol{\xi}}_t,\mathcal{D}_t)$, that is, there exists a positive constant C_9 such that $|W_t(\boldsymbol{\theta}_t;p_t,\hat{\boldsymbol{\xi}}_t,\mathcal{D}_t)-W_t(\hat{\boldsymbol{\theta}}_t;p_t,\hat{\boldsymbol{\xi}}_t,\mathcal{D}_t)|\leq C_9||\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t||$. We complete the proof by noting that $W_t(\boldsymbol{\theta}_t;p_t,\hat{\boldsymbol{\xi}}_t,\mathcal{D}_t)=\tilde{G}_t(p_t;\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\boldsymbol{\varepsilon})$ and $W_t(\hat{\boldsymbol{\theta}}_t;p_t,\hat{\boldsymbol{\xi}}_t,\mathcal{D}_t)=\hat{G}_t(p_t;\hat{\boldsymbol{\theta}}_t,\hat{\boldsymbol{\xi}}_t,\mathbf{e})$, and invoking Proposition 2 with $K_9=\sqrt{K_3}C_9$ and $K_{10}=K_4$. Q.E.D.

Proof of Proposition 6. By Propositions 4 and 5, there exist constants C_{10} and C_{11} such that

$$\mathbb{P}^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}\left\{\left|G^u(p_t;\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)-\hat{G}_t(p_t;\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\mathbf{e})\right|\geq C_{10}\sqrt{\frac{\log M_t}{M_t}}\right\}\leq\frac{C_{11}}{M_t}$$

for any $p_t\in\mathscr{P}$, where $C_{10}=K_7+K_9$ and $C_{11}=K_8+K_{10}$. By the design of our policy, the step size between grid points on the discretized price space $\mathscr{P}_d$ in period t is given by $\iota_t=\rho\sqrt{(\log M_t)/M_t}$. Thus, we deduce that

$$\mathbb{P}^\pi_{\boldsymbol{\theta},\boldsymbol{\xi}}\left\{\max_{p_t\in\mathscr{P}_d}\left|G^u(p_t;\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t)-\hat{G}_t(p_t;\boldsymbol{\theta}_t,\hat{\boldsymbol{\xi}}_t,\mathbf{e})\right|\geq C_{10}\sqrt{\frac{\log M_t}{M_t}}\right\}\leq\frac{2C_{11}\overline{d}(\mathscr{P})}{M_t\iota_t}, \qquad \text{(A.36)}$$

where $\overline{d}(\mathscr{P}) = p_{\max} - p_{\min}$. We hereby suppress the dependence on $\boldsymbol{\theta}_t$ and $\hat{\boldsymbol{\xi}}_t$ for notational ease. Suppose that $\hat{p}_t - p_t^u > \iota_t$ (otherwise, $|p_t^u - \hat{p}_t| \le \iota_t$, and the result follows trivially). Naturally, $\hat{p}_t \in \mathscr{P}_d$ by the policy design. Let $\check{p} = \max\{p \in \mathscr{P}_d : p < \hat{p}_t\}$. Hence, we have $p_t^u < \check{p} < \hat{p}_t$. By assumption, $G^u(\cdot)$ has a non-vanishing first-order derivative, so

$$|G^u(\hat{p}_t) - G^u(\check{p})| \ge \underline{G'} |\hat{p}_t - \check{p}|, \tag{A.37}$$

where $\underline{G'} = \min\{(G^u)'(p) : p \in \mathscr{P}\} > 0$. Consider the following high probability event:

$$\mathscr{A}_d = \left\{ \max_{p_t \in \mathscr{P}_d} |G^u(p_t) - \hat{G}_t(p_t)| \le C_{10} \sqrt{\frac{\log M_t}{M_t}} \right\}.$$

On $\mathscr{A}_d$, we have

$$G^u(\hat{p}_t) + C_{10}\sqrt{\frac{\log M_t}{M_t}} \ge \hat{G}_t(\hat{p}_t) \ge \hat{G}_t(\check{p}) \ge G^u(\check{p}) - C_{10}\sqrt{\frac{\log M_t}{M_t}},$$

where the second inequality follows because $\hat{p}_t$ is the maximizer of $\hat{G}_t(\cdot)$ on the discretized grid $\mathscr{P}_d$. Rearranging terms, we deduce that

$$0 < G^u(\check{p}) - G^u(\hat{p}_t) \le 2C_{10}\sqrt{\frac{\log M_t}{M_t}}. \tag{A.38}$$

Combining (A.37) and (A.38), we obtain the following:

$$|\hat{p}_t - \check{p}| \le \frac{\sqrt{2C_{10}}}{\underline{G'}} \left(\frac{\log M_t}{M_t}\right)^{1/2}. \tag{A.39}$$

Similar argument applies to the case where $p_t^u - \hat{p}_t > \iota_t$ and the same conclusion follows on $\mathscr{A}_d$. Consequently,

$$|p_t^u - \hat{p}_t| \le |p_t^u - \check{p}| + |\check{p} - \hat{p}_t| \le \rho \left(\frac{\log M_t}{M_t}\right)^{1/2} + \frac{\sqrt{2C_{10}}}{\underline{G'}} \left(\frac{\log M_t}{M_t}\right)^{1/2} \le K_{11} \left(\frac{\log M_t}{M_t}\right)^{1/2},$$

where $K_{11} = \frac{\sqrt{2C_{10}}}{\underline{G'}} + \rho$. Because this inequality holds on $\mathscr{A}_d$, we thus conclude by (A.36) that

$$\mathbb{P}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}} \left\{ |p_t^u - \hat{p}_t| \le K_{11}\sqrt{\frac{\log M_t}{M_t}} \right\} \ge \mathbb{P}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}(\mathscr{A}_d) \ge 1 - \frac{K_{12}}{\sqrt{M_t \log M_t}},$$

where $K_{12} = 2C_{11}\overline{d}(\mathscr{P})/\rho$. Q.E.D.

Proof of Proposition 7. Given $\hat{p}_t, \hat{\boldsymbol{\xi}}_t$, we have by Lemma A.1 that

$$y_t^u(\hat{p}_t) = \inf\{y : \mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[(1-q_t)F_\varepsilon((1-q_t)y - g(\hat{\mathbf{X}}_t^\top \boldsymbol{\theta}_t))] \ge \mathscr{Q}(\hat{p}_t, \hat{\boldsymbol{\xi}}_t)\},$$

and

$$\tilde{y}_t^u(\hat{p}_t) = \inf\{y : \mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[(1-q_t)\hat{F}_\varepsilon^t((1-q_t)y - g(\hat{\mathbf{X}}_t^\top \boldsymbol{\theta}_t))] \ge \mathscr{Q}(\hat{p}_t, \hat{\boldsymbol{\xi}}_t)\},$$

where $\mathscr{Q}(\hat{p}_t, \hat{\boldsymbol{\xi}}_t) = [(b+\hat{p}_t)(1 - \mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[q_t]) - (w\mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[q_t] + c)]/(h - c + b + \hat{p}_t)$. Fix $\delta > 0$, and note that

$$\mathbb{P}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\{y_t^u(\hat{p}_t) - \tilde{y}_t^u(\hat{p}_t) > \delta\} \le \mathbb{P}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}} \left\{ \mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[(1-q_t)F_\varepsilon((1-q_t)(\tilde{y}_t^u(\hat{p}_t) + \delta) - g(\hat{\mathbf{X}}_t^\top \boldsymbol{\theta}_t))] < \mathscr{Q}(\hat{p}_t, \hat{\boldsymbol{\xi}}_t) \right\}$$

$$\leq \mathbb{P}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\Big\{\mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[(1-q_t)F_{\varepsilon}((1-q_t)(\tilde{y}^u_t(\hat{p}_t)+\delta)-g(\hat{\mathbf{X}}_t^{\top}\boldsymbol{\theta}_t))] \\ < \mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[(1-q_t)\hat{F}^t_{\varepsilon}((1-q_t)\tilde{y}^u_t(\hat{p}_t)-g(\hat{\mathbf{X}}_t^{\top}\boldsymbol{\theta}_t))]\Big\}$$

$$\leq \mathbb{P}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\Big\{\mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[(1-q_t)F_{\varepsilon}((1-q_t)\tilde{y}^u_t(\hat{p}_t)-g(\hat{\mathbf{X}}_t^{\top}\boldsymbol{\theta}_t))]+\underline{f}\delta \\ < \mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[(1-q_t)\hat{F}^t_{\varepsilon}((1-q_t)\tilde{y}^u_t(\hat{p}_t)-g(\hat{\mathbf{X}}_t^{\top}\boldsymbol{\theta}_t))]\Big\}$$

$$\leq \exp\left(-\frac{4M_t\underline{f}^2\delta^2}{1-\mathbb{E}_{q|\hat{\boldsymbol{\xi}}_t}[q_t]}\right) \leq \exp(-4M_t\underline{f}^2\delta^2),$$

for some positive constant $\underline{f}$. In a similar fashion, we arrive at the same inequality for the case where $y^u_t(\hat{p}_t)-\tilde{y}^u_t(\hat{p}_t)<-\delta$, and thereby conclude that

$$\mathbb{P}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\{|y^u_t(\hat{p}_t)-\tilde{y}^u_t(\hat{p}_t)|>\delta\} \leq 2\exp(-4M_t\underline{f}^2\delta^2).$$

Taking $\delta=\sqrt{\log M_t/M_t}/(2\underline{f})$ leads to the following:

$$\mathbb{P}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left\{|y^u_t(\hat{p}_t)-\tilde{y}^u_t(\hat{p}_t)|>\frac{1}{2\underline{f}}\sqrt{\frac{\log M_t}{M_t}}\right\} \leq \frac{2}{M_t}. \tag{A.40}$$

In addition, there exists a positive constant C_{12} such that $|\tilde{y}^u_t(\hat{p}_t)-\hat{y}^u_t| \leq C_{12}||\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t||$ by the reasoning of Proposition 5. We then deduce from by Proposition 2 that

$$\mathbb{P}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left\{|\tilde{y}^u_t(\hat{p}_t)-\hat{y}^u_t| \geq C_{12}\sqrt{\frac{K_3\log M_t}{M_t}}\right\} \leq \frac{K_4}{M_t}. \tag{A.41}$$

Combining (A.40) and (A.41) leads to the desired result with $K_{13}=1/(2\underline{f})+C_{12}\sqrt{K_3}$ and $K_{14}=2+K_4$. Q.E.D.

Proof of Theorem 2. We prove each part separately.

(a) Given $\boldsymbol{\theta}_t,\boldsymbol{\xi}$, we deduce from the triangle inequality that

$$\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[Q(p^u_t,y^u_t)-Q(\hat{p}_t,\hat{y}^u_t)\right] \leq \mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left|Q(p^u_t,y^u_t)-Q(\hat{p}_t,y^u_t(\hat{p}_t))\right| + \mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left|Q(\hat{p}_t,y^u_t(\hat{p}_t))-Q(\hat{p}_t,\hat{y}^u_t)\right|. \tag{A.42}$$

We first examine the first term on the right-hand side of (A.42). For the DDPO-N policy, the Lipschitz assumption on the function $Q(\cdot,\cdot)$ yields the following:

$$\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left|Q(p^u_t,y^u_t)-Q(\hat{p}_t,y^u_t(\hat{p}_t))\right| = \mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left|G^u(p^u_t)-G^u(\hat{p}_t)\right| \leq \overline{G}'\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}|p^u_t-\hat{p}_t|\,,$$

where $\overline{G}'=\max\{(G^u)'(p):p\in\mathscr{P}\}$. By Proposition 6, there exist positive constants C_{13} and C_{14} such that

$$\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{j=0}^{\mathscr{C}}\sum_{t=n\hat{\tau}_j^{+}+1}^{n\hat{\tau}_j^{-}}|Q(p^u_t,y^u_t)-Q(\hat{p}_t,y^u_t(\hat{p}_t))|\,\mathbb{I}\{t\notin\mathscr{X}\}\right]$$

$$\leq \overline{G}'C_{13}(\log T)^{1/2}\sum_{j=0}^{\mathscr{C}}\sum_{t=n\hat{\tau}_j^{+}+1}^{n\hat{\tau}_j^{-}}\left(\frac{1}{m(\lceil t/n\rceil-\hat{\tau}_j^{+})}\right)^{1/2}$$

$$
\begin{aligned}
&\leq (\mathscr{C}+1)\overline{G}' C_{13}(\log T)^{1/2}\frac{n}{m^{1/2}}\int_{\hat{\tau}_j^+}^{\hat{\tau}_j^-}\left(\frac{1}{\tau}\right)^{1/2}\mathrm{d}\tau\\
&\leq 2(\mathscr{C}+1)\overline{G}' C_{13}(\log T)^{1/2}\frac{n}{m^{1/2}}\left(\frac{T}{n}\right)^{1/2}\\
&\leq 2(\mathscr{C}+1)\overline{G}' C_{13}(\log T)^{1/2}\frac{(2\kappa T^{2/3})^{1/2}}{(\kappa T^{1/3})^{1/2}}T^{1/2}\\
&= C_{14}T^{2/3}(\log T)^{1/2},
\end{aligned}
$$

for all $T\geq 3$, where $C_{13}=K_{11}+K_{12}(p_{\max}-p_{\min})$ and $C_{14}=2\sqrt{2}(\mathscr{C}+1)\overline{G}' C_{13}$. Let us now examine the second term on the right-hand side of (A.42). Again by the Lipschitz assumption on $Q(\cdot,\cdot)$, we have

$$
\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left|Q(\hat{p}_t,y_t^u(\hat{p}_t))-Q(\hat{p}_t,\hat{y}_t^u)\right|\leq \mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left|H(y_t^u(\hat{p}_t))-H(\hat{y}_t^u)\right|\leq \overline{H}'\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}|y_t^u(\hat{p}_t)-\hat{y}_t^u|\,,
$$

where $\overline{H}'=\max\{H'(y):y\in\mathscr{Y}\}$. By Proposition 7, there exist positive constant C_{15} and C_{16} such that

$$
\begin{aligned}
&\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{j=0}^{\mathscr{C}}\sum_{t=n\hat{\tau}_j^++1}^{n\hat{\tau}_j^-}|Q(\hat{p}_t,y_t^u(\hat{p}_t))-Q(\hat{p}_t,y_t^u)|\,\mathbb{I}\{t\notin\mathscr{X}\}\right]\\
&\qquad\leq \overline{H}' C_{15}(\log T)^{1/2}\sum_{j=0}^{\mathscr{C}}\sum_{t=n\hat{\tau}_j^++1}^{n\hat{\tau}_j^-}\left(\frac{1}{m(\lceil t/n\rceil-\hat{\tau}_j^+)}\right)^{1/2}\\
&\qquad\leq 2(\mathscr{C}+1)\overline{H}' C_{15}(\log T)^{1/2}\frac{n}{m^{1/2}}\left(\frac{T}{n}\right)^{1/2}\\
&\qquad\leq C_{16}T^{2/3}(\log T)^{1/2},
\end{aligned}
$$

for all $T\geq 3$, where $C_{15}=K_{13}+K_{14}(y_{\max}-y_{\min})$ and $C_{16}=2\sqrt{2}(\mathscr{C}+1)\overline{H}' C_{15}$. Combining the above results with $K_{15}=C_{14}+C_{16}$ completes the proof.

(b) For the DDPO-E policy, the existence of the density function of the noise distribution implies that $Q(\cdot,\cdot)$ has vanishing first order partial derivatives at the local extrema. Hence, following the same decomposition argument as in part (a) with application of Propositions 2 and 3, we deduce that there exist positive constants C_{17} and C_{18} such that

$$
\begin{aligned}
\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left|Q(p_t^u,y_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi},\boldsymbol{\varphi})-Q(\hat{p}_t,\hat{y}_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi},\boldsymbol{\varphi})\right|&\leq \overline{G}''\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}|p_t^u-\hat{p}_t|^2+\overline{H}''\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}|y_t^u(\hat{p}_t)-\hat{y}_t^u|^2\\
&\leq C_{17}\left(\|\boldsymbol{\theta}_t-\hat{\boldsymbol{\theta}}_t\|^2+\|\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_t\|^2+\|\boldsymbol{\varphi}-\hat{\boldsymbol{\varphi}}_t\|^2\right)\\
&\leq C_{18}\frac{\log M_t}{M_t},
\end{aligned}
$$

where $\overline{G}'' = \max\{(G^u)''(p) : p \in \mathscr{P}\}$ and $\overline{H}'' = \max\{H''(y) : y \in \mathscr{Y}\}$. Therefore,

$$\begin{aligned}
&\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{j=0}^{\mathscr{C}}\sum_{t=n\hat{\tau}_j^{+}+1}^{n\hat{\tau}_j^{-}} |Q(p_t^u, y_t^u; \boldsymbol{\theta}_t, \boldsymbol{\xi}, \boldsymbol{\varphi}) - Q(\hat{p}_t, \hat{y}_t^u; \boldsymbol{\theta}_t, \boldsymbol{\xi}, \boldsymbol{\varphi})| \, \mathbb{I}\{t \notin \mathscr{X}\}\right] \\
&\quad \le C_{18}(\log T)^{1/2} \sum_{j=0}^{\mathscr{C}} \sum_{t=n\hat{\tau}_j^{+}+1}^{n\hat{\tau}_j^{-}} \frac{1}{m(\lceil t/n \rceil - \hat{\tau}_j^{+})} \\
&\quad \le (\mathscr{C}+1) C_{18}(\log T)^{1/2} \frac{n}{m} \int_{\hat{\tau}_j^{+}}^{\hat{\tau}_j^{-}} \frac{1}{\tau} \mathrm{d}\tau \\
&\quad \le (\mathscr{C}+1) C_{18}(\log T)^{1/2} \frac{2\kappa T^{1/2}}{\kappa \log T} \log T \\
&\quad = K_{16} T^{1/2} \log T
\end{aligned}$$

for all $T \ge 3$, where $K_{16} = 2(\mathscr{C}+1)C_{18}$. Q.E.D.

Proof of Proposition 8. We prove each part separately.

(a) Suppose that $\tau_{j+1}^* - \tau_j^* > 2$, where $\tau_j^* = \lfloor (t_j^* - 1)/n \rfloor$ (otherwise, the result is trivial for all $j = 1, 2, \ldots, \mathscr{C}$). Consider the set

$$A_j = \bigcup_{\tau = L(\tau_j^*)}^{\tau_j^*} \left\{ \boldsymbol{\theta}_s = \boldsymbol{\theta}_t \neq \boldsymbol{\theta}_{n(\tau_j^*+1)+1} \text{ for all } s, t \in \mathscr{X}_{1\tau} \cup \mathscr{X}_{2\tau} \right\}.$$

On the event A_j, there exists a cycle $k_0 = L(\tau_j^*), \ldots, \tau_j^*$ such that for all $t \in \mathscr{X}_{1k_0} \cup \mathscr{X}_{2k_0}$, we have $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{nk_0+1} \neq \boldsymbol{\theta}_{n(\tau_j^*+1)+1}$. Let $\mathbf{u}^* = \boldsymbol{\theta}_{n(\tau_j^*+1)+1}$ and $\mathbf{u}_0 = \boldsymbol{\theta}_{nk_0+1}$. Then, the probability of no detection in cycle $k^* = \tau_j^* + 1, \ldots, \tau_{j+1}^* - 2$ on A_j is given by

$$\begin{aligned}
\mathbb{P}^{\pi}_{\boldsymbol{\theta}}\{\chi_{k^*+1} = 0, A_j\} &= \mathbb{P}^{\pi}_{\boldsymbol{\theta}}\left\{ \sup_{i,\tau} \left\{ \left| \bar{D}_{ik^*} - \bar{D}_{i\tau} \right| : L(\tau^*) \le \tau < k^* \right\} \le \eta, A_j \right\} \\
&\le \mathbb{P}^{\pi}_{\boldsymbol{\theta}}\left\{ \left| \bar{D}_{ik^*} - \bar{D}_{ik_0} \right| \le \eta \text{ for } i = 1, 2, A_j \right\} \\
&= \mathbb{P}^{\pi}_{\boldsymbol{\theta}}\left\{ \frac{1}{m} \left| \sum_{t \in \mathscr{X}_{ik^*}} D_t - \sum_{s \in \mathscr{X}_{ik_0}} D_s \right| \le \eta \text{ for } i = 1, 2, A_j \right\} \\
&= \mathbb{P}^{\pi}_{\boldsymbol{\theta}}\left\{ \frac{1}{m} \left| \sum_{t \in \mathscr{X}_{ik^*}} \left(g(\tilde{\mathbf{X}}_i^{\mathsf{T}} \mathbf{u}^*) + \varepsilon_t \right) - \sum_{s \in \mathscr{X}_{ik_0}} \left(g(\tilde{\mathbf{X}}_i^{\mathsf{T}} \mathbf{u}_0) + \varepsilon_s \right) \right| \le \eta \text{ for } i = 1, 2, A_j \right\} \\
&= \mathbb{P}^{\pi}_{\boldsymbol{\theta}}\left\{ \frac{1}{m} \left| m \left[g(\tilde{\mathbf{X}}_i^{\mathsf{T}} \mathbf{u}^*) - g(\tilde{\mathbf{X}}_i^{\mathsf{T}} \mathbf{u}_0) \right] + \sum_{t \in \mathscr{X}_{ik^*}} \varepsilon_t - \sum_{s \in \mathscr{X}_{ik_0}} \varepsilon_s \right| \le \eta \text{ for } i = 1, 2, A_j \right\} \\
&\le \mathbb{P}^{\pi}_{\boldsymbol{\theta}}\left\{ \frac{1}{m} \left| \sum_{t \in \mathscr{X}_{ik^*}} \varepsilon_t - \sum_{s \in \mathscr{X}_{ik_0}} \varepsilon_s \right| \ge \left| g(\tilde{\mathbf{X}}_i^{\mathsf{T}} \mathbf{u}^*) - g(\tilde{\mathbf{X}}_i^{\mathsf{T}} \mathbf{u}_0) \right| - \eta \text{ for } i = 1, 2, A_j \right\} \\
&= \mathbb{P}^{\pi}_{\boldsymbol{\theta}}\left\{ |\bar{\varepsilon}_{ik^*} - \bar{\varepsilon}_{ik_0}| \ge \underline{g}' \left| \tilde{\mathbf{X}}_i^{\mathsf{T}} (\mathbf{u}^* - \mathbf{u}_0) \right| - \eta \text{ for } i = 1, 2, A_j \right\},
\end{aligned}$$

where $\underline{g}' = \min\{g'(\mathbf{X}^\intercal \boldsymbol{\vartheta}) : p \in \mathscr{P}, \boldsymbol{\vartheta} \in \Theta\}$. The rest of proof on A_j then follows from Lemma 4 of Keskin and Zeevi (2017).

On the event A_j^c, for each cycle $k \in [L(\tau_j^*), \tau_j^*)$, either there is a change-point in the experimentation period $\mathscr{X}_k = \mathscr{X}_{1k} \cup \mathscr{X}_{2k}$ or the demand parameter in $\mathscr{X}_k$ takes the same value as $\boldsymbol{\theta}_{t_j}^*$. Let $\mathscr{K}_{1j}$ and $\mathscr{K}_{2j}$ denote the above two sets of cycles, respectively. Given $p_t, \hat{\boldsymbol{\xi}}$, and $\mathcal{D}_t$ (as defined in the proof of Lemma 5) in period t in cycle $\tau \in (\tau_j^*, \hat{\tau}_j^+)$, we have

$$\begin{aligned}
&\tilde{G}_t(p_t; \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t, \boldsymbol{\varepsilon}) - W_t(\boldsymbol{\theta}_t; p_t, \hat{\boldsymbol{\xi}}_t, \mathcal{D}_t) \\
&= \frac{1}{2M_t} \sum_{k \in \mathscr{K}_{1j}} \sum_{s \in \mathscr{X}_k} \mathbb{E}_{q|\boldsymbol{\xi}_t} \Big\{ (h-c)[(1-q_t)y_t^u - g(\mathbf{X}_t^\intercal \boldsymbol{\theta}_t) - (D_s - g(\mathbf{X}_s^\intercal \boldsymbol{\theta}_t))]^+ \\
&\qquad + (b+p_t)[g(\mathbf{X}_t^\intercal \boldsymbol{\theta}_t) + (D_s - g(\mathbf{X}_s^\intercal \boldsymbol{\theta}_t)) - (1-q_t)y_t^u]^+ \\
&\qquad - (h-c)[(1-q_t)y_t^u - g(\mathbf{X}_t^\intercal \boldsymbol{\theta}_t) - (D_s - g(\mathbf{X}_s^\intercal \boldsymbol{\theta}_s))]^+ \\
&\qquad - (b+p_t)[g(\mathbf{X}_t^\intercal \boldsymbol{\theta}_t) + (D_s - g(\mathbf{X}_s^\intercal \boldsymbol{\theta}_s)) - (1-q_t)y_t^u]^+ \Big\}.
\end{aligned}$$

Thus, as the cardinality of $\mathscr{K}_{1j}$ is bounded by the total number of change-points, $\mathscr{C}$, we have

$$\begin{aligned}
|\tilde{G}_t(p_t; \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t, \boldsymbol{\varepsilon}) - \hat{G}_t(p_t; \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t, \mathbf{e})| &\le |W_t(\boldsymbol{\theta}_t; p_t, \hat{\boldsymbol{\xi}}_t, \mathcal{D}_t) - W_t(\hat{\boldsymbol{\theta}}_t; p_t, \hat{\boldsymbol{\xi}}_t, \mathcal{D}_t)| \\
&\quad + \frac{1}{2M_t} \sum_{k \in \mathscr{K}_{1j}} \sum_{s \in \mathscr{X}_k} [(h-c)\overline{g}' d_* + (b+p_t)\overline{g}' d_*] \\
&\le C_9 ||\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t|| + \frac{C_{19}}{M_t},
\end{aligned}$$

where $C_{19} = \mathscr{C} m(h - c + b + p_{\max})\overline{g}' d_*$. Therefore, there exists a constant C_{20} such that

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\xi}}^{\pi} \left\{ |\tilde{G}_t(p_t; \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t, \boldsymbol{\varepsilon}) - \hat{G}_t(p_t; \boldsymbol{\theta}_t, \hat{\boldsymbol{\xi}}_t, \mathbf{e})| \ge C_{20} \sqrt{\frac{\log M_t}{M_t}}, A_j^c \right\} \\
&\le \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\xi}}^{\pi} \left\{ C_9 ||\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t|| + \frac{C_{19}}{M_t} \ge C_{20} \sqrt{\frac{\log M_t}{M_t}}, A_j^c \right\} \\
&\le \mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\xi}}^{\pi} \left\{ ||\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t|| \ge K_3 \sqrt{\frac{\log M_t}{M_t}}, A_j^c \right\} \le \frac{K_4}{M_t},
\end{aligned}$$

where the last inequality follows from Proposition 2. Thus, the results of Propositions 6 and 7 continue to hold for $t \notin \mathscr{X}$ in cycle $\tau \in [L(\tau)_j^*, \hat{\tau}_j^+)$ for all $j = 0, \ldots, \mathscr{C}$ on A_j^c. Then, as in the proof of Theorem 2, similar decomposition arguments apply for both DDPO policies.

(b) To bound the detection error due to early false alarms, we simply note that the average demand difference for two cycles k and k' in the price experimentation periods (i.e., $\bar{D}_{ik} - \bar{D}_{ik'}$ for $i = 1, 2$) reduces to the difference between average demand noises (i.e., $\bar{\varepsilon}_{ik} - \bar{\varepsilon}_{ik'}$ for $i = 1, 2$) because there is no change-point between cycles $\tau_j^* + 1$ and τ_{j+1}^* and thus the mean demand $g(\mathbf{X}_t^\intercal \boldsymbol{\theta}_t)$ cancels out, irrespective of its functional form. Therefore, the desired result follows from the proof of Lemma 5 of Keskin and Zeevi (2017) in this case. Q.E.D.

Proof of Theorem 1. Decomposing the regret, we deduce that

$$\begin{aligned}
\Delta^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}(T) = \mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}&\left[\sum_{t=1}^{T}[Q(p_t^*,y_t^*;\boldsymbol{\theta}_t,\boldsymbol{\xi})-Q(\hat{p}_t,\hat{y}_t;\boldsymbol{\theta}_t,\boldsymbol{\xi})]\mathbb{I}\{t\in\mathscr{X}\}\right]\\
&+\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{t=1}^{T}[Q(p_t^*,y_t^*;\boldsymbol{\theta}_t,\boldsymbol{\xi})-Q(\hat{p}_t,\hat{y}_t;\boldsymbol{\theta}_t,\boldsymbol{\xi})]\mathbb{I}\{t\notin\mathscr{X}\}\right]\\
=\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}&\left[\sum_{t=1}^{T}[Q(p_t^*,y_t^*;\boldsymbol{\theta}_t,\boldsymbol{\xi})-Q(\hat{p}_t,\hat{y}_t;\boldsymbol{\theta}_t,\boldsymbol{\xi})]\mathbb{I}\{t\in\mathscr{X}\}\right]\\
&+\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{t=1}^{T}[Q(p_t^*,y_t^*;\boldsymbol{\theta}_t,\boldsymbol{\xi})-Q(\check{p}_t,\check{y}_t;\boldsymbol{\theta}_t,\boldsymbol{\xi})]\,\mathbb{I}\{t\notin\mathscr{X}\}\right]\\
&+\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{t=1}^{T}[Q(\check{p}_t,\check{y}_t;\boldsymbol{\theta}_t,\boldsymbol{\xi})-Q(p_t^u,y_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi})]\,\mathbb{I}\{t\notin\mathscr{X}\}\right]\\
&+\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{j=0}^{\mathscr{C}}\sum_{t=n\tau_j^*+1}^{n\hat{\tau}_j^+}[Q(p_t^u,y_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi})-Q(\hat{p}_t,\hat{y}_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi})]\,\mathbb{I}\{t\notin\mathscr{X}\}\right]\\
&+\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{j=0}^{\mathscr{C}}\sum_{t=n\hat{\tau}_j^++1}^{n\hat{\tau}_j^-}[Q(p_t^u,y_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi})-Q(\hat{p}_t,\hat{y}_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi})]\,\mathbb{I}\{t\notin\mathscr{X}\}\right]\\
&+\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{j=0}^{\mathscr{C}}\sum_{t=n\hat{\tau}_j^-+1}^{n\tau_{j+1}^*}[Q(p_t^u,y_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi})-Q(\hat{p}_t,\hat{y}_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi})]\,\mathbb{I}\{t\notin\mathscr{X}\}\right]\\
&+\mathbb{E}^{\pi}_{\boldsymbol{\theta},\boldsymbol{\xi}}\left[\sum_{t=1}^{T}[Q(\hat{p}_t,\hat{y}_t^u;\boldsymbol{\theta}_t,\boldsymbol{\xi})-Q(\hat{p}_t,\hat{y}_t;\boldsymbol{\theta}_t,\boldsymbol{\xi})]\,\mathbb{I}\{t\notin\mathscr{X}\}\right].
\end{aligned}$$

By assumption, $Q(p,y)$ is Lipschitz in both arguments in the compact rectangle $\mathscr{P}\times\mathscr{Y}$. Thus, $Q(p_t^*,y_t^*;\boldsymbol{\theta}_t,\boldsymbol{\xi})-Q(\hat{p}_t,\hat{y}_t;\boldsymbol{\theta}_t,\boldsymbol{\xi})$ is bounded above by a constant, and thus the first term on the right-hand side above is of order $2m\lceil T/n\rceil$, which is $O(T^{2/3})$ for DDPO-N and $O(T^{1/2}\log T)$ for DDPO-E. The second, third, and last terms are $O(1)$ by Proposition 1. Combining Theorem 2 and Proposition 8 for the remaining terms completes the proof for both parts of the theorem. Q.E.D.

Appendix B: Robustness Studies

B.1. Effect of the Number of Change-points

This subsection examines how the number of change-points, $\mathscr{C}$, affects the regret performance of the DDPO policy. To that end, we consider various problem instances where $\mathscr{C}$ scales with T logarithmically, sub-linearly (at different rates ranging from $T^{0.1}$ to $T^{0.9}$), and linearly. Specifically, we scale up the 3 change-points in 365 periods calibrated from the ginger data in store A such that $\mathscr{C}$ is in the order of $\log T$ (denoted as $\mathscr{C} \approx \log T$), or in the order of T^r (denoted as $\mathscr{C} \approx T^r$), where $r \in \{0.1, 0.2, \ldots, 1\}$. To maintain the 3 change-points in 365 periods while adding more change-points in longer time horizons, we repeat the 3 calibrated change-points cyclically by concatenating the 365 periods containing them until there are $\mathscr{C}$ change-points. For instance, if $T = 2000$ and $r = 0.5$, then there are $\mathscr{C} = 7$ change-points given by $\{61, 135, 210, 365, 426, 500, 575\}$ over $365 \times 2 = 730$ periods. We then scale them up in proportion to T, given by $\{61, 135, 210, 365, 426, 500, 575\} \times T/730$ rounded to the nearest integers.

Table B.1 displays the estimated growth rates of regret over $T \in \{2000, 4000, \ldots, 20000\}$ under both versions of the DDPO policy, as $\mathscr{C}$ increases in T at the rates shown in the first column (all other problem parameters are the same as in §4.2.1). When $\mathscr{C}$ scales with T logarithmically or sub-linearly at a rate less than or equal to $T^{0.6}$, the regret performance of our policy is robust to the number of change-points (the T-period regret is approximately in the order of $T^{2/3}$ for DDPO-N and $T^{1/2}$ for DDPO-E). As $\mathscr{C}$ scales up with T at faster rates, the regret performance slightly deteriorates. It is worth noting that the estimates in Table B.1 are based on real-life data, which do not necessarily correspond to a worst-case scenario. In particular, when the number of change-points scales linearly in T, it is possible to construct a worst-case scenario with linearly growing regret.

Table B.1 Effect of the Number of Change-points on the Growth Rate of Regret

Number of Change-points	T-period Regret of DDPO-N	T-period Regret of DDPO-E
$\mathscr{C} \approx \log T$	$O(T^{0.67}(\log T)^{1/2})$	$O(T^{0.50} \log T)$
$\mathscr{C} \approx T^{0.1}$	$O(T^{0.66}(\log T)^{1/2})$	$O(T^{0.52} \log T)$
$\mathscr{C} \approx T^{0.2}$	$O(T^{0.68}(\log T)^{1/2})$	$O(T^{0.51} \log T)$
$\mathscr{C} \approx T^{0.3}$	$O(T^{0.68}(\log T)^{1/2})$	$O(T^{0.52} \log T)$
$\mathscr{C} \approx T^{0.4}$	$O(T^{0.67}(\log T)^{1/2})$	$O(T^{0.52} \log T)$
$\mathscr{C} \approx T^{0.5}$	$O(T^{0.66}(\log T)^{1/2})$	$O(T^{0.51} \log T)$
$\mathscr{C} \approx T^{0.6}$	$O(T^{0.67}(\log T)^{1/2})$	$O(T^{0.52} \log T)$
$\mathscr{C} \approx T^{0.7}$	$O(T^{0.68}(\log T)^{1/2})$	$O(T^{0.53} \log T)$
$\mathscr{C} \approx T^{0.8}$	$O(T^{0.71}(\log T)^{1/2})$	$O(T^{0.53} \log T)$
$\mathscr{C} \approx T^{0.9}$	$O(T^{0.73}(\log T)^{1/2})$	$O(T^{0.55} \log T)$
$\mathscr{C} \approx T$	$O(T^{0.73}(\log T)^{1/2})$	$O(T^{0.56} \log T)$

B.2. Effect of the Beta Distribution Parameters

This subsection examines the effect of the beta distribution parameters on the growth rate of regret under the DDPO policy. For the first parameter, λ, we consider the values in $\{0.4, 1, 1.5\}$, which correspond to different shapes of the beta distribution. For each value of λ, we consider a set of values for the second parameter, ν, so that the mean of the beta distribution is higher than, roughly equal to, or lower than that of the beta distribution calibrated from our real-life data set.

Table B.2 shows the estimated growth rate of regret over $T \in \{2000, 4000, \ldots, 20000\}$ under both versions of the DDPO policy, based on the ginger data in store A. Except for the perishability parameter $\boldsymbol{\xi} = (\lambda, \nu)$, all other parameters are the same as in §4.2.1. The results on the table indicate that the regret performance of our policy is robust to the values of the beta distribution parameters.

Table B.2 Effect of the Beta Distribution Parameters on the Growth Rate of Regret

λ	ν	T-period Regret of DDPO-N	T-period Regret of DDPO-E
0.4	40	$O(T^{0.65}(\log T)^{1/2})$	$O(T^{0.50}\log T)$
	44	$O(T^{0.64}(\log T)^{1/2})$	$O(T^{0.50}\log T)$
	48	$O(T^{0.65}(\log T)^{1/2})$	$O(T^{0.50}\log T)$
1	105	$O(T^{0.65}(\log T)^{1/2})$	$O(T^{0.50}\log T)$
	110	$O(T^{0.65}(\log T)^{1/2})$	$O(T^{0.50}\log T)$
	115	$O(T^{0.65}(\log T)^{1/2})$	$O(T^{0.50}\log T)$
1.5	160	$O(T^{0.65}(\log T)^{1/2})$	$O(T^{0.50}\log T)$
	165	$O(T^{0.65}(\log T)^{1/2})$	$O(T^{0.50}\log T)$
	170	$O(T^{0.65}(\log T)^{1/2})$	$O(T^{0.50}\log T)$

B.3. Effect of the Cost Parameters

In this subsection, we consider various combinations of the cost parameters h, b, and w, and study their effect on the growth rate of regret and the annual profit, based on the ginger data in store A. As in §4.2.1, the average wholesale price for ginger is $c = 8.6215$, and the average profit margin is 2.1572. For the lost-sales penalty b, we consider three scenarios: 20%, 50%, and 100% of the average profit margin, divided by 365 to obtain a unit cost per day. This gives rise to three choices of b: 0.0012, 0.003, 0.0059. As mentioned in §4.2.1, according to the supermarket chain manager, the holding cost h and the disposal cost w jointly account for about 20% of the wholesale price c on an annual basis. Hence, for each value of b, we consider three scenarios for the pair (h, w): $(5\%, 15\%)$, $(10\%, 10\%)$, $(15\%, 5\%)$ of the wholesale price c, divided by 365 to obtain a unit cost per day. All other problem parameters remain the same as in §4.2.1.

Consistent with our theoretical results, the regret over $T \in \{2000, 4000, \ldots, 20000\}$ is $O(T^{0.66}(\log T)^{1/2})$ under DDPO-N and $O(T^{0.51}\log T)$ under DDPO-E for every combination of the cost parameters. The annual profits of DDPO-N and DDPO-E in all of the aforementioned settings are in Table B.3. The results indicate that the annual profit of the DDPO policy is also robust to the values of the cost parameters.

Table B.3 Effect of the Cost Parameters on Profits

b	h	w	Annual Profit of DDPO-N (RMB)	Annual Profit of DDPO-E (RMB)
0.0012	0.0012	0.0035	2692.0	1485.0
	0.0024	0.0024	2691.0	1484.3
	0.0035	0.0012	2690.1	1483.0
0.003	0.0012	0.0035	2691.5	1483.9
	0.0024	0.0024	2690.5	1483.2
	0.0035	0.0012	2689.7	1481.9
0.0059	0.0012	0.0035	2690.7	1482.1
	0.0024	0.0024	2689.8	1481.4
	0.0035	0.0012	2688.9	1480.0

B.4. Using the Data from All Periods

The DDPO policy uses the data from the first $2m$ periods of each cycle for learning purposes. Let us now consider a modification of the DDPO policy that uses the data from all periods to estimate the unknown parameters. Based on the ginger data in store A and the same policy parameters as in §4.2.1, the regret over $T \in \{2000, 4000, \ldots, 20000\}$ is $O(T^{0.67}(\log T)^{1/2})$ under the modified version of DDPO-N and $O(T^{0.51}\log T)$ under the modified version of DDPO-E—these are roughly the same as the growth rates of regret under the original versions of the DDPO policy (see §4.2.2). The intuition for these results is that the DDPO policy is specifically designed to achieve the best possible growth rate of regret using only the observations from the first $2m$ periods of each cycle. Based on the sample path realized in the ginger data in store A, the annual profits slightly increase by 1.5% under the modified version of DDPO-N, and by 0.8% under the modified version of DDPO-E.

B.5. Histograms of Regret Savings

The histograms in Figure 8 display the percentage regret savings of DDPO-N and DDPO-E relative to the supermarket's decisions across 33 product-store pairs, as described in §4.2.3.

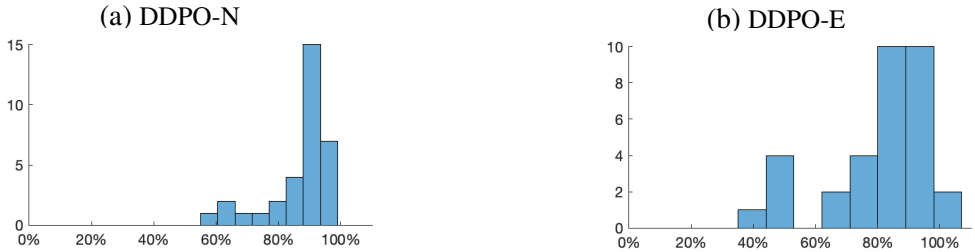

Figure 8 Histograms of Percentage Regret Savings