Dynamic Pricing and Learning with Discounting

Zhichao Feng

Department of Logistics and Maritime Studies, Faculty of Business, The Hong Kong Polytechnic University
zhi-chao.feng@polyu.edu.hk

Milind Dawande, Ganesh Janakiraman, Anyan Qi

Naveen Jindal School of Management, The University of Texas at Dallas
milind@utdalla.edu, ganesh@utdallas.edu, axq140430@utdallas.edu

Abstract

In many practical settings, learning algorithms can take a substantial amount of time to converge, thereby raising the need to understand the role of discounting in learning. We illustrate the impact of discounting on the performance of learning algorithms by examining two classic and representative dynamic-pricing and learning problems studied in Broder and Rusmevichientong (2012) [BR] and Keskin and Zeevi (2014) [KZ]. In both settings, a seller sells a product with unlimited inventory over T periods. The seller initially does not know the parameters of the general choice model in BR (resp., the linear demand curve in KZ). Given a discount factor ρ, the retailer's objective is to determine a pricing policy to maximize the expected discounted revenue over T periods. In both settings, we establish lower bounds on the regret under any policy and show limiting bounds of $\Omega(\sqrt{1/(1-\rho)})$ and $\Omega(\sqrt{T})$ when $T \to \infty$ and $\rho \to 1$, respectively. In the model of BR with discounting, we propose an asymptotically tight learning policy and show that the regret under our policy as well that under the MLE-CYCLE policy in BR is $\mathcal{O}(\sqrt{1/(1-\rho)})$ (resp., $\mathcal{O}(\sqrt{T})$) when $T \to \infty$ (resp., $\rho \to 1$). In the model of KZ with discounting, we present sufficient conditions for a learning policy to guarantee asymptotic optimality, and show that the regret under any policy satisfying these conditions is $\mathcal{O}(\log(1/(1-\rho))\sqrt{1/(1-\rho)})$ (resp., $\mathcal{O}(\log T\sqrt{T})$) when $T \to \infty$ (resp., $\rho \to 1$). We show that three different policies – namely, the two variants of the greedy Iterated-Least-Squares policy in KZ and a different policy that we propose – achieve this upper bound on the regret. We numerically examine the behavior of the regret under our policies as well as those in BR and KZ in the presence of discounting. We also analyze a setting in which the discount factor per period is a function of the number of decision periods in the planning horizon.

Keywords: *dynamic pricing*, *learning*, *discounting*, *regret minimization*

Funding: Zhichao Feng received support from the National Natural Science Foundation of China [Grant 72201256].

1 Introduction

In the classic revenue-management setting where a revenue-maximizing retailer sells a new product over a planning horizon, customers arrive sequentially and the retailer needs to dynamically adjust its retail price to learn the demand curve of the product, i.e., the relationship between the demand and the retail price; see, for example, Broder and Rusmevichientong (2012) [henceforth, BR] and Keskin and Zeevi (2014) [henceforth, KZ]. Depending on the industry, the arrival rate of consumers (via online and/or off-line channels) for the product(s) whose demand curve is to be learned, can be drastically different. For example, for a typical product sold (say) in a grocery store, the retailer can hope to access thousands of consumers each day. In contrast, the arrival rate of customers for expensive and niche products is typically much lower. For instance, only a handful of customers typically visit a Lamborghini dealership in a day; in 2019, Lamborghini sold a total of 4,296 units of the Super SUV Lamborghini Urus across 165 dealers in the world (which is about 0.08 units sold per dealership per day on average)[1]. When the consumer arrival rate is high, the required observations can be made quickly, thus enabling learning algorithms to converge fast. On the other hand, when the arrival rate is low, then learning can take a substantial amount, e.g., months, to converge, thus raising the need to understand the role of discounting in learning algorithms. In a similar vein, complex learning tasks can also require a significant amount of time; e.g., the learning of the correlated demand curves of a large number of substitutable products, again arguing for the need to incorporate discounting in learning. Events such as the recent significant interest-rate hikes by the U.S. Federal Reserve further raise the need to understand the impact of discounting in learning[2].

Intuitively, it is clear that discounting can fundamentally influence the tradeoff between exploration and exploitation that learning algorithms often exploit. Our goal is to illustrate the impact of discounting on the performance of learning algorithms. To this end, we examine two classic and representative settings that are analyzed in BR and KZ, and incorporate a discount factor in the learning algorithms for these settings. In particular, we consider two asymptotic regimes without assuming any specific relationship between the length of the planning horizon and the discount factor – first, consistent with the analysis in BR and KZ, where the length of the planning horizon approaches infinity, and the second, where the discount factor approaches 1. For each setting, we first use the classical notion of regret to analyze the performance of the algorithms proposed in these two papers with respect to the discount factor. This analysis requires the development of a lower bound of the regret under *any* policy and an upper bound on the regret under the specific algorithms proposed in these papers. Then, we propose new algorithms that explicitly incorporate the discount factor and obtain upper bounds on the regret under these algorithms. We also numerically examine the behavior

[1] https://www.best-selling-cars.com/brands/2019-global-lamborghini-sales-worldwide/

[2] https://www.bloomberg.com/news/articles/2022-07-13/fed-could-weigh-historic-100-basis-point-hike-after-cpi-scorcher#xj4y7vzkg

of the regret under these algorithms.

It is important to note that while the significance of incorporating discounting naturally increases as the duration of the planning horizon increases, discounting can be explicitly incorporated in *any* learning setting, without consideration for the length of the planning horizon. For instance, for any given planning horizon that has been divided into decision periods, one can define the precise change in the value of money with time by examining a specific relationship between the discount rate per period and the discount rate over the planning horizon. Of course, the significance of incorporating discounting will likely vary depending on the specifics of the business context under consideration. To this end, we also examine an alternate setting where a given planning horizon (say, a quarter) is split into smaller decision periods. This naturally results in the discount factor per period being an increasing function of the number of periods. In particular, as the number of periods tends to infinity, the discount factor per period tends to 1. Here, we consider another asymptotic regime where the number of periods approaches infinity, keeping the demand volume per period the same, and, as before, derive a lower bound on the regret under any policy and an upper bound on the regret under the algorithms in BR and KZ as well as our new algorithms.

1.1 Summary of Contributions

We incorporate discounting in two models: the model in BR and the one in KZ. Both papers study the dynamic pricing of products with unlimited inventory, where the seller needs to learn the unknown parameters of a demand function. In both settings, we first establish a lower bound on the regret under any pricing policy – this bound is a function of both the time horizon T and the discount factor ρ – and then obtain asymptotically optimal policies. We also analyze a setting in which the discount factor per period is a specific function of the number of decision periods in the planning horizon.

In Section 2, we incorporate discounting in the model analyzed in BR, where the probability that a customer purchases a product at a given price is characterized by certain parameters that are unknown and should be learnt. The authors show that the regret under any pricing policy is $\Omega(\sqrt{T})$, and propose a pricing policy, referred to as MLE-CYCLE, that achieves a regret of $\mathcal{O}(\sqrt{T})$. In the presence of discounting, we establish (in Theorem 1) a lower bound of $\Omega\left(\sqrt{\frac{1}{1-\rho}}\right)$ (resp., $\Omega(\sqrt{T})$) on the regret under *any* pricing policy, when $T \to \infty$ (resp., $\rho \to 1$). Then, we propose a pricing policy and show (in Theorem 2) that it achieves a regret of $\mathcal{O}\left(\sqrt{\frac{1}{1-\rho}}\right)$ (resp., $\mathcal{O}(\sqrt{T})$) when $T \to \infty$ (resp., $\rho \to 1$). In addition, we show (in Theorem 3) that under discounting, the MLE-CYCLE policy in BR also achieves a regret of $\mathcal{O}\left(\sqrt{\frac{1}{1-\rho}}\right)$ when $T \to \infty$. In Section 2.3, we numerically examine the behavior of the regret under our policy as well as the MLE-CYCLE policy in the presence of discounting.

In Section 3, we incorporate discounting in the single-product setting of KZ. The authors consider a linear demand model, where the demand in each time period consists of two parts: a deterministic part, which is a linear function (with unknown parameters) of the price, and a random demand shock. They show that the regret under *any* pricing policy is $\Omega(\sqrt{T})$. The authors then consider the well-known greedy Iterated Least Squares (ILS) policy (Anderson and Taylor 1976) and show that, under certain conditions, a variant of this policy achieves a regret of $\mathcal{O}(\log T\sqrt{T})$. In the presence of discounting, we establish (in Theorem 4) a lower bound of $\Omega\left(\sqrt{\frac{1}{1-\rho}}\right)$ (resp., $\Omega(\sqrt{T})$) on the regret under *any* policy, when $T \to \infty$ (resp., $\rho \to 1$). We then modify the conditions in KZ in the presence of discounting and show (in Theorem 5) that if a policy satisfies our conditions, then it achieves a regret of $\mathcal{O}\left(\log\left(\frac{1}{1-\rho}\right)\sqrt{\frac{1}{1-\rho}}\right)$ (resp., $\mathcal{O}(\log T\sqrt{T})$), when $T \to \infty$ (resp., $\rho \to 1$). Next, we show that three different policies – namely, the two variants of the greedy ILS policy in KZ and a different policy that we propose – achieve this upper bound on the regret. Finally, in Section 3.3, we numerically examine the performance of the above policies.

Section 4 assumes a specific relationship between ρ and T. We show that for the models in BR and KZ, the regret under any policy is $\Omega(\sqrt{T})$ (Propositions A.1 and A.4). For the model in BR, we show that the regret under our policy as well that under the MLE-CYCLE policy in BR is $\mathcal{O}(\sqrt{T})$ (Propositions A.2 and A.3). For the model of KZ, we show that the regret is $\mathcal{O}(\log T\sqrt{T})$ under three policies – namely, the two variants of the greedy Iterated-Least-Squares policy in KZ and a different policy that we propose (Propositions A.5 and A.6).

1.2 Technical Highlights

We first briefly explain the highlights of our technical analysis when we incorporate discounting in the model of BR. To establish a lower bound on the regret under *any* policy in terms of the time horizon T and discount factor ρ, similar to BR, we apply the Kullback-Leibler (KL) divergence as a measure of the difference between two distributions. BR establish two lower bounds (Lemmas 3.3 and 3.4 in that paper) on the regret that are functions of the KL divergence and T. Based on these two lower bounds, they choose two parameter values that depend on T, and show that the regret is $\Omega(\sqrt{T})$ for at least one of these two parameter values. In our analysis, however, we need to take into account the discounting effect and establish two *new* lower bounds (Lemmas A.2 and A.3) on the regret that are functions of the KL divergence, T, and ρ. Using these two bounds, we choose three parameter values, which depend on both T and ρ, and show that for at least one of these three values, the regret is $\Omega\left(\sqrt{\frac{1}{1-\rho}}\right)$ (resp., $\Omega(\sqrt{T})$) when $T \to \infty$ (resp., $\rho = 1$).

The incorporation of the discount factor alters the analysis in a non-trivial manner. For instance, BR prove a lower bound on the regret in Lemma 3.3 of their paper by (i) first establishing a lower bound on the regret in each time period using the conditional KL divergence in that time period and (ii) then applying the Chain Rule to show a lower bound on the cumulative regret using the total KL divergence. In our analysis, however, we cannot apply the Chain Rule directly as we must consider the cumulative *discounted* regret, which is bounded from below by a constant times the cumulative discounted conditional KL divergence; we note that the Chain Rule can only be applied to the summation of the conditional KL divergence. Therefore, we need to isolate the discount factor from the cumulative discounted conditional KL divergence – to achieve this, it becomes necessary to decompose and reorganize the cumulative discounted conditional KL divergence.

As mentioned earlier, in our analysis of the setting in BR, Theorems 2 and 3 establish, respectively, asymptotic bounds on the regret under our proposed policy and under the MLE-CYCLE policy in BR. The structure of our policy is different from that of the policy in BR. Specifically, MLE-CYCLE operates in cycles, with each cycle consisting of an exploration phase and an exploitation phase. Our policy consists of one exploration phase at the beginning of the time horizon and one exploitation phase for the remainder of the time horizon. The key to guarantee the asymptotic upper bounds on the regret under our policy is to choose an appropriate length of the exploration phase to balance the exploration-exploitation tradeoff.

In our analysis of the setting in KZ, to obtain a lower bound on the regret under *any* policy in terms of both the time horizon T and discount factor ρ, we generalize Lemma 1 in KZ by incorporating ρ (Lemma A.7). To establish the upper bound on the regret under discounting in Theorem 5, we use a decomposition approach that is significantly different from the technique used in KZ: For a constant N, we partition the cumulative regret into two parts, namely the regret in the first N periods and that in the remaining $T - N$ periods, and first establish an upper bound on the cumulative regret that is a function of T, ρ, and N. Our eventual upper bound is then obtained by making an appropriate choice of N. Finally, the upper bound on the regret in Theorem 6 also requires non-trivial technical work. In particular, Lemma A.9 establishes an upper bound on the difference between the optimal price under the true parameters of the demand function and the "greedy" price (based on our estimates) offered in the exploitation phase of our policy.

1.3 Literature Review

Our work is related to the stream of literature on demand learning with discounting. Studies in this stream examine dynamic pricing policies in Bayesian learning settings with discounted rewards; unlike our work, the focus in these papers is not on studying the impact of discounting in their policies or on analyzing the regret under their policies with respect to the discount factor. We briefly summarize a few representative papers in this literature. Chen and Wang (1999) study the dynamic pricing of a single asset over an infinite horizon with discounted rewards. The willingness-to-pay is drawn from one of two possible distributions and is learnt via Bayesian updating. The authors characterize an optimal pricing policy that incorporates updated beliefs in each period. Zhang and Chen (2006) investigate joint pricing and inventory control with Bayesian learning of a component of the demand that does not depend on the selling price. The authors solve a Bayesian dynamic program and characterize an optimal policy to maximize the finite-horizon expected discounted profit. Araman and Caldentey (2009) and Farias and Van Roy (2010) consider dynamic pricing problems faced by a retailer with finite inventory, and aim to maximize expected discounted revenue over an infinite time horizon. In both problems, the willingness-to-pay distribution is known while the customer arrival rate is unknown. Both papers formulate Bayesian dynamic programs, propose pricing heuristics, and analyze their performance. Mason and Välimäki (2011) focus on revenue maximization in a posted-price, infinite-horizon setting with discounted rewards. They assume a commonly-known willingness-to-pay distribution and an unknown arrival-rate of customers, which can be either high or low and is learnt in a Bayesian fashion. The authors study the structural properties of the optimal price. Kwon et al. (2012) consider a firm seeking to maximize its expected discounted profit over an infinite time horizon using markdown pricing. They model the cumulative demand as a Brownian motion with an unknown drift that is either high or low and is learnt via Bayesian updating. The authors characterize the optimal initial price, markdown price, and markdown time.

Our work also contributes to the recent fast-growing literature on learning algorithms, e.g., Broder (2011), Besbes and Zeevi (2015), Qi et al. (2017), Baardman et al. (2019), Chen et al. (2020), Jagabathula et al. (2020), Keskin et al. (2020), Lei et al. (2023), Mintz et al. (2020), Zhang et al. (2020), Besbes et al. (2021), Chen et al. (2021), Lyu et al. (2021), Zhang et al. (2021), Keskin and Li (2023), and Feng et al. (2023). The papers in this domain propose and analyze learning algorithms under a variety of business environments (without explicitly incorporating discounting), and examine the behavior of the regret. For an excellent review of earlier work in this domain, we refer the reader to den Boer (2015).

2 Incorporating Discounting in the Model in BR (2012)

We briefly introduce the model in BR; to the extent possible, we retain the same notation. Consider a retailer selling a product with infinite inventory over a time horizon of T periods. In period $t \geq 1$, the retailer determines the price p_t and a customer decides whether to purchase the product or not at that price. Let $d(p; \boldsymbol{z})$ denote the probability that the customer will purchase a product at price p, where $p \in \mathcal{P} = [p_{\min}, p_{\max}] \subseteq \mathbb{R}_+$, $\boldsymbol{z} \in \mathcal{Z} \subseteq \mathbb{R}^n$ is a vector of unknown parameters, and $\mathcal{Z}$ is a compact and convex parameter set. We assume that $d(p; \boldsymbol{z})$ is nonincreasing in p for all $\boldsymbol{z} \in \mathcal{Z}$. Then, the single-period expected revenue $r(p; \boldsymbol{z})$ under price p is $r(p; \boldsymbol{z}) = d(p; \boldsymbol{z})p$.

We now incorporate discounting. Let $\rho \in [0, 1)$ denote the discount factor and let $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$. Then, the expected discounted revenue over T periods is:

$$R(\boldsymbol{z}, \mathcal{C}, T, \rho) = \sum_{t=1}^{T} \rho^{t-1} r(p_t; \boldsymbol{z}).$$

When $\boldsymbol{z}$ is known, we assume that $r(\cdot; \boldsymbol{z})$ has a unique maximizer in $\mathcal{P}$ for each $\boldsymbol{z} \in \mathcal{Z}$, and denote it by $p^*(\boldsymbol{z})$. We make the following two assumptions (Assumptions 1 and 2 below), reproduced verbatim here from BR, on the problem class $\mathcal{C}$.

Assumption 1 *There exists positive constants $d_{\min}$, $d_{\max}$, L, and c_r, such that*

(a) $0 < d_{\min} \leq d(p; \boldsymbol{z}) \leq d_{\max} < 1$ *for all $p \in \mathcal{P}$ and $\boldsymbol{z} \in \mathcal{Z}$.*

(b) The revenue function $p \mapsto r(p; \boldsymbol{z})$ has a unique maximizer $p^*(\boldsymbol{z}) \in \mathcal{P}$.

(c) The function $\boldsymbol{z} \mapsto p^*(\boldsymbol{z})$ is L-Lipschitz, that is, $|p^*(\boldsymbol{z}) - p^*(\bar{\boldsymbol{z}})| \leq L\|\boldsymbol{z} - \bar{\boldsymbol{z}}\|$ for all $\boldsymbol{z}, \bar{\boldsymbol{z}} \in \mathcal{Z}$.

(d) The revenue function $p \mapsto r(p; \boldsymbol{z})$ is twice differentiable with $\sup_{p \in \mathcal{P}, \boldsymbol{z} \in \mathcal{Z}} |r''(p; \boldsymbol{z})| \leq c_r$.

For $t \geq 1$, let $y_t = 1$ if a customer purchases a product in period t, and $y_t = 0$ otherwise. Let $\boldsymbol{y_t} = (y_1, y_2, \cdots, y_t) \in \{0, 1\}^t$ denote the purchasing history in the first t periods. When $\boldsymbol{z}$ is unknown, we define a pricing policy $\psi = (\psi_1, \psi_2, \cdots)$ as a sequence of functions, where $\psi_t : \{0, 1\}^{t-1} \to \mathcal{P}$ sets the price in period t based on $\boldsymbol{y_{t-1}}$, i.e., the purchasing history of the first $t - 1$ periods. For any policy ψ and $\boldsymbol{z} \in \mathcal{Z}$, let $\boldsymbol{Y_t^{\psi, z}} = (Y_1, Y_2, \cdots, Y_t)$ denote the random outcome of the first t periods when policy ψ is used and the vector of the underlying true parameters is $\boldsymbol{z}$. Then, the probability distribution of $\boldsymbol{Y}_{\boldsymbol{t}}^{\boldsymbol{\psi},\boldsymbol{z}}$ is given by

$$Q_t^{\psi,\boldsymbol{z}}(\boldsymbol{y_t}) = \prod_{l=1}^{t} d(\psi_l(\boldsymbol{y_{l-1}});\boldsymbol{z})^{y_l}(1-d(\psi_l(\boldsymbol{y_{l-1}});\boldsymbol{z}))^{1-y_l} \text{ for all } \boldsymbol{y_t}=(y_1,y_2,\cdots,y_t)\in\{0,1\}^t.$$

We also define the distribution of customer responses to a sequence of fixed prices $\boldsymbol{p}=(p_1,\cdots,p_k)\in\mathcal{P}^k$ for some $k\in\mathbb{N}$:

$$Q^{\boldsymbol{p},\boldsymbol{z}}(\boldsymbol{y}) = \prod_{l=1}^{k} d(p_l;\boldsymbol{z})^{y_l}(1-d(p_l;\boldsymbol{z}))^{1-y_l},$$

where $\boldsymbol{y}\in\{0,1\}^k$. We make the following statistical assumption:

Assumption 2 *There exists a vector of exploration prices $\bar{\boldsymbol{p}}\in\mathcal{P}^k$ for some $k\in\mathbb{N}$ such that the family of distributions $\{Q^{\bar{\boldsymbol{p}},\boldsymbol{z}}:\boldsymbol{z}\in\mathcal{Z}\}$ is identifiable; that is, $Q^{\bar{\boldsymbol{p}},\boldsymbol{z}}(\cdot)\neq Q^{\bar{\boldsymbol{p}},\bar{\boldsymbol{z}}}(\cdot)$ whenever $\boldsymbol{z}\neq\bar{\boldsymbol{z}}$. Moreover, there exists a constant $c_f>0$ depending only on the problem class $\mathcal{C}$ and $\bar{\boldsymbol{p}}$ such that the smallest eigenvalue of matrix $\boldsymbol{I}(\bar{\boldsymbol{p}},\boldsymbol{z})$, denoted by $\lambda_{\min}\{\boldsymbol{I}(\bar{\boldsymbol{p}},\boldsymbol{z})\}$, satisfies $\lambda_{\min}\{\boldsymbol{I}(\bar{\boldsymbol{p}},\boldsymbol{z})\}\geq c_f$ for all $\boldsymbol{z}\in\mathcal{Z}$, where $\boldsymbol{I}(\bar{\boldsymbol{p}},\boldsymbol{z})$ denotes the Fisher information matrix that is defined as follows:*

$$[\boldsymbol{I}(\bar{\boldsymbol{p}},\boldsymbol{z})]_{i,j} = \mathbb{E}\left[-\frac{\partial^2}{\partial z_i\partial z_j}\log Q^{\bar{\boldsymbol{p}},\boldsymbol{z}}(\boldsymbol{Y})\right] = \sum_{k=1}^{n}\frac{\{(\partial/\partial z_i)d(\bar{p}_k,\boldsymbol{z})\}\times\{(\partial/\partial z_j)d(\bar{p}_k,\boldsymbol{z})\}}{d(\bar{p}_k,\boldsymbol{z})(1-d(\bar{p}_k,\boldsymbol{z}))}.$$

We refer the reader to Section 2 of BR for three common parametric demand models, namely the logit, linear, and exponential demand models, that satisfy Assumptions 1 and 2. For a problem class $\mathcal{C}$ satisfying Assumptions 1 and 2, we define the regret under an arbitrary policy:

Regret: Let P_t^{ψ} denote the random price in period t under policy ψ. For a problem class $\mathcal{C}$, a vector of parameter $\boldsymbol{z}\in\mathcal{Z}$, and a discount factor $\rho\in[0,1)$, the cumulative regret over T periods under ψ is:

$$\text{Regret}(\boldsymbol{z},\mathcal{C},T,\rho,\psi) = \sum_{t=1}^{T}\rho^{t-1}\mathbb{E}[r(p^*(\boldsymbol{z});\boldsymbol{z})-r(P_t^{\psi};\boldsymbol{z})].$$

2.1 Lower Bound on the Regret

Theorem 1 below establishes a lower bound on the regret under any pricing policy.

Theorem 1 *Define a problem class $\mathcal{C}_{LB}=(\mathcal{P},\mathcal{Z},d)$ by letting $\mathcal{P}=[3/4,5/4]$, $\mathcal{Z}=[1/3,1]$, and $d(p;z)=1/2+z-zp$. Then, for any policy ψ, there exists $\hat{\rho}\in[0,1)$, $\hat{T}\in\mathbb{N}$, a parameter $z\in\mathcal{Z}$, and a constant $K_0>0$ independent of ρ and T, such that for any $\rho\geq\hat{\rho}$ and $T\geq\hat{T}$, we have*

$$\textit{Regret}(z,\mathcal{C}_{LB},T,\rho,\psi)\geq K_0\left(\sqrt{\frac{\rho}{1-\rho}}(1-\rho^{T-1})+\sqrt{\frac{\rho^T(1-\rho^{T-1})}{1-\rho}}\right).$$

Further,

$$\lim_{\rho\to 1} \mathit{Regret}(z, \mathcal{C}_{LB}, T, \rho, \psi) = \Omega\left(\sqrt{T}\right) \quad \mathit{and} \quad \lim_{T\to\infty} \mathit{Regret}(z, \mathcal{C}_{LB}, T, \rho, \psi) = \Omega\left(\sqrt{\tfrac{1}{1-\rho}}\right).$$

2.2 An Asymptotically Tight Learning Algorithm

In this section, we propose a pricing policy and establish matching upper bounds on the regret under this policy. That is, as $T \to \infty$ (resp., $\rho \to 1$), our policy offers an upper bound of $\mathcal{O}\left(\sqrt{\frac{1}{1-\rho}}\right)$ (resp., $\mathcal{O}(\sqrt{T})$) on the regret. In addition, we show that the regret under the MLE-CYCLE policy in BR is also $\mathcal{O}\left(\sqrt{\frac{1}{1-\rho}}\right)$ when $T \to \infty$. We also numerically examine the regret under these two policies in Section 2.3.

Our pricing policy consists of only two phases: a single exploration phase at the start of the time horizon followed by a single exploitation phase. Recall from Assumption 2 that the vector of exploration prices is $\bar{\boldsymbol{p}} \in \mathcal{P}^k$. We first define the length of the exploration and exploitation phases. Let $\tau = \left[\sqrt{\frac{1-\rho^T}{1-\rho}}\right]$, where $[x]$ denotes the integer nearest to x. Then, the length of the exploration phase (resp., exploitation phase) is $k\tau$ (resp., $T - k\tau$). We now formally describe our policy below.

Inputs: A problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ and exploration prices $\bar{\boldsymbol{p}} = (\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_k) \in \mathcal{P}^k$.
Description: Divide the T periods into two phases: one for exploration and the other for exploitation.

- **Exploration Phase:** In time period $t = 1, 2, \cdots, k\tau$:

  Offer the product at exploration prices $\bar{\boldsymbol{p}} = (\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_k)$ sequentially. For $s = 1, \cdots, \tau$, let $\boldsymbol{Y}(s) = (Y_{(s-1)k+1}, \cdots Y_{sk})$ denote the outcomes in periods $(s-1)k+1, \cdots, sk$, when the prices in $\bar{\boldsymbol{p}}$ are offered for the s^{th} time. Let $\boldsymbol{Z}(\tau)$ denote the maximum likelihood estimate (MLE) based on the observed customer responses during the exploration phase; that is,

  $$\boldsymbol{Z}(\tau) = \arg\max_{\boldsymbol{z}\in\mathcal{Z}} \prod_{s=1}^{\tau} Q^{\bar{\boldsymbol{p}},\boldsymbol{z}}(\boldsymbol{Y}(s)).$$

- **Exploitation Phase:** In time period $t = k\tau + 1, k\tau + 2, \cdots, T$:

  Offer price $p^*(\boldsymbol{Z}(\tau))$ based on the estimate $\boldsymbol{Z}(\tau)$.

Let $\hat{\psi}$ denote our pricing policy above. Theorem 2 establishes an upper bound on the regret under $\hat{\psi}$ as well as limiting values of this bound as $T \to \infty$ and as $\rho \to 1$.

Theorem 2 *For any problem class $\mathcal{C}$ satisfying Assumptions 1 and 2 with corresponding exploration prices $\bar{\boldsymbol{p}} \in \mathcal{P}^k$, there exist constants $K_1, K_2 > 0$ independent of ρ and T, such that for all $\boldsymbol{z} \in \mathcal{Z}$,*

$T \in \mathbb{N}$ and $\rho \in [0,1)$, the policy $\hat{\psi}$ satisfies

$$\textit{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho, \hat{\psi}) \leq K_1 \frac{1-\rho^{k\tau}}{1-\rho} + K_2 \frac{\rho^{k\tau} - \rho^T}{(1-\rho)\tau},$$

where $\tau = \left\lceil \sqrt{\frac{1-\rho^T}{1-\rho}} \right\rceil$. Further,

$\lim_{\rho \to 1} \textit{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho, \hat{\psi}) = \mathcal{O}\left(\sqrt{T}\right)$ *and* $\lim_{T \to \infty} \textit{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho, \hat{\psi}) = \mathcal{O}\left(\sqrt{\frac{1}{1-\rho}}\right)$.

Theorems 1 and 2 together indicate that

$$\lim_{\rho \to 1} \text{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho, \hat{\psi}) = \Theta\left(\sqrt{T}\right) \quad \text{and} \quad \lim_{T \to \infty} \text{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho, \hat{\psi}) = \Theta\left(\sqrt{\frac{1}{1-\rho}}\right).$$

Theorem 3 establishes an upper bound on the regret under the MLE-CYCLE policy in BR when discounting is taken into consideration. We also show that the regret is $\mathcal{O}\left(\sqrt{\frac{1}{1-\rho}}\right)$ when $T \to \infty$ under that policy. We first briefly summarize the MLE-CYCLE policy.

The MLE-CYCLE policy operates in cycles, with each cycle consisting of an exploration phase and an exploitation phase. The length of the exploration phase is fixed, say k periods, while the length of the exploitation phase increases linearly in the number of cycles; specifically, the length of the exploitation phase of cycle h is h periods. During a cycle's exploration phase, the product is offered at exploration prices $\bar{\boldsymbol{p}} \in \mathcal{P}^k$ sequentially, and then the maximum-likelihood estimate of the underlying demand-curve parameters is computed based on the observed customer responses. During the cycle's exploitation phase, the "optimal" price based on the current estimates of the parameters is offered.

Theorem 3 *Let $\check{\psi}$ denote the MLE-CYCLE policy in BR. In the presence of discounting, for any problem class $\mathcal{C}$ satisfying Assumptions 1 and 2 with corresponding exploration prices $\bar{\boldsymbol{p}} \in \mathcal{P}^k$, there exists a constant K_3, independent of ρ and T, such that for all $\boldsymbol{z} \in \mathcal{Z}$, $T \geq 2$, and $\rho \in [0,1)$, the policy $\check{\psi}$ satisfies*

$$\textit{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho, \check{\psi}) \leq K_3 \left(1 + \sum_{h=1}^{\lfloor \sqrt{2T} \rfloor} \rho^{h^2/2}\right).$$

Further,

$\lim_{\rho \to 1} \textit{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho, \check{\psi}) = \mathcal{O}\left(\sqrt{T}\right)$ *and* $\lim_{T \to \infty} \textit{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho, \check{\psi}) = \mathcal{O}\left(\sqrt{\frac{1}{1-\rho}}\right)$.

2.3 Numerical Experience

In this section, we numerically examine the behavior of the regret under our policy and under the MLE-CYCLE policy in BR. First, we consider a linear demand model in the setting of BR. We let the demand distribution be $d(p;\boldsymbol{z}) = z_1 - z_2 p$, where $p \in \mathcal{P} = [0.75, 1.83]$ and $\boldsymbol{z} \in \mathcal{Z} = [1.1, 1.3] \times [0.4, 0.6]$. We set the exploration prices to be $\bar{\boldsymbol{p}} = (0.8, 1.8)$, and the total number of periods to be $T = 40000$. We consider the following nine scenarios of the true demand parameters:

$$(z_1, z_2) \in \Big\{(1.15, 0.45), (1.15, 0.5), (1.15, 0.55), (1.2, 0.45), (1.2, 0.5), (1.2, 0.55), (1.25, 0.45),\\ (1.25, 0.5), (1.25, 0.55)\Big\}.$$

We also vary $\log_{10}\left(\frac{1}{1-\rho}\right)$ from 1 to 6, in increments of 0.1; correspondingly, the discount factor ρ ranges from 0.9 to 0.999999. For each combination of (z_1, z_2) and ρ, we compute the average regret under our policy and that under MLE-CYCLE in BR over 100 simulations. Let Regret_1 (resp., Regret_2) denote the average regret under MLE-CYCLE (resp., our policy).

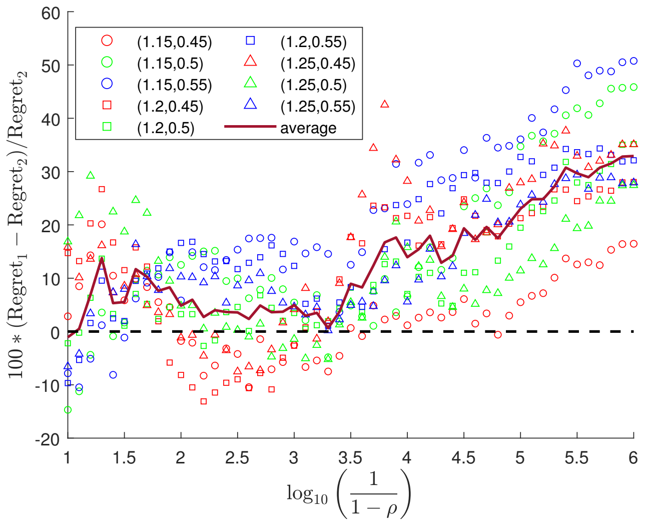
Next, we consider a logit demand model in the setting of BR. Here, we let the demand distribution be $d(p;\boldsymbol{z}) = \frac{e^{-z_1 p - z_2}}{1+e^{-z_1 p - z_2}}$ for $p \in \mathcal{P} = [0.5, 8]$ and $\boldsymbol{z} \in \mathcal{Z} = [0.2, 2] \times [-1, 1]$. We set the exploration prices to be $\bar{\boldsymbol{p}} = (0.5, 4.25)$ and the total number of periods to be $T = 40000$. We consider the following nine scenarios of the true demand parameters:

$$(z_1, z_2) \in \{(1.2, -1), (1.2, -0.5), (1.2, 0), (1.3, -1), (1.3, -0.5), (1.3, 0), (1.4, -1), (1.4, -0.5), (1.4, 0)\},$$

and vary $\log_{10}\left(\frac{1}{1-\rho}\right)$ from 1 to 6, in increments of 0.1. For each combination of (z_1, z_2) and ρ, we compute the average regret under our policy and that under MLE-CYCLE over 100 simulations. Let Regret_3 (resp., Regret_4) denote the average regret under MLE-CYCLE (resp., our policy).

Figure 1 (resp., Figure 2) plots the relative difference between the average regret under policy MLE-CYCLE in BR and and that under our policy for the linear (resp., logit) model. For both figures, each scenario corresponds to a given choice $\boldsymbol{z} = (z_1, z_2)$ of the true demand parameters and the red line plots the average of the nine scenarios. A consistent observation in both the figures, and across the nice scenarios in each figure, is that our policy performs better when the discount factor is sufficiently close to 1. This is a combined consequence of two reasons: First, we note that when ρ is sufficiently close to 1, the effect of discounting is minimal in the sense that the revenue earned later in time is nearly as important as that earned earlier. Second, our policy obtains better estimates of the unknown parameters by front-loading all the exploration – this benefits us later in the exploitation phase. Specifically, in the comparison of our policy with policy MLE-CYCLE in BR, while the exploitation phase under our policy starts later in time than that under MLE-CYCLE, the estimates of the unknown parameters of the demand distribution in the exploitation phase under

Figure 1: The relative percentage difference between the average regret under policy MLE-CYCLE and that under our policy for a linear model.

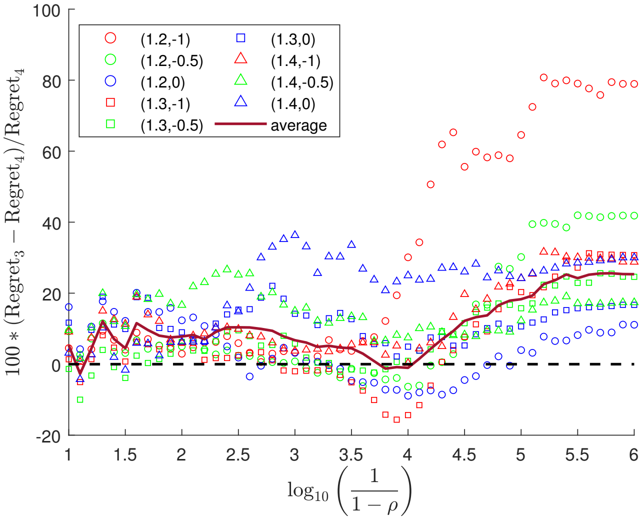


Note: Regret_1 is the average regret under MLE-CYCLE and Regret_2 is the average regret under our policy.

our policy are based on a higher number of observations and are thus more accurate than those in the early exploitation periods under MLE-CYCLE. Consequently, the expected revenue earned in the exploitation phase under our policy is higher than that under MLE-CYCLE, leading to a better performance. In Appendix E, we also demonstrate the robustness of the superior performance of our policy when ρ is sufficient close to 1, by examining the relative difference between the regret under MLE-CYCLE and that under our policy for different values of the time horizon T.

Figures 1 and 2 also show that when the discount factor is modest (i.e., not very close to 1; say 0.9), our policy typically performs better than policy MLE-CYCLE in BR. However, this behavior is relatively less sharp (as compared to the case when the discount factor is very close to 1): MLE-CYCLE may sometimes perform better on instances in which the expected revenue under the exploration prices is sufficiently close to the (clairvoyant) optimal revenue. The reasoning is as follows. First, recall that our policy explicitly takes the discount factor into consideration; therefore, unlike MLE-CYCLE, the length of the exploration phase in our policy depends on the discount factor. Table A.1 in Appendix D shows the number of exploration periods as a function of the discount factor for our policy and for policy MLE-CYCLE. As shown in Table A.1, when the discount factor is modest, since the revenue obtained in the latter periods is of little consequence, it is favorable to start exploitation early, and therefore, our policy spends relatively less amount of time on exploration. With a shorter exploration

Figure 2: The relative percentage difference between the average regret under policy MLE-CYCLE and that under our policy for a logit model.



Note: Regret_3 is the average regret under MLE-CYCLE and Regret_4 is the average regret under our policy.

phase, our policy switches to exploitation early while MLE-CYCLE continues with exploration. Thus, the shorter exploration phase in our policy results in a better relative performance for most of the instances in our test bed. However, in some cases, the shorter exploration phase in our policy leads to estimates of the unknown parameters that are not sufficiently close to their true values. In such cases, if the expected revenue under the exploration prices is sufficiently close to the optimal revenue, then the expected revenue obtained by MLE-CYCLE under the exploration prices could be higher than the expected revenue obtained by our policy (in the corresponding exploitation periods) under the greedy prices based on our estimates. Consequently, MLE-CYCLE may perform better since it spends relatively more time on exploration.

3 Incorporating Discounting in the Model in KZ (2014)

We now consider the single-product setting in KZ. Consider a firm selling a product over a time horizon of T periods. Given price $p_t \in [l, u]$ in period t, the demand in that period is $D_t = \alpha + \beta \cdot p_t + \epsilon_t$, where α, β are unknown parameters, and $\{\epsilon_t\}_{t=1,\cdots,T}$ are unobservable demand shocks that are independent and identically distributed random variables with mean zero and variance σ^2. In addition, there exists a positive constant x_0 such that $\mathbb{E}[e^{\epsilon_t \cdot x}] < \infty$ for all $|x| \leq x_0$. Let $\theta = (\alpha, \beta) \in \Theta$ be the vector of the unknown demand parameters, where $\Theta \subseteq \mathbb{R}^2$ is a compact rectangle. Note that $\beta < 0$; we assume that $\beta \in [b_{\min}, b_{\max}]$ for $b_{\min} < b_{\max} < 0$. Then, the seller's expected single-period revenue is

$$r_\theta(p) := p \cdot (\alpha + \beta p) \text{ for } \theta \in \Theta \text{ and } p \in [l, u].$$

Let $\varphi(\theta)$ be the optimal price that maximizes the expected single-period revenue, i.e., $\varphi(\theta) := \arg\max\{r_\theta(p) : p \in [l, u]\}$. We assume that $\varphi(\theta)$ is in the interior of $[l, u]$; thus, $\varphi(\theta) = \frac{-\alpha}{2\beta}$. Let $r^*_\theta := r_\theta(\varphi(\theta))$ denote the seller's optimal single-period revenue.

Let $H_t := (D_1, p_1; D_2, p_2; \cdots ; D_t, p_t)$ denote the history of demands and prices in the first t periods. Let $\pi = (\pi_1, \pi_2 \cdots)$ denote a non-anticipating policy, where π_t is a mapping from $\mathbb{R}^{2t-2}$ into $[l, u]$ that sets the price in period t based on H_{t-1}. Then, in the presence of a discount factor $\rho \in [0, 1)$, the seller's T-period expected discounted revenue under policy π is

$$R^\pi_\theta(T, \rho) = \mathbb{E}\left\{\sum_{t=1}^{T} \rho^{t-1} r_\theta(\pi_t(H_{t-1}))\right\}.$$

The T-period regret is then defined as

$$\Delta^\pi_\theta(T, \rho) := \sum_{t=1}^{T} \rho^{t-1} r^*_\theta - R^\pi_\theta(T, \rho).$$

Let $\Delta^\pi(T, \rho) := \sup_{\theta\in\Theta} \Delta^\pi_\theta(T, \rho)$ be the worst-case regret over Θ.

3.1 A Lower Bound on the Regret

Theorem 4 below establishes a lower bound on the regret under *any* pricing policy.

Theorem 4 *There exists a finite positive constant K_4 that is independent of ρ and T such that*

$$\Delta^\pi(T, \rho) \geq K_4 \sqrt{\frac{\rho(1 - \rho^{2T-2})}{1 - \rho^2}} \quad \textit{for any policy } \pi, \rho \in [0, 1), \ \textit{and } T \geq 3.$$

Further, $\lim_{\rho\to 1} \Delta^\pi(T, \rho) = \Omega\left(\sqrt{T}\right)$ *and* $\lim_{T\to\infty} \Delta^\pi(T, \rho) = \Omega\left(\sqrt{\frac{1}{1-\rho}}\right)$.

3.2 Asymptotically Optimal Learning Algorithms

In Section 3.2.1, we present sufficient conditions (in Theorem 5) under which a learning policy achieves an upper bound of $\mathcal{O}\left(\sqrt{\frac{1}{1-\rho}} \log \frac{1}{1-\rho}\right)$ (resp., $\mathcal{O}(\sqrt{T} \log T)$) on the regret when $T \to \infty$ (resp., $\rho \to 1$). In Section 3.2.2, we show that, in the presence of discounting, the two policies in KZ satisfy our conditions in Theorem 5. Thus, they are asymptotically optimal. In Section 3.2.3, we propose another policy and show (in Theorem 6) that the regret under our policy is also $\mathcal{O}\left(\sqrt{\frac{1}{1-\rho}} \log \frac{1}{1-\rho}\right)$ (resp., $\mathcal{O}(\sqrt{T} \log T)$) when $T \to \infty$ (resp., $\rho \to 1$).

3.2.1 Sufficient Conditions for Asymptotic Optimality under Discounting

In the presence of discounting, we modify the conditions in Theorem 2 of KZ, and establish an $\mathcal{O}\left(\sqrt{\frac{1}{1-\rho}}\log\frac{1}{1-\rho}\right)$ (resp., $\mathcal{O}(\sqrt{T}\log T)$) upper bound on the regret when $T \to \infty$ (resp., $\rho \to 1$), under any policy satisfying these conditions.

We first define the well-known *Greedy Iterated Least Squares* (ILS) policy (Anderson and Taylor 1976). Given the history of demands and prices in the first t periods, let $SSE_t(\theta) = \sum_{s=1}^{t}(D_s - \alpha - \beta \cdot p_s)^2$, where $\theta = (\alpha, \beta)$, and let $\hat{\theta}_t := \arg\min_\theta\{SSE_t(\theta)\}$ denote the least-squares estimate of θ based on these observations. Recall that θ lies in the compact rectangle Θ. Let $\vartheta_t := \arg\min_{\vartheta\in\Theta}\{\|\vartheta - \hat{\theta}_t\|\}$ denote the truncated estimate and assume that the optimal price based on ϑ_t (i.e., $\varphi(\vartheta_t)$) is an interior point of $[l, u]$. Then, the greedy ILS policy is defined as one that charges price $p_t = \varphi(\vartheta_{t-1})$ in period t. That is, the greedy ILS policy estimates the unknown demand parameters at the beginning of each time period, and then charges the "greedy" or "myopic" price based on these estimates.

Under the greedy ILS policy, parameter estimates may get stuck at values that are not the true ones; this is referred to as incomplete learning (den Boer and Zwart 2014 and KZ). Therefore, one needs to modify the greedy ILS and impose additional conditions to ensure that the regret is asymptotically optimal. KZ show that price deviations can help gather information to learn the unknown parameters. Consider a policy π and let p_s^π denote the price in period s under this policy. For the problem without discounting, KZ characterize two conditions under which π is asymptotically optimal. In condition (i) of Theorem 2 in KZ, the authors impose a lower bound on the sum of the squared price deviations under policy π, denoted by $J_t^\pi := \sum_{s=1}^{t}(p_s^\pi - \bar{p}_t^\pi)^2$, where $\bar{p}_t^\pi = t^{-1}\sum_{s=1}^{t} p_s^\pi$. In condition (ii), they impose an upper bound on the sum of the squared difference between the price under policy π and the greedy price. In the presence of discounting, condition (i) of Theorem 5 below is the same as that of Theorem 2 in KZ. However, condition (ii) of their Theorem 2 needs to be modified to capture the sum of the *discounted* squared difference between the price under policy π and the greedy price.

Theorem 5 *Let κ_0, κ_1, κ_2 be finite positive constants and $\kappa_3 \in \mathbb{N}$. Let π be a pricing policy that satisfies*

(i) $J_t^\pi \geq \kappa_0\sqrt{t}$, and

(ii) $\sum_{s=t_1}^{t_2} \rho^s(\varphi(\vartheta_s) - p_{s+1}^\pi)^2 \leq \kappa_1 t_1 + \kappa_2 \sum_{s=\kappa_3 t_1^2}^{t_2} \rho^s s^{-1/2}$,[3]

[3] Note that in condition (ii), if $t_2 < \kappa_3 t_1^2$, then the term $\sum_{s=\kappa_3 t_1^2}^{t_2} \rho^s s^{-1/2}$ is 0.

almost surely for all $t \geq 2$ and $t_2 \geq t_1$, where p^π_{s+1} is the price in period $s+1$ under policy π and $\varphi(\vartheta_s)$ is the greedy price in period $s+1$. Then, there exist positive constants K_5, K_6, K_7, and $N \in \mathbb{N}$, that are independent of ρ and T, such that

$$\Delta^\pi(T,\rho) \leq K_5 \frac{1-\rho^N}{1-\rho} + K_6 N \log N + K_7 \frac{\log N}{N} \frac{1-\rho^{T-N^2}}{1-\rho},$$

for $T \geq 3$ and $\rho \in [0,1)$. Further, for $N = \left\lfloor \sqrt{\frac{1-\rho^T}{1-\rho}} \right\rfloor$, we have

$$\lim_{\rho \to 1} \Delta^\pi(T,\rho) = \mathcal{O}\left(\sqrt{T}\log T\right) \quad \textit{and} \quad \lim_{T\to\infty} \Delta^\pi(T,\rho) = \mathcal{O}\left(\sqrt{\frac{1}{1-\rho}} \log\left(\frac{1}{1-\rho}\right)\right).$$

3.2.2 Analysis of the ILS Variants in Keskin and Zeevi (2014) with Discounting

We now present two variants of greedy ILS policy in KZ. We show that, in the presence of discounting, these two variants satisfy our conditions in Theorem 5 and are, therefore, asymptotically optimal. For notational simplicity, we henceforth drop the superscript π from p^π_t, $\bar{p}^\pi_t$, and J^π_t.

Example 1: *Constrained Iterated Least Squares* (CILS). Let $\delta_s := \varphi(\vartheta_{s-1}) - \bar{p}_{s-1}$ denote the difference between the greedy ILS price in period s and the average price in the first $s-1$ periods. For a positive constant c_1, a CILS policy (referred to as CILS(c_1)) charges the following prices:

$$p_t = \begin{cases} \bar{p}_{t-1} + \operatorname{sgn}(\delta_t) c_1 t^{-1/4} & \text{if } |\delta_t| < c_1 t^{-1/4} \\ \varphi(\vartheta_{t-1}) & \text{otherwise.} \end{cases}$$

Any policy π in the CILS family $\{\text{CILS}(c_1) : c_1 > 0\}$ satisfies the conditions of Theorem 5. KZ show that the sum of squared price deviations J_t satisfies $J_t \geq \frac{1}{4}c_1^2 t^{1/2}$. Thus, condition (i) of Theorem 5 is satisfied for $\kappa_0 = \frac{1}{4}c_1^2$. Moreover, note that the deviation from the greedy ILS price satisfies $|\varphi(\vartheta_{s-1}) - p_s| \leq c_1 s^{-1/4}$, and thus $\sum_{s=t_1}^{t_2} \rho^s (\varphi(\vartheta_s) - p_{s+1})^2 \leq c_1^2 \sum_{s=t_1}^{t_2} \rho^s (s+1)^{-1/2} \leq 2c_1^2 t_1 + c_1^2 \sum_{s=t_1^2}^{t_2} \rho^s s^{-1/2}$. Therefore, condition (ii) is satisfied with $\kappa_1 = 2c_1^2$, $\kappa_2 = c_1^2$, and $\kappa_3 = 1$. Consequently, any policy in the CILS family achieves the performance guarantee in Theorem 5. ■

Example 2: *ILS with Deterministic Testing* (ILS-d). Let $\tilde{p}_1, \tilde{p}_2$ be two distinct prices in $[l,u]$, and let $\{\mathcal{F}_{1,t}\}$, $\{\mathcal{F}_{2,t}\}$ be two sequences of sets satisfying the following conditions. For each $i \in \{1,2\}$ and $t \in \mathbb{N}$, $\mathcal{F}_{i,t} \subseteq \mathcal{F}_{i,t+1}$, where $\mathcal{F}_{1,t}$, $\mathcal{F}_{2,t}$ are disjoint subsets of $\{1,2,\cdots,t\}$. Further, $\mathcal{F}_{1,t}$ (resp., $\mathcal{F}_{2,t}$) contains $\lfloor\sqrt{t}\rfloor$ (resp., $\lfloor\sqrt{t-1}\rfloor$) distinct elements. An ILS-d policy with experimental prices $\tilde{p}_1$ and $\tilde{p}_2$ (referred to as ILS-d$(\tilde{p}_1, \tilde{p}_2)$) charges the following prices:

$$p_t = \begin{cases} \tilde{p}_1 & \text{if } t \in \mathcal{F}_{1,t} \\ \tilde{p}_2 & \text{if } t \in \mathcal{F}_{2,t} \\ \varphi(\vartheta_{t-1}) & \text{otherwise.} \end{cases}$$

Thus, $\tilde{p}_1$ is used in period $t \in \mathcal{F}_{1,T} = \{1, 4, 9, 16, \cdots, (\lfloor\sqrt{T}\rfloor)^2\}$ and $\tilde{p}_2$ is used in period $t \in \mathcal{F}_{2,T} = \{2, 5, 10, 17, \cdots, (\lfloor\sqrt{T-1}\rfloor)^2 + 1\}$. Next, we show that any policy π in the ILS-d family $\{\text{ILS-d}(\tilde{p}_1, \tilde{p}_2) : \tilde{p}_1 \neq \tilde{p}_2, \tilde{p}_1 \in [l, u], \tilde{p}_2 \in [l, u]\}$ satisfies the conditions of Theorem 5. Note that at least $\lfloor\sqrt{t-1}\rfloor$ experiments are conducted with $\tilde{p}_1$ and $\tilde{p}_2$ each in the first t periods; this implies that $J_t = \sum_{s=1}^{t}(p_s - \bar{p}_t)^2 \geq \sum_{s \in \mathcal{F}_{1,t} \cup \mathcal{F}_{2,t}}(p_s - \bar{p}_t)^2 \geq \kappa_0\sqrt{t}$ for $\kappa_0 = \frac{1}{4}(\tilde{p}_1 - \tilde{p}_2)^2$. In addition, we have

$$\begin{aligned} & \sum_{s=t_1}^{t_2} \rho^s(\varphi(\vartheta_s) - p_{s+1})^2 \\ \leq & \sum_{s=t_1}^{t_2} \rho^s(\lfloor\sqrt{s+1}\rfloor - \lfloor\sqrt{s}\rfloor)(u-l)^2 + \sum_{s=t_1}^{t_2} \rho^s(\lfloor\sqrt{s}\rfloor - \lfloor\sqrt{s-1}\rfloor)(u-l)^2 \\ \leq & \ (u-l)^2\Big[\sum_{s=t_1}^{t_1^2-1}(\lfloor\sqrt{s+1}\rfloor - \lfloor\sqrt{s}\rfloor) + \sum_{s=t_1^2}^{t_2} \rho^s(\sqrt{s+1} - \sqrt{s}) + \sum_{s=t_1}^{t_1^2}(\lfloor\sqrt{s}\rfloor - \lfloor\sqrt{s-1}\rfloor) + \\ & \sum_{s=t_1^2+1}^{t_2} \rho^s(\sqrt{s} - \sqrt{s-1})\Big] \\ \leq & \ 2(u-l)^2 \left[t_1 + \sum_{s=t_1^2}^{t_2} \rho^s s^{-1/2}\right]. \end{aligned}$$

Note that when $s = t_1^2$ (resp., $s = t_1^2{+}1$), we have $\sqrt{s} = t_1 \in \mathbb{N}$ (resp., $\sqrt{s-1} = t_1 \in \mathbb{N}$), which implies the second inequality. Then, condition (ii) holds with $\kappa_1 = 2(u-l)^2$, $\kappa_2 = 2(u-l)^2$, and $\kappa_3 = 1$. ■

3.2.3 A Different ILS Variant with Discounting

We note that the conditions in Theorem 5 are sufficient, but not necessary for establishing the claimed results. We now present a policy that does not satisfy the conditions in Theorem 5, and show that the regret under this policy is also $\mathcal{O}\left(\sqrt{\frac{1}{1-\rho}}\log\frac{1}{1-\rho}\right)$ (resp., $\mathcal{O}(\sqrt{T}\log T)$) when $T \to \infty$ (resp., $\rho \to 1$).

Our pricing policy is also a variant of the greedy ILS policy. Recall from Section 2.2 that $\tau = \left\lceil\sqrt{\frac{1-\rho^T}{1-\rho}}\right\rceil$. For $c_2 \in \mathbb{N}$, we conduct price experiments in the first $2c_2\tau$ periods using two distinct prices $\tilde{p}_1$ and $\tilde{p}_2$. In the remaining $T - 2c_2\tau$ periods, we offer the greedy ILS price $\varphi(\vartheta_{t-1})$. Specifically, our policy, which we denote by $\hat{\pi}$, charges the following prices:

$$p_t = \begin{cases} \tilde{p}_1 & \text{if } t \in \{1, 3, \cdots, 2c_2\tau - 1\}, \\ \tilde{p}_2 & \text{if } t \in \{2, 4, \cdots, 2c_2\tau\}, \\ \varphi(\vartheta_{t-1}) & \text{otherwise.} \end{cases}$$

We show that our policy is asymptotically optimal. Let $\eta = \log\left(\frac{1-\rho^T}{1-\rho}\right)$. Then, we have

Theorem 6 *There exist positive constants $\check{\rho} \in [0, 1)$, $\check{T} \in \mathbb{N}$, and positive constants K_8 and K_9 that are independent of ρ and T, such that*

$$\Delta^{\hat{\pi}}(T, \rho) \leq K_8 \frac{1 - \rho^{2c_2\tau}}{1 - \rho} + K_9 \frac{\eta}{\tau} \frac{\rho^{2c_2\tau} - \rho^T}{1 - \rho},$$

for $\rho \geq \check{\rho}$ and $T \geq \check{T}$. Further,

$$\lim_{\rho \to 1} \Delta^{\hat{\pi}}(T, \rho) = \mathcal{O}\left(\sqrt{T} \log T\right) \quad \textit{and} \quad \lim_{T \to \infty} \Delta^{\hat{\pi}}(T, \rho) = \mathcal{O}\left(\sqrt{\frac{1}{1 - \rho}} \log\left(\frac{1}{1 - \rho}\right)\right).$$

Note that if $\rho = 1$, then for $\tau = [\sqrt{T}]$, policies CILS, ILS-d, and our policy $\hat{\pi}$, all satisfy the conditions in Lemma 1 below, which generalizes the conditions in Theorem 2 of KZ and ensures that the regret under any policy that satisfies these conditions is $\mathcal{O}(\sqrt{T} \log T)$.

Lemma 1 *Suppose $\rho = 1$. Let κ_0, κ_1 be finite positive constants, and let π be a pricing policy that satisfies*

(i) $J_t \geq \kappa_0 \sqrt{t}$, and

(ii) $\sum_{s=0}^{T-1} (\varphi(\vartheta_s) - p_{s+1})^2 \leq \kappa_1 \sqrt{T}$

almost surely for all $t \geq 2$, where p_{s+1} is the price in period $s+1$ under policy π and $\varphi(\vartheta_s)$ is the greedy price in period $s + 1$. Then, the regret $\Delta^{\pi}(T, 1)$ under policy π satisfies $\Delta^{\pi}(T, 1) = \mathcal{O}(\sqrt{T} \log T)$.

3.3 Numerical Experience

We numerically examine the behavior of the regret under our policy, the CILS policy, and the ILS-d policy. Given a price $p \in [l, u] = [0.75, 2]$, demand is normally distributed with mean $\alpha + \beta p$ and standard deviation $\sigma = 0.1$, where $\theta = (\alpha, \beta) \in \Theta = [1, 1.4] \times [-0.64, -0.36]$. We set the experimental prices in policy ILS-d and our policy to be $(\tilde{p}_1, \tilde{p}_2) = (0.75, 1.75)$, and the total number of periods to be $T = 40000$. The constant c_1 in CILS is set to 0.55 and the constant c_2 in our policy is set to 1. We consider the following nine scenarios of the true demand parameters:

$$\theta = (\alpha, \beta) \in \Big\{(1.15, -0.45), (1.15, -0.5), (1.15, -0.55), (1.2, -0.45), (1.2, -0.5), (1.2, -0.55),$$

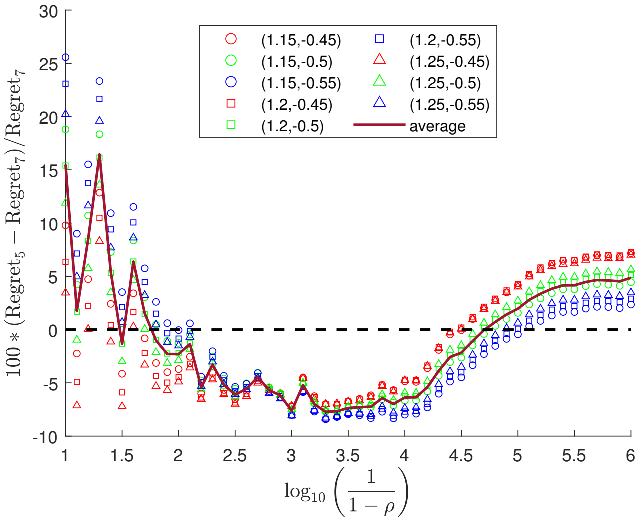
$$(1.25, -0.45), (1.25, -0.5), (1.25, -0.55)\Big\},$$

and vary $\log_{10}\left(\frac{1}{1-\rho}\right)$ from 1 to 6, in increments of 0.1. For each combination of θ and ρ, we compute the average regret under our policy (resp., ILS-d and CILS) over 100 simulations. Let Regret_7 (resp., Regret_5 and Regret_6) denote the average regret under our policy (resp., ILS-d and CILS). Figure 3 plots the relative difference between the average regret under ILS-d and that under our policy. Recall from Section 3.2.2 that under ILS-d, $\tilde{p}_1$ is used in period $t \in \mathcal{F}_{1,T} = \{1, 4, 9, 16, \cdots, 40000\}$ and $\tilde{p}_2$ is used in period $t \in \mathcal{F}_{2,T} = \{2, 5, 10, 17, \cdots, 39602\}$. That is, the exploration periods are $\mathcal{F}_{1,T} \cup \mathcal{F}_{2,T} = \{1, 2, 4, 5, 9, 10, 16, 17, \cdots, 39602, 40000\}$. The structure of ILS-d is similar to that of MLE-CYCLE, in the sense that both policies alternate between exploration and exploitation in cycles. As in Figures 1 and 2, Figure 3 illustrates the better performance of our policy when the discount factor is sufficiently close to 1, confirming the robustness of our finding. Note that when the discount factor ρ is sufficiently close to 1, the effect of discounting is minimal in the sense that the revenue earned later in time is nearly as important as that earned earlier. As compared to the ILS-d policy, our policy obtains better estimates of the unknown parameters by front-loading all the exploration. Consequently, the expected revenue earned in the exploitation phase under our policy is higher than that under ILS-d, leading to a better performance. We also demonstrate the robustness of the superior performance of our policy when ρ is sufficient close to 1, by examining the relative difference between the regret under ILS-d and that under our policy for different T (see Appendix E). Figure 3 also shows that when ρ is modest, while our policy typically performs better than policy ILS-d, this behavior is relatively less sharp – the underlying reason is the same as that discussed earlier in Section 2.3 for the comparison between our policy and MLE-CYCLE. Table A.1 in Appendix D shows the number of exploration periods as a function of the discount factor for our policy and for policy ILS-d. Figure 4 plots the relative difference between the average regret under the CILS policy and that under our policy, and again underlines the better performance of our policy for discount factors sufficiently close to 1.

4 Dependent Discount Factor and Duration of Planning Horizon

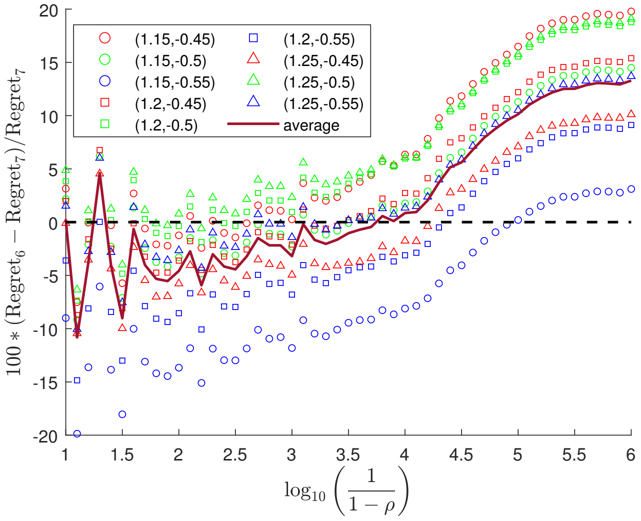
Recall that our analysis in Sections 2 and 3 did not assume any relationship between the discount factor, ρ, and the duration of planning horizon, T. As a result, our expressions for the regret bounds were functions of both ρ and T. For any given planning horizon which has been divided into decision periods, one can define the precise change in the value of money with time by examining a specific relationship between the discount rate per period and the discount rate over the planning horizon.

Figure 3: The relative percentage difference between the average regret under policy ILS-d and that under our policy.



Note: Regret_5 is the average regret under ILS-d and Regret_7 is the average regret under our policy.

Figure 4: The relative percentage difference between the average regret under policy CILS and that under our policy.



Note: Regret_6 is the average regret under CILS and Regret_7 is the average regret under our policy.

Motivated by this fact, we now revisit our earlier analysis by assuming a specific expression of the discount factor per period as a function of the number of decision periods in the planning horizon. To this end, we examine the following setting: Consider a planning horizon (say, one quarter) over which the discount rate is $\rho_0 \in (0, 1)$. The planning horizon is divided into T decision periods. Thus, the effective discount rate per decision period is $\rho(T) = (\rho_0)^{1/T}$. Note that as T goes to infinity, the discount factor in each period, $\rho(T)$, goes to 1. In our analysis, we also consider an asymptotic regime where the number of decision periods approaches infinity, keeping the demand volume per period the same – this can be interpreted as a business context where the quarterly sales volume increases. As before, we derive a lower bound on the regret under any policy and an upper bound on the regret under the algorithms in BR and KZ as well as our new algorithms. We show that for the models in BR and KZ, the regret under any policy is $\Omega(\sqrt{T})$. For the model in BR, we show that the regret under our policy as well that under the MLE-CYCLE policy in BR is $\mathcal{O}(\sqrt{T})$. For the model of KZ, we show that the regret is $\mathcal{O}(\log T\sqrt{T})$ under three policies – namely, the two variants of the greedy Iterated-Least-Squares policy in KZ and a different policy that we propose. These results are presented in Propositions A.1 through A.6 in Appendix C.

5 Concluding Remarks

We offer the following three directions in which future research can proceed:

- In this paper, we investigate the impact of discounting on the performance of learning algorithms by examining two classic and representative dynamic-pricing and learning problems studied in BR and KZ. In both settings, we study the dynamic pricing of a single product with stationary unknown demand in the presence of discounting. The generalization of our analysis to multiple products is an important question. KZ extend their single-product results to the case of multiple products with substitutable demands – this is a challenging extension of single-product pricing. BR focus on the single-product setting but their results can also be extended easily to multiple products with independent demands; if, instead, the demands for the products are dependent, then their analysis does not seem to extend in a straightforward manner. We believe that the generalization of our work to multiple products with dependent demands is a challenging and fruitful direction in which future work can proceed.

- Keskin and Zeevi (2017) consider a linear demand model with unknown and time-varying parameters, and study dynamic pricing and demand learning under changing demand environments;

i.e., under non-stationary demand. They derive a lower bound on the regret under any policy, given a finite variation “budget”, and propose families of dynamic-pricing policies that achieve a matching upper bound on the regret. They also illustrate how the manner in which the demand environment changes matters: the decision maker can earn more revenue when the demand changes in “bursts” rather than when it changes “smoothly”. In the presence of discounting, a changing demand environment imposes another challenge on the design of dynamic-pricing policies, since the decision maker now also needs to consider the timing of the demand changes to exploit the knowledge of the underlying parameters that vary over time and with different patterns. We believe that our analysis in this paper can serve as a stepping stone for understanding how non-stationary demand can affect dynamic pricing and demand learning in the presence of discounting.

- It will be interesting to investigate how discounting can be incorporated in algorithms such as CILS (Keskin and Zeevi 2014) that simultaneously conduct exploration and exploitation.

Acknowledgements

The authors are grateful to the area editor, the associate editor, and two anonymous reviewers for their many constructive comments and suggestions that significantly improved the paper in both content and exposition.

Author Biographies

Zhichao Feng is an assistant professor at the Department of Logistics and Maritime Studies, Faculty of Business, The Hong Kong Polytechnic University. His research interests are in revenue management and pricing, online advertising, queueing theory and its applications, and process analysis.

Milind Dawande is the Mike Redeker Distinguished Professor of Management at the Naveen Jindal School of Management, The University of Texas at Dallas. His research interests are in optimization theory and its applications to problems in manufacturing and service operations management.

Ganesh Janakiraman is the Ashbel Smith Professor in the operations management area of the Naveen Jindal School of Management, The University of Texas at Dallas. The primary research methodologies he employs are stochastic, dynamic optimization, and mechanism design. He applies these to inventory theory, sourcing, and real-time bidding in online advertising, etc.

Anyan Qi is an associate professor of operations management at the Naveen Jindal School of Management at The University of Texas at Dallas. The primary research methodologies he employs are data-driven optimization, game theory, and behavioral experiments. He applies these to learning algorithms, strategic procurement, socially responsible operations, and new product development.

References

Anderson, T. W., and J. B. Taylor. 1976. Some experimental results on the statistical properties of least squares estimates in control problems. *Econometrica: Journal of the Econometric Society*:1289–1302.

Araman, V. F., and R. Caldentey. 2009. Dynamic Pricing for Nonperishable Products with Demand Learning. *Operations Research* 57 (5): 1169–1188.

Baardman, L., E. Fata, A. Pani, and G. Perakis. 2019. Learning Optimal Online Advertising Portfolios with Periodic Budgets. *Available at SSRN 3346642*.

Besbes, O., Y. Fonseca, and I. Lobel. 2021. Contextual Inverse Optimization: Offline and Online Learning. *Available at SSRN 3863366*.

Besbes, O., and A. Zeevi. 2015. On the (Surprising) Sufficiency of Linear Models for Dynamic Pricing with Demand Learning. *Management Science* 61 (4): 723–739.

Borovkov, A. 1998. *Mathematical Statistics*. Amsterdam: Gordon and Breach.

Broder, J. 2011. Online Algorithms for Revenue Management. Doctoral Thesis, Cornell University, Ithaca, NY.

Broder, J., and P. Rusmevichientong. 2012. Dynamic Pricing under a General Parametric Choice Model. *Operations Research* 60 (4): 965–980.

Chen, B., X. Chao, and Y. Wang. 2020. Data-Based Dynamic Pricing and Inventory Control with Censored Demand and Limited Price Changes. *Operations Research* 68 (5): 1445–1456.

Chen, Q., S. Jasin, and I. Duenyas. 2021. Joint Learning and Optimization of Multi-Product Pricing with Finite Resource Capacity and Unknown Demand Parameters. *Operations Research* 69 (2): 560–573.

Chen, Y., and R. Wang. 1999. Learning Buyers' Valuation Distribution in Posted-Price Selling. *Economic Theory* 14 (2): 417–428.

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 B. Zwart. 2014. Simultaneously Learning and Optimizing Using Controlled Variance Pricing. *Management Science* 60 (3): 770–783.

Farias, V. F., and B. Van Roy. 2010. Dynamic Pricing with a Prior on Market Response. *Operations Research* 58 (1): 16–29.

Feng, Z., M. Dawande, G. Janakiraman, and A. Qi. 2023. An Asymptotically Tight Learning Algorithm for Mobile-Promotion Platforms. *Management Science* 69 (3): 1536–1554.

Jagabathula, S., L. Subramanian, and A. Venkataraman. 2020. A Conditional Gradient Approach for Nonparametric Estimation of Mixing Distributions. *Management Science* 66 (8): 3635–3656.

Keskin, N. B., and M. Li. 2023. Selling Quality-Differentiated Products in a Markovian Market with Unknown Transition Probabilities. *Operations Research*: Forthcoming.

Keskin, N. B., Y. Li, and N. Sunar. 2020. Data-Driven Clustering and Feature-Based Retail Electricity Pricing with Smart Meters. *Available at SSRN*.

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

Keskin, N. B., and A. Zeevi. 2017. Chasing Demand: Learning and Earning in a Changing Environment. *Mathematics of Operations Research* 42 (2): 277–307.

Kwon, H. D., S. A. Lippman, and C. S. Tang. 2012. Optimal Markdown Pricing Strategy with Demand Learning. *Probability in the Engineering and Informational Sciences* 26 (1): 77–104.

Lei, Y. M., S. Miao, and R. Momot. 2023. Privacy-Preserving Personalized Revenue Management. *Management Science*:Forthcoming.

Lyu, C., H. Zhang, and L. Xin. 2021. UCB-Type Learning Algorithms for Lost-Sales Inventory Models with Lead Times. *Available at SSRN 3944354*.

Mason, R., and J. Välimäki. 2011. Learning about the Arrival of Sales. *Journal of Economic Theory* 146 (4): 1699–1711.

Mintz, Y., A. Aswani, P. Kaminsky, E. Flowers, and Y. Fukuoka. 2020. Nonstationary Bandits with Habituation and Recovery Dynamics. *Operations Research* 68 (5): 1493–1516.

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

Zhang, H., X. Chao, and C. Shi. 2020. Closing the Gaps: An Online Learning Algorithm for Lost-Sales Inventory Systems with Lead Times. *Management Science* 66 (5): 1962–1980.

Zhang, J., and J. Chen. 2006. Bayesian Solution to Pricing and Inventory Control under Unknown Demand Distribution. *Operations Research Letters* 34 (5): 517–524.

Zhang, M., H. S. Ahn, and J. Uichanco. 2021. Data-Driven Pricing for a New Product. *Operations Research*: Forthcoming.

Online Appendix – *Dynamic Pricing and Learning with Discounting*

Zhichao Feng

Department of Logistics and Maritime Studies, Faculty of Business, The Hong Kong Polytechnic University

zhi-chao.feng@polyu.edu.hk

Milind Dawande, Ganesh Janakiraman, Anyan Qi

Naveen Jindal School of Management, The University of Texas at Dallas

milind@utdalla.edu, ganesh@utdallas.edu, axq140430@utdallas.edu

Appendix A Proofs of the Results in Section 2

A.1 Proof of Theorem 1

Preliminary Results

We first show some preliminary results, which are used in the proof of Theorem 1. In our analysis, we use a common quantitative measure of uncertainty known as the *KL divergence*.

Definition A.1 *(Definition 2.26 in Cover and Thomas 1999). For any probability measures Q_0 and Q_1 on a discrete sample space $\mathcal{Y}$, the KL divergence of Q_0 and Q_1 is*

$$\mathcal{K}(Q_0; Q_1) = \sum_{y \in \mathcal{Y}} Q_0(y) \log \left(\frac{Q_0(y)}{Q_1(y)} \right).$$

Broder and Rusmevichientong (2012) show the following properties of the problem class $\mathcal{C}_{LB}$ defined in the statement of Theorem 1, which are used to prove Lemmas A.2 and A.3 below.

Lemma A.1 *(Lemma EC.1.1 of Broder and Rusmevichientong 2012) For all $p \in \mathcal{P}$ and $z \in \mathcal{Z}$,*

1. $p^*(z) = \frac{1+2z}{4z}$.
2. $p^*(z_0) = 1$ *for* $z_0 = 1/2$.
3. $d(p^*(z_0); z) = 1/2$ *for all* $z \in \mathcal{Z}$.
4. $r(p^*(z); z) - r(p; z) \geq \frac{1}{3}(p^*(z) - p)^2$.
5. $|p^*(z) - p^*(z_0)| \geq \frac{1}{4}|z - z_0|$.
6. $|d(p; z) - d(p; z_0)| \leq |p^*(z_0) - p||z - z_0|$.

We recall that P_t^{ψ} is the random price in period t under policy ψ. For notational simplicity, we henceforth drop the superscript ψ from P_t^{ψ}.

Lemma A.2 *For $z_0 = 1/2$, $z \in \mathcal{Z}$, $T \geq 1$, and any policy ψ,*

$$\textit{Regret}(z_0, \mathcal{C}_{LB}, T, \rho, \psi) \geq \frac{1}{16(z_0 - z)^2} \left[\rho^{T-1} \mathcal{K}(Q_T^{\psi, z_0}; Q_T^{\psi, z}) + (1-\rho) \sum_{t=1}^{T-1} \rho^{t-1} \mathcal{K}(Q_t^{\psi, z_0}; Q_t^{\psi, z}) \right].$$

Proof of Lemma A.2: To show the lemma, we use the following Chain Rule for KL divergence (Theorem 2.5.3, Cover and Thomas 1999):

$$\mathcal{K}(Q_T^{\psi, z_0}; Q_T^{\psi, z}) = \sum_{t=1}^{T} \mathcal{K}\left(Q_t^{\psi, z_0}; Q_t^{\psi, z} \middle| \boldsymbol{Y_{t-1}} \right),$$

where $\mathcal{K}\left(Q_t^{\psi, z_0}; Q_t^{\psi, z} \middle| \boldsymbol{Y_{t-1}} \right) := \sum_{\boldsymbol{y_t} \in \{0,1\}^t} Q_t^{\psi, z_0}(\boldsymbol{y_t}) \log \left(\frac{Q_t^{\psi, z_0}(y_t | \boldsymbol{y_{t-1}})}{Q_t^{\psi, z}(y_t | \boldsymbol{y_{t-1}})} \right)$ is the conditional KL divergence. Similar to Broder and Rusmevichientong (2012), we show that

$$\mathcal{K}\left(Q_t^{\psi, z_0}; Q_t^{\psi, z} \middle| \boldsymbol{Y_{t-1}} \right) \leq 16(z_0 - z)^2 \mathbb{E}[r(p^*(z_0); z_0) - r(P_t; z_0)].$$

Thus, we have

$$\begin{aligned}
&\text{Regret}(z_0, \mathcal{C}_{LB}, T, \rho, \psi) \\
= &\sum_{t=1}^{T} \rho^{t-1} \mathbb{E}[r(p^*(z_0); z_0) - r(P_t; z_0)] \\
\geq &\sum_{t=1}^{T} \rho^{t-1} \frac{1}{16(z_0 - z)^2} \mathcal{K}\left(Q_t^{\psi, z_0}; Q_t^{\psi, z} \middle| \boldsymbol{Y_{t-1}} \right) \\
= &\frac{1}{16(z_0 - z)^2} \left[\sum_{t=1}^{T} \mathcal{K}\left(Q_t^{\psi, z_0}; Q_t^{\psi, z} \middle| \boldsymbol{Y_{t-1}} \right) - \sum_{t=2}^{T} (1-\rho) \mathcal{K}\left(Q_t^{\psi, z_0}; Q_t^{\psi, z} \middle| \boldsymbol{Y_{t-1}} \right) - \right. \\
&\left. \sum_{t=3}^{T} (\rho - \rho^2) \mathcal{K}\left(Q_t^{\psi, z_0}; Q_t^{\psi, z} \middle| \boldsymbol{Y_{t-1}} \right) - \cdots - \sum_{t=T}^{T} (\rho^{T-2} - \rho^{T-1}) \mathcal{K}\left(Q_t^{\psi, z_0}; Q_t^{\psi, z} \middle| \boldsymbol{Y_{t-1}} \right) \right] \\
= &\frac{1}{16(z_0 - z)^2} \left[\mathcal{K}(Q_T^{\psi, z_0}; Q_T^{\psi, z}) - (1-\rho)(\mathcal{K}(Q_T^{\psi, z_0}; Q_T^{\psi, z}) - \mathcal{K}(Q_1^{\psi, z_0}; Q_1^{\psi, z})) - \right. \\
&\left. (\rho - \rho^2)(\mathcal{K}(Q_T^{\psi, z_0}; Q_T^{\psi, z}) - \mathcal{K}(Q_2^{\psi, z_0}; Q_2^{\psi, z})) - \cdots - (\rho^{T-2} - \rho^{T-1})(\mathcal{K}(Q_T^{\psi, z_0}; Q_T^{\psi, z}) - \mathcal{K}(Q_{T-1}^{\psi, z_0}; Q_{T-1}^{\psi, z})) \right] \\
= &\frac{1}{16(z_0 - z)^2} \left[\rho^{T-1} \mathcal{K}(Q_T^{\psi, z_0}; Q_T^{\psi, z}) + (1-\rho) \sum_{t=1}^{T-1} \rho^{t-1} \mathcal{K}(Q_t^{\psi, z_0}; Q_t^{\psi, z}) \right].
\end{aligned}$$

The third equality holds by the Chain Rule for KL-divergence. ■

Lemma A.3 *For $z_0 = 1/2$, $z \in \mathcal{Z}$, $T \geq 2$, and any policy ψ,*

$$\textit{Regret}(z_0, \mathcal{C}_{LB}, T, \rho, \psi) + \textit{Regret}(z, \mathcal{C}_{LB}, T, \rho, \psi) \geq \frac{1}{6(12)^2} (z_0 - z)^2 \sum_{t=1}^{T-1} \rho^t e^{-\mathcal{K}(Q_t^{\psi, z_0}; Q_t^{\psi, z})}.$$

Proof of Lemma A.3 uses Lemma A.4 below.

Lemma A.4 *(Theorem 2.2, Tsybakov 2009) Let Q_0 and Q_1 be two probability distributions on a finite space $\mathcal{Y}$, with $Q_0(y), Q_1(y) > 0$ for all $y \in \mathcal{Y}$. Then for any function $J : \mathcal{Y} \to \{0, 1\}$,*

$$Q_0\{J = 1\} + Q_1\{J = 0\} \geq \frac{1}{2} e^{-\mathcal{K}(Q_0;Q_1)},$$

where $\mathcal{K}(Q_0; Q_1)$ denotes the KL divergence of Q_0 and Q_1.

Proof of Lemma A.3: We first define two intervals $C_{z_0} \subset \mathcal{P}$ and $C_z \subset \mathcal{P}$ by

$$C_{z_0} = \left\{ p : |p^*(z_0) - p| \leq \frac{1}{12}|z_0 - z| \right\} \text{ and } C_z = \left\{ p : |p^*(z) - p| \leq \frac{1}{12}|z_0 - z| \right\}.$$

By property 5 in Lemma A.1, i.e., $|p^*(z_0) - p^*(z)| \geq \frac{1}{4}|z_0 - z|$, C_{z_0} and C_z are disjoint. By property 4 in Lemma A.1, for each $\hat{z} \in \{z_0, z\}$, if $p \in \mathcal{P} \setminus C_{\hat{z}}$, then

$$r(p^*(\hat{z}); \hat{z}) - r(p; \hat{z}) \geq \frac{1}{3}(p - p^*(\hat{z}))^2 \geq \frac{1}{3(12)^2}(z_0 - z)^2.$$

Let $P_1, P_2, \cdots, P_T$ denote the sequence of random prices under policy ψ. Let $Pr_z\{A\}$ (resp., $Pr_{z_0}\{A\}$) denote the probability that event A occurs when the underlying parameter is z (resp., z_0). Then

$$\begin{aligned}
&\text{Regret}(z_0, \mathcal{C}_{LB}, T, \rho, \psi) + \text{Regret}(z, \mathcal{C}_{LB}, T, \rho, \psi) \\
\geq\ & \sum_{t=1}^{T-1} \rho^t \mathbb{E}[r(p^*(z_0); z_0) - r(P_{t+1}; z_0)] + \sum_{t=1}^{T-1} \rho^t \mathbb{E}[r(p^*(z); z) - r(P_{t+1}; z)] \\
\geq\ & \frac{1}{3(12)^2}(z_0 - z)^2 \sum_{t=1}^{T-1} \rho^t \left(Pr_{z_0}\{P_{t+1} \notin C_{z_0}\} + Pr_z\{P_{t+1} \notin C_z\} \right) \\
\geq\ & \frac{1}{3(12)^2}(z_0 - z)^2 \sum_{t=1}^{T-1} \rho^t \left(Pr_{z_0}\{P_{t+1} \in C_z\} + Pr_z\{P_{t+1} \notin C_z\} \right) \\
\geq\ & \frac{1}{6(12)^2}(z_0 - z)^2 \sum_{t=1}^{T-1} \rho^t e^{-\mathcal{K}(Q_t^{\psi,z_0}; Q_t^{\psi,z})}.
\end{aligned}$$

The last inequality holds by Lemma A.4. ■

Proof of Theorem 1

Let $z_1 = z_0 + \left(\frac{1-\rho}{\rho}\right)^{1/4}$. Then for $\rho \geq \frac{16}{17}$, we have $z_1 \in \mathcal{Z}$. Using Lemmas A.2 and A.3, we have

$$\begin{aligned}
&2(\text{Regret}(z_0, \mathcal{C}_{LB}, T, \rho, \psi) + \text{Regret}(z_1, \mathcal{C}_{LB}, T, \rho, \psi)) \\
\geq\ & \text{Regret}(z_0, \mathcal{C}_{LB}, T, \rho, \psi) + (\text{Regret}(z_0, \mathcal{C}_{LB}, T, \rho, \psi) + \text{Regret}(z_1, \mathcal{C}_{LB}, T, \rho, \psi)) \\
\geq\ & \frac{1}{16}\sqrt{\frac{\rho}{1-\rho}} \left[\rho^{T-1} \mathcal{K}(Q_T^{\psi,z_0}; Q_T^{\psi,z_1}) + (1-\rho) \sum_{t=1}^{T-1} \rho^{t-1} \mathcal{K}(Q_t^{\psi,z_0}; Q_t^{\psi,z_1}) \right] + \\
& \frac{1}{6(12)^2}\sqrt{\frac{1-\rho}{\rho}} \sum_{t=1}^{T-1} \rho^t e^{-\mathcal{K}(Q_t^{\psi,z_0}; Q_t^{\psi,z_1})}
\end{aligned}$$

$$
\begin{aligned}
&\geq \frac{1}{6(12)^2}\sqrt{(1-\rho)\rho}\left[\sum_{t=1}^{T-1}\rho^{t-1}\mathcal{K}(Q_t^{\psi,z_0};Q_t^{\psi,z_1})+\sum_{t=1}^{T-1}\rho^{t-1}e^{-\mathcal{K}(Q_t^{\psi,z_0};Q_t^{\psi,z_1})}\right]\\
&= \frac{1}{6(12)^2}\sqrt{(1-\rho)\rho}\sum_{t=1}^{T-1}\rho^{t-1}\left[\mathcal{K}(Q_t^{\psi,z_0};Q_t^{\psi,z_1})+e^{-\mathcal{K}(Q_t^{\psi,z_0};Q_t^{\psi,z_1})}\right]\\
&\geq \frac{1}{6(12)^2}\sqrt{(1-\rho)\rho}\sum_{t=1}^{T-1}\rho^{t-1}\\
&= \frac{1}{6(12)^2}\sqrt{\frac{\rho}{1-\rho}}(1-\rho^{T-1}). \qquad \text{(A-1)}
\end{aligned}
$$

The second inequality holds by Lemmas A.2 and A.3. The third inequality holds by the the fact that the KL divergence is nonnegative. The last inequality holds since $x+e^{-x}\geq 1$ for all $x \geq 0$.

Let $z_2 = z_0 + \left(\frac{(1-\rho)\rho^{T-2}}{1-\rho^{T-1}}\right)^{1/4}$. Note that

$$
\lim_{T\to\infty}\lim_{\rho\to 1}\left(\frac{(1-\rho)\rho^{T-2}}{1-\rho^{T-1}}\right)^{1/4} = 0.
$$

Thus, there exists $\hat{\rho}\in(\frac{16}{17},1)$ and $\hat{T}\in\mathbb{N}$ such that for all $\rho\geq\hat{\rho}$ and $T\geq\hat{T}$, $\left(\frac{(1-\rho)\rho^{T-2}}{1-\rho^{T-1}}\right)^{1/4}\leq 1/2$ and $z_2\in\mathcal{Z}$.

Note that for $z\in\mathcal{Z}$, $\mathcal{K}(Q_t^{\psi,z_0};Q_t^{\psi,z})$ is non-decreasing in t because

$$
\mathcal{K}(Q_t^{\psi,z_0};Q_t^{\psi,z}) = \mathcal{K}(Q_{t-1}^{\psi,z_0};Q_{t-1}^{\psi,z}) + \mathcal{K}(Q_t^{\psi,z_0};Q_t^{\psi,z}\Big|\,\boldsymbol{Y_{t-1}}) \geq \mathcal{K}(Q_{t-1}^{\psi,z_0};Q_{t-1}^{\psi,z}).
$$

The equality holds by the Chain Rule for KL divergence (Theorem 2.5.3, Cover and Thomas 1999). The inequality holds since $\mathcal{K}(Q_t^{\psi,z_0};Q_t^{\psi,z}\Big|\,\boldsymbol{Y_{t-1}})$ is nonnegative (Theorem 2.6.3, Cover and Thomas 1999).

Using Lemmas A.2 and A.3, we have

$$
\begin{aligned}
&2(\text{Regret}(z_0,\mathcal{C}_{LB},T,\rho,\psi)+\text{Regret}(z_2,\mathcal{C}_{LB},T,\rho,\psi))\\
\geq\;& \text{Regret}(z_0,\mathcal{C}_{LB},T,\rho,\psi)+(\text{Regret}(z_0,\mathcal{C}_{LB},T,\rho,\psi)+\text{Regret}(z_2,\mathcal{C}_{LB},T,\rho,\psi))\\
\geq\;& \frac{1}{16}\sqrt{\frac{1-\rho^{T-1}}{(1-\rho)\rho^{T-2}}}\left[\rho^{T-1}\mathcal{K}(Q_T^{\psi,z_0};Q_T^{\psi,z_2})+(1-\rho)\sum_{t=1}^{T-1}\rho^{t-1}\mathcal{K}(Q_t^{\psi,z_0};Q_t^{\psi,z_2})\right]+\\
&\frac{1}{6(12)^2}\sqrt{\frac{(1-\rho)\rho^{T-2}}{1-\rho^{T-1}}}\sum_{t=1}^{T-1}\rho^{t}e^{-\mathcal{K}(Q_t^{\psi,z_0};Q_t^{\psi,z_2})}\\
\geq\;& \frac{1}{16}\sqrt{\frac{1-\rho^{T-1}}{(1-\rho)\rho^{T-2}}}\rho^{T-1}\mathcal{K}(Q_T^{\psi,z_0};Q_T^{\psi,z_2})+\frac{1}{6(12)^2}\sqrt{\frac{(1-\rho)\rho^{T-2}}{1-\rho^{T-1}}}\sum_{t=1}^{T-1}\rho^{t}e^{-\mathcal{K}(Q_T^{\psi,z_0};Q_T^{\psi,z_2})}\\
=\;& \frac{1}{16}\sqrt{\frac{\rho^T(1-\rho^{T-1})}{1-\rho}}\mathcal{K}(Q_T^{\psi,z_0};Q_T^{\psi,z_2})+\frac{1}{6(12)^2}\sqrt{\frac{(1-\rho)\rho^{T-2}}{1-\rho^{T-1}}}\frac{\rho(1-\rho^{T-1})}{1-\rho}e^{-\mathcal{K}(Q_T^{\psi,z_0};Q_T^{\psi,z_2})}\\
\geq\;& \frac{1}{6(12)^2}\sqrt{\frac{\rho^T(1-\rho^{T-1})}{1-\rho}}\left[\mathcal{K}(Q_T^{\psi,z_0};Q_T^{\psi,z_2})+e^{-\mathcal{K}(Q_T^{\psi,z_0};Q_T^{\psi,z_2})}\right]\\
\geq\;& \frac{1}{6(12)^2}\sqrt{\frac{\rho^T(1-\rho^{T-1})}{1-\rho}}. \qquad \text{(A-2)}
\end{aligned}
$$

The second inequality holds by Lemmas A.2 and A.3. The third inequality holds by the the fact that the KL divergence is nonnegative (see Theorem 2.6.3 in Cover and Thomas 1999) and $\mathcal{K}(Q_t^{\psi,z_0}; Q_t^{\psi,z_2})$ is non-decreasing in t. The last inequality holds since $x + e^{-x} \geq 1$ for all $x \geq 0$.

Combining (A-1) and (A-2), we have

$$\begin{aligned}
& 2(\mathrm{Regret}(z_0, \mathcal{C}_{LB}, T, \rho, \psi) + \mathrm{Regret}(z_1, \mathcal{C}_{LB}, T, \rho, \psi) + \mathrm{Regret}(z_2, \mathcal{C}_{LB}, T, \rho, \psi)) \\
\geq\ & (\mathrm{Regret}(z_0, \mathcal{C}_{LB}, T, \rho, \psi) + \mathrm{Regret}(z_1, \mathcal{C}_{LB}, T, \rho, \psi)) + (\mathrm{Regret}(z_0, \mathcal{C}_{LB}, T, \rho, \psi) + \mathrm{Regret}(z_2, \mathcal{C}_{LB}, T, \rho, \psi)) \\
\geq\ & \frac{1}{(12)^3}\left(\sqrt{\frac{\rho}{1-\rho}}(1-\rho^{T-1}) + \sqrt{\frac{\rho^T(1-\rho^{T-1})}{1-\rho}}\right).
\end{aligned}$$

Then we have

$$\begin{aligned}
& \max_{z \in \{z_0, z_1, z_2\}} \mathrm{Regret}(z, \mathcal{C}_{LB}, T, \rho, \psi) \\
\geq\ & \frac{\mathrm{Regret}(z_0, \mathcal{C}_{LB}, T, \rho, \psi) + \mathrm{Regret}(z_1, \mathcal{C}_{LB}, T, \rho, \psi) + \mathrm{Regret}(z_2, \mathcal{C}_{LB}, T, \rho, \psi)}{3} \\
\geq\ & K_0\left(\sqrt{\frac{\rho}{1-\rho}}(1-\rho^{T-1}) + \sqrt{\frac{\rho^T(1-\rho^{T-1})}{1-\rho}}\right),
\end{aligned}$$

where $K_0 = \frac{1}{6(12)^3}$.

Let $f(\rho, T) = K_0\left(\sqrt{\frac{\rho}{1-\rho}}(1-\rho^{T-1}) + \sqrt{\frac{\rho^T(1-\rho^{T-1})}{1-\rho}}\right)$. Then we have

$$\begin{aligned}
& \lim_{\rho \to 1} f(\rho, T) = K_0\sqrt{T-1} = \Omega(\sqrt{T}), \text{ and} \\
& \lim_{T \to \infty} f(\rho, T) = K_0\sqrt{\frac{\rho}{1-\rho}} = \Omega(\sqrt{\frac{1}{1-\rho}}).
\end{aligned}$$

■

A.2 Proof of Theorem 2

Lemmas A.5 and A.6 below are used in the proof of Theorem 2. Similar to the proof of Lemma 3.7 in Broder and Rusmevichientong 2012), it is straightforward to obtain Lemma A.5 using the tail inequality for MLE based on IID Samples in Theore

Lemma A.5 ***(Mean-Squared Errors for MLE Based on IID Samples, Borovkov 1998)*** *For any $\tau \geq 1$, there exists a constant K_{mle} depending only on the exploration prices $\bar{\boldsymbol{p}}$ and the problem class $\mathcal{C}$ such that*

$$\mathbb{E}[\|\boldsymbol{Z}(\tau) - \boldsymbol{z}\|]^2 \leq \frac{K_{mle}}{\tau}.$$

Lemma A.6 is reproduced verbatim from Corollary 2.4 of Broder and Rusmevichientong (2012).

Lemma A.6 *For any problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ satisfying Assumption 1 and for any $\boldsymbol{z}, \hat{\boldsymbol{z}} \in \mathcal{Z}$,*

$$r(p^*(\boldsymbol{z}); \boldsymbol{z}) - r(p^*(\hat{\boldsymbol{z}}); \boldsymbol{z}) \leq c_r L^2 \|\boldsymbol{z} - \hat{\boldsymbol{z}}\|^2.$$

Proof of Theorem 2: First, we show an upper bound on the regret incurred during the exploration phase. Recall from Assumption 1 that the revenue function is twice differentiable. In addition, the pricing interval $\mathcal{P}$ is compact; thus, there exists a constant K_1 depending only on the problem class $\mathcal{C}$ such that $r(p^*(\boldsymbol{z}); \boldsymbol{z}) - r(p; \boldsymbol{z}) \leq K_1$ for all $p \in \mathcal{P}$ and $\boldsymbol{z} \in \mathcal{Z}$. Thus, the regret incurred during the exploration phase satisfies

$$\sum_{s=1}^{\tau} \sum_{l=1}^{k} \rho^{(s-1)k+l-1} \mathbb{E}[r(p^*(\boldsymbol{z}); \boldsymbol{z}) - r(\bar{p}_l; \boldsymbol{z})] \leq \frac{1 - \rho^{k\tau}}{1 - \rho} K_1. \tag{A-3}$$

Next, we show an upper bound on the regret incurred during the exploitation phase. During the exploitation phase, we use price $p^*(\boldsymbol{Z}(\tau))$ and we offer this price for all $T - k\tau$ periods. It follows from Lemmas A.5 and A.6 that

$$\mathbb{E}\left[r(p^*(\boldsymbol{z}); \boldsymbol{z}) - r(p^*(\boldsymbol{Z}(\tau)); \boldsymbol{z})\right] \leq c_r L^2 \mathbb{E}\left[\|\boldsymbol{z} - \boldsymbol{Z}(\tau)\|^2\right] \leq c_r L^2 \frac{K_{mle}}{\tau}.$$

Thus, the regret incurred during the exploitation phase satisfies

$$\sum_{t=k\tau+1}^{T} \rho^{t-1} \mathbb{E}[r(p^*(\boldsymbol{z}); \boldsymbol{z}) - r(p^*(\boldsymbol{Z}(\tau)); \boldsymbol{z})] \leq \frac{\rho^{k\tau} - \rho^T}{(1 - \rho)\tau} c_r L^2 K_{mle}. \tag{A-4}$$

Let $K_2 = c_r L^2 K_{mle}$. Combining (A-3) and (A-4), the cumulative regret under policy $\hat{\psi}$ satisfies

$$\text{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho, \hat{\psi}) \leq K_1 \frac{1 - \rho^{k\tau}}{1 - \rho} + K_2 \frac{\rho^{k\tau} - \rho^T}{(1 - \rho)\tau}.$$

Let $g(\rho, T) = K_1 \frac{1-\rho^{k\tau}}{1-\rho} + K_2 \frac{\rho^{k\tau} - \rho^T}{(1-\rho)\tau}$, where $\tau = \left[\sqrt{\frac{1-\rho^T}{1-\rho}}\right]$. Then we have $\lim_{\rho \to 1} \tau = [\sqrt{T}]$ and

$$\lim_{\rho \to 1} g(\rho, T) = \lim_{\rho \to 1} \left(K_1 \frac{1 - \rho^{k[\sqrt{T}]}}{1 - \rho} + K_2 \frac{\rho^{k[\sqrt{T}]} - \rho^T}{(1 - \rho)[\sqrt{T}]} \right) = K_1 k[\sqrt{T}] + K_2 \left(\frac{T}{[\sqrt{T}]} - k \right) = \mathcal{O}(\sqrt{T}).$$

Note that $\lim_{T \to \infty} \tau = [\sqrt{1/(1 - \rho)}]$. Thus, we have

$$\lim_{T \to \infty} g(\rho, T) = K_1 \frac{1 - \rho^{k[\sqrt{1/(1-\rho)}]}}{1 - \rho} + K_2 \frac{\rho^{k[\sqrt{1/(1-\rho)}]}}{(1 - \rho)[\sqrt{1/(1 - \rho)}]} = \mathcal{O}\left(\sqrt{\frac{1}{1 - \rho}}\right).$$

■

A.3 Proof of Theorem 3

Proof of Theorem 3: Note that the MLE-CYCLE policy $\check{\psi}$ operates in cycles. Broder and Rusmevichientong (2012) show that the regret incurred in each cycle is bounded from above by a constant, denoted by K_3. Note that cycle h starts in period $\frac{h^2+h(2k-1)-2k+2}{2}$ and the total number of cycles is no more than $\lfloor\sqrt{2T}\rfloor$ for $T \geq 2$. Thus, we have

$$\text{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho, \check{\psi}) \leq \sum_{h=1}^{\lfloor\sqrt{2T}\rfloor} \rho^{[h^2+h(2k-1)-2k]/2} K_3 \leq \left(1 + \sum_{h=1}^{\lfloor\sqrt{2T}\rfloor} \rho^{h^2/2}\right) K_3.$$

and

$$\begin{aligned}\text{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho, \check{\psi}) \leq K_3 + \sum_{h=1}^{\lfloor\sqrt{2T}\rfloor} \rho^{h^2/2} K_3 \leq K_3 + K_3 \int_0^{\sqrt{2T}} \rho^{h^2/2} dh &= K_3 + K_3 \int_0^{\sqrt{2T}} e^{\log(\rho)h^2/2} dh \\ &= K_3 + K_3 \sqrt{\frac{1}{2}} \int_0^T e^{\log(\rho)x} \sqrt{\frac{1}{x}} dx.\end{aligned}$$

The last equality holds by letting $x = h^2/2$. When $T \to \infty$,

$$\begin{aligned}\lim_{T\to\infty} K_3 \left(1 + \sqrt{\frac{1}{2}} \int_0^T e^{\log(\rho)x} \sqrt{\frac{1}{x}} dx\right) &= K_3 \left(1 + \sqrt{\frac{1}{2}} \sqrt{\frac{1}{-\log\rho}} \int_0^\infty e^{-w} w^{-1/2} dw\right) \\ &= K_3 \left(1 + \sqrt{\frac{\pi}{2}} \sqrt{\frac{1}{-\log\rho}}\right) \\ &= \mathcal{O}\left(\sqrt{\frac{1}{-\log\rho}}\right) = \mathcal{O}\left(\sqrt{\frac{1}{1-\rho}}\right).\end{aligned}$$

The first equality holds by letting $w = -\log(\rho)x$. The second equality holds since $\int_0^\infty e^{-w} w^{-1/2} dw = \Gamma(1/2) = \sqrt{\pi}$. When $\rho \to 1$,

$$\lim_{\rho\to 1} \left(1 + \sum_{h=1}^{\lfloor\sqrt{2T}\rfloor} \rho^{h^2/2}\right) K_3 \leq K_3(\sqrt{2T} + 1) = \mathcal{O}\left(\sqrt{T}\right).$$

■

Appendix B Proofs of the Results in Section 3

B.1 Proof of Theorem 4

Let p_t denote the random price in period t under policy π and $\mathcal{F}_t = \sum_{s=1}^t \begin{bmatrix} 1 & p_s \\ p_s & p_s^2 \end{bmatrix}$ denote the Fisher information matrix. It is straightforward that Lemma A.7 holds using Lemma 1 in Keskin and Zeevi (2014), which is used to show Theorem 4.

Lemma A.7 *There exist positive constants μ_0 and μ_1 such that*

$$\sup_{\theta\in\Theta} \left\{\sum_{t=2}^T \rho^{t-1} \mathbb{E}(p_t - \varphi(\theta))^2\right\} \geq \sum_{t=2}^T \rho^{t-1} \frac{\mu_0}{\mu_1 + \sup_{\theta\in\Theta}\{C(\theta)\mathbb{E}[\mathcal{F}_{t-1}]C(\theta)^\intercal\}}, \tag{A-5}$$

where $C(\cdot)$ is a 1×2 matrix function on Θ such that $C(\theta) = [-\varphi(\theta)\ 1]$.

Proof of Theorem 4: By the definition of $C(\theta)$, we have $C(\theta)\mathbb{E}[\mathcal{F}_{t-1}]C(\theta)^\intercal = \sum_{s=1}^{t-1}\mathbb{E}(p_s-\varphi(\theta))^2$. Thus, inequality (A-5) in Lemma A.7 is equivalent to the following:

$$\sup_{\theta\in\Theta}\left\{\sum_{t=2}^{T}\rho^{t-1}\mathbb{E}(p_t-\varphi(\theta))^2\right\} \geq \sum_{t=2}^{T}\rho^{t-1}\frac{\mu_0}{\mu_1+\sup_{\theta\in\Theta}\{\sum_{s=1}^{t-1}\mathbb{E}(p_s-\varphi(\theta))^2\}}, \tag{A-6}$$

By the definition of regret, we have

$$\begin{aligned}\Delta^\pi(T,\rho) &= \sup_{\theta\in\Theta}\left\{\sum_{t=1}^{T}\rho^{t-1}\mathbb{E}(r^*_\theta - r_\theta(p_t))\right\}\\ &= \sup_{\theta\in\Theta}\left\{-\beta\sum_{t=1}^{T}\rho^{t-1}\mathbb{E}(p_t-\varphi(\theta))^2\right\}\\ &\geq |b_{\max}|\sup_{\theta\in\Theta}\left\{\sum_{t=1}^{T}\rho^{t-1}\mathbb{E}(p_t-\varphi(\theta))^2\right\}.\end{aligned}$$

The second equality holds since $r^*_\theta - r_\theta(p_t) = \varphi(\theta)(\alpha+\beta\varphi(\theta)) - p_t(\alpha+\beta p_t)$ and we replace α with $-2\beta\varphi(\theta)$. Using (A-6), we have

$$\begin{aligned}\Delta^\pi(T,\rho) &\geq b^2_{\max}\mu_0\sum_{t=2}^{T}\rho^{t-1}\frac{1}{\mu_1|b_{\max}|+|b_{\max}|\sup_{\theta\in\Theta}\{\sum_{s=1}^{t-1}\mathbb{E}(p_s-\varphi(\theta))^2\}}\\ &\geq K_{10}\sum_{t=2}^{T}\rho^{t-1}\frac{1}{K_{11}|b_{\max}|\sup_{\theta\in\Theta}\{\sum_{s=1}^{t-1}\mathbb{E}(p_s-\varphi(\theta))^2\}}\\ &= K_{10}\sum_{t=2}^{T}\frac{\rho^{2t-3}}{K_{11}|b_{\max}|\sup_{\theta\in\Theta}\{\sum_{s=1}^{t-1}\rho^{t-2}\mathbb{E}(p_s-\varphi(\theta))^2\}}\\ &\geq K_{10}\sum_{t=2}^{T}\frac{\rho^{2t-3}}{K_{11}|b_{\max}|\sup_{\theta\in\Theta}\{\sum_{s=1}^{t-1}\rho^{s-1}\mathbb{E}(p_s-\varphi(\theta))^2\}}\\ &\geq K_{10}\sum_{t=2}^{T}\frac{\rho^{2t-3}}{K_{11}\Delta^\pi(t-1,\rho)}\\ &\geq K_{10}\sum_{t=2}^{T}\frac{\rho^{2t-3}}{K_{11}\Delta^\pi(T,\rho)} = \frac{K_{10}}{K_{11}\Delta^\pi(T,\rho)}\frac{\rho(1-\rho^{2T-2})}{1-\rho^2},\end{aligned}$$

where $K_{10}=\mu_0 b^2_{\max}$ and $K_{11} = 1+\frac{4\mu_1}{(u-l)^2} \geq 1+\frac{\mu_1}{\sup_{\theta\in\Theta}\{\mathbb{E}(p_1-\varphi(\theta))^2\}} \geq 1+\frac{\mu_1}{\sup_{\theta\in\Theta}\{\sum_{s=1}^{t-1}\mathbb{E}(p_s-\varphi(\theta))^2\}}$. The last inequality holds since $\Delta^\pi(t,\rho)$ increases in t.

Then, we have $\Delta^\pi(T,\rho)\geq\sqrt{\frac{K_{10}}{K_{11}}\frac{\rho(1-\rho^{2T-2})}{1-\rho^2}}$. Thus, $\Delta^\pi(T,\rho)\geq K_4\sqrt{\frac{\rho(1-\rho^{2T-2})}{1-\rho^2}}$, where $K_4=\sqrt{K_{10}/K_{11}}$.

When $T\to\infty$,

$$\lim_{T\to\infty}\Delta^\pi(T,\rho)\geq\lim_{T\to\infty}K_4\sqrt{\frac{\rho(1-\rho^{2T-2})}{1-\rho^2}} = K_4\sqrt{\frac{\rho}{1-\rho^2}} = \Omega\left(\sqrt{\frac{1}{1-\rho}}\right).$$

When $\rho\to 1$,

$$\lim_{\rho\to 1}\Delta^\pi(T,\rho)\geq\lim_{\rho\to 1}K_4\sqrt{\frac{\rho(1-\rho^{2T-2})}{1-\rho^2}} = K_4\sqrt{T-1} = \Omega(\sqrt{T}).$$

■

B.2 Proof of Theorem 5

We show Theorem 5 using Lemma A.8 below, which is reproduced verbatim from Lemma 3 of Keskin and Zeevi (2014).

Lemma A.8 *There exist finite positive constants λ and γ such that, under any pricing policy π,*

$$\mathbb{P}^{\pi}_{\theta}\{\|\hat{\theta}_t - \theta\| > \delta, J_t \geq m\} \leq \gamma t \exp(-\lambda(\delta \wedge \delta^2)m) \tag{A-7}$$

for all $\delta, m > 0$, and $t \geq 2$.

Proof of Theorem 5: Using Lemma A.8 and condition (i), Keskin and Zeevi (2014) show that, there exists a constant K_{12} such that

$$\mathbb{E}(\varphi(\theta) - p_{t+1})^2 \leq \frac{K_{12}\log t}{\lambda\kappa_0\sqrt{t}} + 2\mathbb{E}(\varphi(\vartheta_t) - p_{t+1})^2 \text{ for all } t \geq N,$$

where N satisfies $kN\exp(-\frac{1}{2}\lambda\kappa_0\sqrt{N}) \leq 1$. For $N \geq \exp$, we have

$$\begin{aligned}
&\sum_{t=N}^{T-1}\rho^t\mathbb{E}(\varphi(\theta) - p_{t+1})^2 \\
&\leq K_{12}\sum_{t=N}^{T-1}\rho^t\frac{\log t}{\lambda\kappa_0\sqrt{t}} + 2\sum_{t=N}^{T-1}\rho^t\mathbb{E}(\varphi(\vartheta_t) - p_{t+1})^2 \\
&\leq K_{12}\sum_{t=N}^{N^2-1}\rho^t\frac{\log t}{\lambda\kappa_0\sqrt{t}} + K_{12}\sum_{t=N^2}^{T-1}\rho^t\frac{\log t}{\lambda\kappa_0\sqrt{t}} + 2\left(\kappa_1 N + \kappa_2\sum_{s=\kappa_3 N^2}^{T-1}\rho^s s^{-1/2}\right) \\
&\leq 4K_{12}\log N\frac{N}{\lambda\kappa_0} + K_{12}\frac{2\log N}{\lambda\kappa_0 N}\frac{\rho^{N^2}(1-\rho^{T-N^2})}{1-\rho} + 2\kappa_1 N + \frac{2\kappa_2}{\sqrt{\kappa_3}N}\frac{\rho^{\kappa_3 N^2}(1-\rho^{T-\kappa_3 N^2})}{1-\rho} \\
&\leq K_{13}N\log N + K_{14}\frac{\log N}{N}\frac{1-\rho^{T-N^2}}{1-\rho},
\end{aligned}$$

where $K_{13} = \frac{4K_{12}}{\lambda\kappa_0} + 2\kappa_1$ and $K_{14} = \frac{2K_{12}}{\lambda\kappa_0} + \frac{2\kappa_2}{\sqrt{\kappa_3}}$. The third inequality holds since $\log t/\sqrt{t}$ decreases in t for $t \geq \exp^2$. Since $\beta \in [b_{\min}, b_{\max}]$, we have

$$\begin{aligned}
\Delta^{\pi}(T,\rho) = &\sup_{\theta\in\Theta}\left\{-\beta\sum_{t=0}^{T-1}\rho^t\mathbb{E}(p_{t+1} - \varphi(\theta))^2\right\} \\
&\leq |b_{\min}|\sup_{\theta\in\Theta}\left\{\sum_{t=0}^{N-1}\rho^t\mathbb{E}(p_{t+1} - \varphi(\theta))^2 + \sum_{t=N}^{T-1}\rho^t\mathbb{E}(p_{t+1} - \varphi(\theta))^2\right\} \\
&\leq |b_{\min}|\left\{\frac{1-\rho^N}{1-\rho}(u-l)^2 + K_{13}N\log N + K_{14}\frac{\log N}{N}\frac{1-\rho^{T-N^2}}{1-\rho}\right\} \\
&= K_5\frac{1-\rho^N}{1-\rho} + K_6 N\log N + K_7\frac{\log N}{N}\frac{1-\rho^{T-N^2}}{1-\rho},
\end{aligned}$$

where $K_5 = |b_{\min}|(u-l)^2$, $K_6 = |b_{\min}|K_{13}$, and $K_7 = |b_{\min}|K_{14}$. Let $h(\rho, T) = K_5 \frac{1-\rho^N}{1-\rho} + K_6 N \log N + K_7 \frac{\log N}{N} \frac{1-\rho^{T-N^2}}{1-\rho}$ and $N = \left\lfloor \sqrt{\frac{1-\rho^T}{1-\rho}} \right\rfloor$. When $\rho \to 1$, we have $N = \lfloor \sqrt{T} \rfloor$ and

$$\lim_{\rho \to 1} h(\rho, T) = K_5 \lfloor \sqrt{T} \rfloor + K_6 \lfloor \sqrt{T} \rfloor \log \lfloor \sqrt{T} \rfloor + K_7 \log \lfloor \sqrt{T} \rfloor \frac{(T - \lfloor \sqrt{T} \rfloor^2)}{\lfloor \sqrt{T} \rfloor} = \mathcal{O}(\sqrt{T} \log T).$$

When $T \to \infty$, we have $N = \left\lfloor \sqrt{\frac{1}{1-\rho}} \right\rfloor$ and

$$\begin{aligned}\lim_{T \to \infty} h(\rho, T) &= K_5 \frac{1-\rho^N}{1-\rho} + K_6 \left\lfloor \sqrt{\frac{1}{1-\rho}} \right\rfloor \log \left\lfloor \sqrt{\frac{1}{1-\rho}} \right\rfloor + K_7 \frac{1/(1-\rho)}{\left\lfloor \sqrt{1/(1-\rho)} \right\rfloor} \log \left\lfloor \sqrt{\frac{1}{1-\rho}} \right\rfloor \\ &= \mathcal{O}\left(\sqrt{\frac{1}{1-\rho}} \log \left(\frac{1}{1-\rho} \right) \right).\end{aligned}$$

■

B.3 Proof of Theorem 6

We first show Lemma A.9, which will be used to show Theorem 6.

Lemma A.9 *Let $\hat{\pi}$ denote our policy. Then there exist positive constants K_{15}, $\check{T} \in \mathbb{N}$, and $\check{\rho} \in [0,1)$ such that, under policy $\hat{\pi}$, for all $T \geq \check{T}$ and $\rho \geq \check{\rho}$, we have $\mathbb{E}(\varphi(\theta) - \varphi(\vartheta_t))^2 \leq K_{15} \frac{\eta}{\tau}$ for $t \geq 2c_2\tau$.*

Proof of Lemma A.9: For $t \geq 2c_2\tau$, we have $J_t \geq \sum_{s=1}^{2c_2\tau} (p_s - \bar{p}_t)^2 \geq \frac{c_2\tau}{2}(\tilde{p}_1 - \tilde{p}_2)^2 = k_1\tau$, where $k_1 = \frac{c_2}{2}(\tilde{p}_1 - \tilde{p}_2)^2$. By the mean value theorem, we have $|\varphi(\theta) - \varphi(\vartheta_t)| \leq \sqrt{2k_2}\|\theta - \vartheta_t\|$, where $k_2 = \max_{j \in \{1,2\}}\{\max_\theta\{(\partial\varphi(\theta)/\partial\theta_j)^2\}\}$. By monotonicity of expectation, we have

$$\begin{aligned} &\mathbb{E}(\varphi(\theta) - \varphi(\vartheta_t))^2 \\ \leq\ & 2k_2 \mathbb{E}\|\theta - \vartheta_t\|^2 \\ \leq\ & 2k_2 \mathbb{E}\|\theta - \hat{\theta}_t\|^2 \\ =\ & 2k_2 \int_0^\infty \mathbb{P}(\|\theta - \hat{\theta}_t\|^2 > x, J_t \geq k_1\tau) dx \\ \leq\ & \frac{4k_2\eta}{\lambda k_1 \tau} + 2k_2 \int_{\frac{2\eta}{\lambda k_1 \tau}}^\infty \gamma t \exp(-\lambda(\sqrt{x} \wedge x) k_1 \tau) dx \\ \leq\ & \frac{4k_2\eta}{\lambda k_1 \tau} + 2k_2 \left[\int_{\frac{2\eta}{\lambda k_1 \tau}}^1 \gamma t \exp(-\lambda x k_1 \tau) dx + \int_1^\infty \gamma t \exp(-\lambda \sqrt{x} k_1 \tau) dx \right] \\ =\ & \frac{4k_2\eta}{\lambda k_1 \tau} + 2k_2 \left[\frac{\gamma t \exp(-2\eta)}{\lambda k_1 \tau} + \frac{\gamma t \exp(-\lambda k_1 \tau)}{\lambda k_1 \tau} + \frac{2\gamma t \exp(-\lambda k_1 \tau)}{\lambda^2 k_1^2 \tau^2} \right] \\ \leq\ & \frac{4k_2\eta}{\lambda k_1 \tau} + 2k_2 \left[\frac{\gamma T \exp(-2\eta)}{\lambda k_1 \tau} + \frac{\gamma T \exp(-\lambda k_1 \tau)}{\lambda k_1 \tau} + \frac{2\gamma T \exp(-\lambda k_1 \tau)}{\lambda^2 k_1^2 \tau^2} \right]. \end{aligned}$$

The third inequality holds by Lemma A.8. Note that

$$\lim_{\rho \to 1} T \exp(-2\eta) = T^{-1} < \log T = \lim_{\rho \to 1} \eta \text{ for } T \geq 2.$$

$$\lim_{T\to\infty}\lim_{\rho\to 1} T\exp(-\lambda k_1\tau) = \lim_{T\to\infty} T\exp\left(-\lambda k_1\left[\sqrt{T}\right]\right) = 0 < \lim_{T\to\infty}\log T = \lim_{T\to\infty}\lim_{\rho\to 1}\eta,$$
$$\lim_{T\to\infty}\lim_{\rho\to 1} T\exp(-\lambda k_1\tau)/\tau = \lim_{T\to\infty} T\exp\left(-\lambda k_1\left[\sqrt{T}\right]\right)/\left(\left[\sqrt{T}\right]\right) = 0 < \lim_{T\to\infty}\log T = \lim_{T\to\infty}\lim_{\rho\to 1}\eta.$$

Thus, there exists $\check{T}\in\mathbb{N}$ and $\check{\rho}\in[0,1)$ such that for all $T\geq\check{T}$ and $\rho\geq\check{\rho}$, we have

$$\begin{aligned}
&\mathbb{E}(\varphi(\theta)-\varphi(\vartheta_t))^2\\
&\leq \frac{4k_2\eta}{\lambda k_1\tau} + 2k_2\left[\frac{\gamma T\exp(-2\eta)}{\lambda k_1\tau} + \frac{\gamma T\exp(-\lambda k_1\tau)}{\lambda k_1\tau} + \frac{2\gamma T\exp(-\lambda k_1\tau)}{\lambda^2 k_1^2\tau^2}\right]\\
&\leq \frac{4k_2\eta}{\lambda k_1\tau} + 2k_2\left[\frac{\gamma\eta}{\lambda k_1\tau} + \frac{\gamma\eta}{\lambda k_1\tau} + \frac{2\gamma\eta}{\lambda^2 k_1^2\tau}\right]\\
&\leq K_{15}\frac{\eta}{\tau},
\end{aligned}$$

where $K_{15} = \frac{4k_2}{\lambda k_1} + \frac{4k_2\gamma}{\lambda k_1} + \frac{4k_2\gamma}{\lambda^2 k_1^2}$. ■

Proof of Theorem 6: Under policy $\hat{\pi}$, we have

$$\sum_{t=2c_2\tau}^{T-1}\rho^t\mathbb{E}(\varphi(\theta)-p_{t+1})^2 = \sum_{t=2c_2\tau}^{T-1}\rho^t\mathbb{E}(\varphi(\theta)-\varphi(\vartheta_t))^2 \leq K_{15}\sum_{t=2c_2\tau}^{T-1}\rho^t\frac{\eta}{\tau} = K_{15}\frac{\eta}{\tau}\frac{\rho^{2c_2\tau}(1-\rho^{T-2c_2\tau})}{1-\rho}.$$

Then, we have

$$\begin{aligned}
\Delta^{\hat{\pi}}(T,\rho) &\leq |b_{\min}|\sup_{\theta\in\Theta}\left\{\sum_{t=0}^{2c_2\tau-1}\rho^t\mathbb{E}(p_{t+1}-\varphi(\theta))^2 + \sum_{t=2c_2\tau}^{T-1}\rho^t\mathbb{E}(p_{t+1}-\varphi(\theta))^2\right\}\\
&\leq |b_{\min}|(u-l)^2\frac{1-\rho^{2c_2\tau}}{1-\rho} + |b_{\min}|K_{15}\frac{\eta}{\tau}\frac{\rho^{2c_2\tau}(1-\rho^{T-2c_2\tau})}{1-\rho}\\
&= K_8\frac{1-\rho^{2c_2\tau}}{1-\rho} + K_9\frac{\eta}{\tau}\frac{\rho^{2c_2\tau}-\rho^T}{1-\rho},
\end{aligned}$$

where $K_8 = |b_{\min}|(u-l)^2$ and $K_9 = |b_{\min}|K_{15}$. When $\rho\to 1$, we have $\tau = \left[\sqrt{T}\right]$, $\eta = \log T$, and

$$\lim_{\rho\to 1}\Delta^{\hat{\pi}}(T,\rho) \leq 2c_2K_8\tau + K_9\frac{\eta}{\tau}(T-2c_2\tau) = \mathcal{O}(\sqrt{T}\log T).$$

When $T\to\infty$, we have $\tau = \left[\sqrt{\frac{1}{1-\rho}}\right]$, $\eta = \log\left(\frac{1}{1-\rho}\right)$, and

$$\lim_{T\to\infty}\Delta^{\hat{\pi}}(T,\rho) \leq K_8\frac{1-\rho^{2c_2\tau}}{1-\rho} + K_9\frac{\eta}{\tau}\frac{1}{1-\rho} = \mathcal{O}\left(\sqrt{\frac{1}{1-\rho}}\log\left(\frac{1}{1-\rho}\right)\right).$$

■

Proof of Lemma 1: Using Lemma A.8 and condition (i), Keskin and Zeevi (2014) show that, there exists a constant K_{12} such that

$$\mathbb{E}(\varphi(\theta)-p_{t+1})^2 \leq \frac{K_{12}\log t}{\lambda\kappa_0\sqrt{t}} + 2\mathbb{E}(\varphi(\vartheta_t)-p_{t+1})^2 \text{ for all } t\geq N,$$

where N is a constant which satisfies $kN\exp(-\frac{1}{2}\lambda\kappa_0\sqrt{N})\leq 1$. Using condition (ii), we have

$$\sum_{t=N}^{T-1}\mathbb{E}(\varphi(\theta)-p_{t+1})^2$$

$$\begin{aligned} &\leq K_{12}\sum_{t=N}^{T-1}\frac{\log t}{\lambda\kappa_0\sqrt{t}}+2\sum_{t=N}^{T-1}\mathbb{E}(\varphi(\vartheta_t)-p_{t+1})^2 \\ &\leq \frac{2K_{12}}{\lambda\kappa_0}\sqrt{T}\log T+2\kappa_1\sqrt{T}. \end{aligned}$$

Since $\beta \in [b_{\min}, b_{\max}]$, we have

$$\begin{aligned} \Delta^\pi(T,\rho) &= \sup_{\theta\in\Theta}\left\{-\beta\sum_{t=0}^{T-1}\rho^t\mathbb{E}(p_{t+1}-\varphi(\theta))^2\right\} \\ &\leq |b_{\min}|\sup_{\theta\in\Theta}\left\{\sum_{t=0}^{N-1}\mathbb{E}(p_{t+1}-\varphi(\theta))^2+\sum_{t=N}^{T-1}\mathbb{E}(p_{t+1}-\varphi(\theta))^2\right\} \\ &\leq |b_{\min}|\left\{N(u-l)^2+\frac{2K_{12}}{\lambda\kappa_0}\sqrt{T}\log T+2\kappa_1\sqrt{T}\right\} \\ &\leq K_{16}\sqrt{T}\log T, \end{aligned}$$

where $K_{16} = |b_{\min}|(N(u-l)^2+\frac{2K_{12}}{\lambda\kappa_0}+2\kappa_1)$. ■

Appendix C Proofs of the Results in Section 4

In this section, we consider the setting in Section 4 where the effective discount rate per decision period is $\rho(T) = (\rho_0)^{1/T}$ for $\rho_0 \in (0,1)$. We show that for the models in BR and KZ, the regret under any policy is $\Omega(\sqrt{T})$ (Propositions A.1 and A.4). For the model in BR, we show that the regret under our policy as well that under the MLE-CYCLE policy in BR is $\mathcal{O}(\sqrt{T})$ (Propositions A.2 and A.3). For the model of KZ, we show that the regret is $\mathcal{O}(\log T\sqrt{T})$ under three policies – namely, the two variants of the greedy Iterated-Least-Squares policy in KZ and a different policy that we propose (Propositions A.5 and A.6).

Proposition A.1 *Consider the problem class $\mathcal{C}_{LB}$ defined in Theorem 1. For any policy ψ and $\rho(T) = (\rho_0)^{1/T}$, there exists a parameter $z \in \mathcal{Z}$, such that*

$$\textit{Regret}(z, \mathcal{C}_{LB}, T, \rho(T), \psi) = \Omega(\sqrt{T}).$$

Proof of Proposition A.1: Recall from Theorem 1 that there exists a parameter $z \in \mathcal{Z}$ such that

$$\text{Regret}(z, \mathcal{C}_{LB}, T, \rho, \psi) \geq K_0\left(\sqrt{\frac{\rho}{1-\rho}}(1-\rho^{T-1})+\sqrt{\frac{\rho^T(1-\rho^{T-1})}{1-\rho}}\right).$$

For $T \geq 2$ and $\rho(T) = (\rho_0)^{1/T}$, we have

$$\begin{aligned} &\text{Regret}(z, \mathcal{C}_{LB}, T, \rho(T), \psi) \\ &\geq K_0\left(\sqrt{\frac{(\rho_0)^{1/T}}{1-(\rho_0)^{1/T}}}\left(1-(\rho_0)^{(1-1/T)}\right)+\sqrt{\frac{\rho_0\left(1-(\rho_0)^{(1-1/T)}\right)}{1-(\rho_0)^{1/T}}}\right) \end{aligned}$$

$$
\begin{aligned}
&\geq K_0 \left(\sqrt{(\rho_0)^{1/2}\left(1-(\rho_0)^{(1-1/2)}\right)} \sqrt{\frac{1}{1-(\rho_0)^{1/T}}} + \sqrt{\rho_0 \left(1-(\rho_0)^{(1-1/2)}\right)} \sqrt{\frac{1}{1-(\rho_0)^{1/T}}} \right) \\
&= K_0 \left(\sqrt{(\rho_0)^{1/2}\left(1-(\rho_0)^{(1-1/2)}\right)} + \sqrt{\rho_0 \left(1-(\rho_0)^{(1-1/2)}\right)} \right) \sqrt{\frac{1}{1-(\rho_0)^{1/T}}} \\
&= \Omega(\sqrt{T}).
\end{aligned}
$$

■

Proposition A.2 *For any problem class $\mathcal{C}$ satisfying Assumptions 1 and 2 with corresponding exploration prices $\bar{\boldsymbol{p}} \in \mathcal{P}^k$ and $\rho(T) = (\rho_0)^{1/T}$, our policy $\hat{\psi}$ (defined in Section 2.2) satisfies*

$$
\textit{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho(T), \hat{\psi}) = \mathcal{O}\left(\sqrt{T}\right).
$$

Proof of Proposition A.2: Recall from Theorem 2, our policy $\hat{\psi}$ satisfies

$$
\text{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho, \hat{\psi}) \leq K_1 \frac{1-\rho^{k\tau}}{1-\rho} + K_2 \frac{\rho^{k\tau}-\rho^T}{(1-\rho)\tau},
$$

where $\tau = \left[\sqrt{\frac{1-\rho^T}{1-\rho}}\right]$. For $\rho(T) = (\rho_0)^{1/T}$, we have

$$
\begin{aligned}
\text{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho(T), \hat{\psi}) &\leq K_1 \frac{1-(\rho_0)^{k\tau/T}}{1-(\rho_0)^{1/T}} + K_2 \frac{(\rho_0)^{k\tau/T} - \rho_0}{\left(1-(\rho_0)^{1/T}\right)\tau} \\
&= \mathcal{O}\left(\frac{1-(\rho_0)^{-2k\sqrt{1-\rho_0}\sqrt{1-(\rho_0)^{1/T}}/\log \rho_0}}{1-(\rho_0)^{1/T}} \right) + \mathcal{O}\left(\frac{1}{\sqrt{1-(\rho_0)^{1/T}}} \right) \\
&= \mathcal{O}\left(\frac{1}{\sqrt{1-(\rho_0)^{1/T}}} \right) = \mathcal{O}\left(\sqrt{T}\right).
\end{aligned}
$$

■

Proposition A.3 *For any problem class $\mathcal{C}$ satisfying Assumptions 1 and 2 with corresponding exploration prices $\bar{\boldsymbol{p}} \in \mathcal{P}^k$ and $\rho(T) = (\rho_0)^{1/T}$, the MLE-CYCLE policy $\check{\psi}$ in BR satisfies*

$$
\textit{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho(T), \check{\psi}) = \mathcal{O}\left(\sqrt{T}\right).
$$

Proof of Proposition A.3: Recall from Theorem 3, the MLE-CYCLE policy $\check{\psi}$ satisfies

$$
\text{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho, \hat{\psi}) \leq K_3 \left(1 + \sum_{h=1}^{\lfloor \sqrt{2T} \rfloor} \rho^{h^2/2} \right).
$$

For $\rho(T) = (\rho_0)^{1/T}$, we have

$$
\text{Regret}(\boldsymbol{z}, \mathcal{C}, T, \rho(T), \hat{\psi}) \leq K_3 + \sum_{h=1}^{\lfloor \sqrt{2T} \rfloor} (\rho_0)^{h^2/(2T)} K_3
$$

$$
\begin{aligned}
&\leq K_3 + K_3 \int_0^{\sqrt{2T}} (\rho_0)^{h^2/(2T)} dh \\
&= K_3 + K_3 \int_0^{\sqrt{2T}} e^{\log(\rho_0) h^2/(2T)} dh \\
&= K_3 + K_3 \sqrt{\frac{1}{2}} \int_0^{T} e^{\log(\rho_0) x/T} \sqrt{\frac{1}{x}} dx.
\end{aligned}
$$

The last equality holds by letting $x = h^2/2$. When $T \to \infty$,

$$
\begin{aligned}
\lim_{T\to\infty} K_3 \left(1 + \sqrt{\frac{1}{2}} \int_0^{T} e^{\log(\rho_0) x/T} \sqrt{\frac{1}{x}} dx\right) &= \lim_{T\to\infty} K_3 \left(1 + \sqrt{\frac{1}{2}} \sqrt{\frac{1}{-\log(\rho_0)}} \sqrt{T} \int_0^{-\log(\rho_0)} e^{-w} w^{-1/2} dw\right) \\
&\leq K_3 \left(1 + \sqrt{\frac{\pi}{2}} \sqrt{\frac{1}{-\log(\rho_0)}} \sqrt{T}\right) \\
&= \mathcal{O}\left(\sqrt{T}\right).
\end{aligned}
$$

The first equality holds by letting $w = -\log(\rho_0) x/T$. The inequality holds since

$$
\int_0^{-\log(\rho_0)} e^{-w} w^{-1/2} dw < \int_0^{\infty} e^{-w} w^{-1/2} dw = \Gamma(1/2) = \sqrt{\pi}.
$$

■

Next, we consider the model in KZ with discounting where the discount factor $\rho(T) = (\rho_0)^{1/T}$.

Proposition A.4 *For any policy π and $\rho(T) = (\rho_0)^{1/T}$, we have*

$$
\Delta^\pi(T, \rho(T)) = \Omega\left(\sqrt{T}\right).
$$

Proof of Proposition A.4: Recall from Theorem 4, we have

$$
\Delta^\pi(T, \rho) \geq K_4 \sqrt{\frac{\rho(1-\rho^{2T-2})}{1-\rho^2}} \quad \text{for any policy } \pi, \rho \in [0,1), \text{ and } T \geq 3.
$$

For $\rho(T) = (\rho_0)^{1/T}$, we have

$$
\begin{aligned}
&\Delta^\pi(T, \rho(T)) \\
&\geq K_4 \sqrt{\frac{(\rho_0)^{1/T}(1-(\rho_0)^{(2T-2)/T})}{1-(\rho_0)^{2/T}}} \\
&\geq K_4 \sqrt{\rho_0(1-\rho_0)} \sqrt{\frac{1}{1-(\rho_0)^{2/T}}} \\
&= \Omega(\sqrt{T}).
\end{aligned}
$$

■

Proposition A.5 *Let π be a pricing policy that satisfies the conditions in Theorem 5. Then, we have*

$$
\Delta^\pi(T, \rho(T)) = \mathcal{O}\left(\sqrt{T} \log T\right).
$$

Proof of Proposition A.5: Recall from Theorem 5 that the regret under policy π satisfies

$$\Delta^{\pi}(T,\rho) \leq K_5 \frac{1-\rho^N}{1-\rho} + K_6 N \log N + K_7 \frac{\log N}{N} \frac{1-\rho^{T-N^2}}{1-\rho}, \text{ for } N = \left\lfloor \sqrt{\frac{1-\rho^T}{1-\rho}} \right\rfloor .$$

For $\rho(T) = \rho_0^{1/T}$, we have

$$\begin{aligned}
&\Delta^{\pi}(T,\rho(T)) \\
\leq\; & K_5 \frac{1-(\rho_0)^{N/T}}{1-(\rho_0)^{1/T}} + K_6 N \log N + K_7 \frac{\log N}{N} \frac{1-(\rho_0)^{(T-N^2)/T}}{1-(\rho_0)^{1/T}} \\
=\; & \mathcal{O}\left(\frac{1-(\rho_0)^{-2\sqrt{1-\rho_0}\sqrt{1-(\rho_0)^{1/T}}/\log \rho_0}}{1-(\rho_0)^{1/T}} \right) + \mathcal{O}\left(\frac{1}{\sqrt{1-(\rho_0)^{1/T}}} \log\left(\frac{1}{1-(\rho_0)^{1/T}} \right) \right) + \\
& \mathcal{O}\left(\frac{1}{\sqrt{1-(\rho_0)^{1/T}}} \log\left(\frac{1}{1-(\rho_0)^{1/T}} \right) \right) \\
=\; & \mathcal{O}\left(\frac{1}{\sqrt{1-(\rho_0)^{1/T}}} \log\left(\frac{1}{1-(\rho_0)^{1/T}} \right) \right) = \mathcal{O}\left(\log T \sqrt{T} \right).
\end{aligned}$$

■

Proposition A.6 *Our policy $\hat{\pi}$ in Section 3.2.3 satisfies*

$$\Delta^{\hat{\pi}}(T,\rho(T)) = \mathcal{O}\left(\sqrt{T} \log T \right).$$

Proof of Proposition A.6: Recall from Theorem 6 that our policy $\hat{\pi}$ satisfies

$$\Delta^{\hat{\pi}}(T,\rho) \leq K_8 \frac{1-\rho^{2c_2\tau}}{1-\rho} + K_9 \frac{\eta}{\tau} \frac{\rho^{2c_2\tau}-\rho^T}{1-\rho},$$

where $\tau = \left\lceil \sqrt{\frac{1-\rho^T}{1-\rho}} \right\rceil$ and $\eta = \log\left(\frac{1-\rho^T}{1-\rho}\right)$. For $\rho(T) = (\rho_0)^{1/T}$, we have

$$\begin{aligned}
&\Delta^{\hat{\pi}}(T,\rho(T)) \\
\leq\; & K_8 \frac{1-(\rho_0)^{2c_2\tau/T}}{1-(\rho_0)^{1/T}} + K_9 \frac{\eta}{\tau} \frac{(\rho_0)^{2c_2\tau/T}-\rho_0}{1-(\rho_0)^{1/T}} \\
=\; & O\left(\frac{1-(\rho_0)^{-4c_2\sqrt{1-\rho_0}\sqrt{1-(\rho_0)^{1/T}}/\log \rho_0}}{1-(\rho_0)^{1/T}} \right) + \mathcal{O}\left(\frac{1}{\sqrt{1-(\rho_0)^{1/T}}} \log\left(\frac{1-\rho_0}{1-(\rho_0)^{1/T}} \right) \right) \\
=\; & \mathcal{O}\left(\frac{1}{\sqrt{1-(\rho_0)^{1/T}}} \log\left(\frac{1}{1-(\rho_0)^{1/T}} \right) \right) = \mathcal{O}\left(\log T \sqrt{T} \right).
\end{aligned}$$

■

Appendix D Number of Exploration Periods as a Function of the Discount Factor

Table A.1 below shows the number of exploration periods as a function of the discount factor ρ for our policies, for policy MLE-CYCLE in BR, and for policy ILS-d in KZ.

Table A.1: The number of exploration periods as a function of ρ under our policy, policy MLE-CYCLE in BR, and policy ILS-d in KZ, for $T = 40,000$.

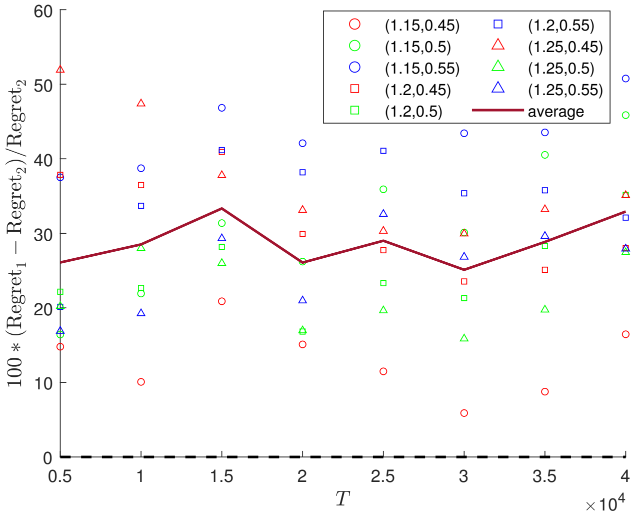
$\log_{10}\left(\frac{1}{1-\rho}\right)$	ρ	our policy	MLE-CYCLE	ILS-d
1	0.9	6	562	399
2	0.99	20	562	399
3	0.999	64	562	399
4	0.9999	198	562	399
5	0.99999	364	562	399
6	0.999999	396	562	399

We do not report the number of exploration periods for CILS in KZ because there is no clear boundary between exploration and exploitation under the CILS policy, thus making it difficult to determine the exact length of exploration. We now elaborate. Recall that under the CILS policy, in each period t, we first compute the difference between the greedy ILS price $\varphi(\vartheta_{t-1})$ in period t and the average price $\bar{p}_{t-1}$ in the first $t-1$ periods, denoted by $\delta_t = \varphi(\vartheta_{t-1}) - \bar{p}_{t-1}$. Then, for a positive constant c_1, the CILS policy charges $\bar{p}_{t-1} + \text{sgn}(\delta_t)c_1 t^{-1/4}$ if $|\delta_t| < c_1 t^{-1/4}$ and $\varphi(\vartheta_{t-1})$ otherwise. When $|\delta_t| \geq c_1 t^{-1/4}$, the CILS policy uses the greedy price $\varphi(\vartheta_{t-1})$ for exploitation. However, when $|\delta_t| < c_1 t^{-1/4}$, it is unclear whether the price $\bar{p}_{t-1} + \text{sgn}(\delta_t)c_1 t^{-1/4}$ is used purely exploration or exploitation. On the one hand, we exploit the average price $\bar{p}_{t-1}$, which is dynamically updated as time t increases and is sufficiently close to the greedy price when t is large. On the other hand, while we use a price deviation ($\text{sgn}(\delta_t)c_1 t^{-1/4}$) from the average price $\bar{p}_{t-1}$ for exploration, this deviation decreases with time t and is relatively small, so that the deviation from the greedy or "exploitation" price is not too much. That is, the CILS policy focuses more on exploitation and less on exploration as time t increases. When t is sufficiently large, the offered price can be very close to the greedy price. Therefore, there is no clear or simple answer to whether the price is used for exploration or exploitation. Alternatively, one can say that the price is used for both exploration and exploitation, and balances the tradeoff between the two. Since there is no clear definition for exploration periods in CILS, we do not report that number for the CILS policy in Table A.1.

Appendix E Additional Numerical Experience

We show the robustness of the superior performance of our policy when ρ is sufficient close to 1. In particular, for $\rho = 0.999999$, we numerically examine the behavior of the regret under different policies with respect to the time horizon T by varying T from 5000 to 40000, in increments of 5000, in the settings of both BR (see Section 2.3) and KZ (see Section 3.3). Figure A.1 (resp., Figure A.2) plots the relative difference between the average regret under policy MLE-CYCLE in BR and that under our policy for the linear (resp., logit) model.

Figure A.1: The relative percentage difference between the average regret under policy MLE-CYCLE and that under our policy for a linear model ($\rho = 0.999999$).



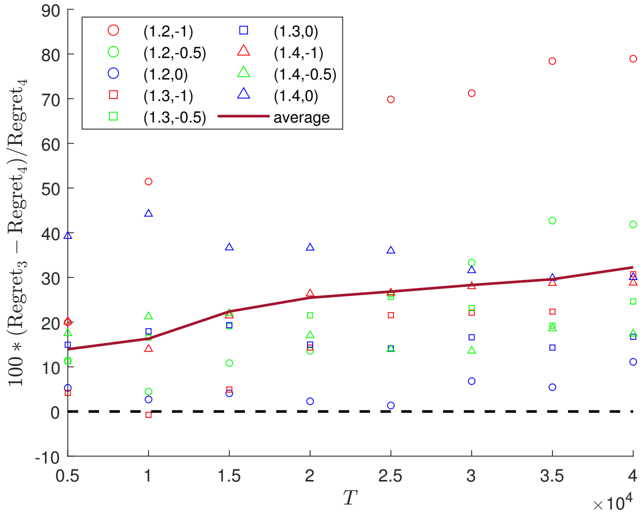
Note: Regret_1 is the average regret under MLE-CYCLE and Regret_2 is the average regret under our policy.

Figure A.3 (resp., Figure A.4) plots the relative difference between the average regret under policy ILS-d (resp., CILS) in KZ and that under our policy. We also provide a companion table (Table A.2) to Figures A.1, A.2, and A.3 to show the number of exploration periods as a function of T for our policy, for policy MLE-CYCLE in BR, and for policy ILS-d in KZ. As seen in Table A.2, the exploration length of our policy is similar to that of ILS-d. As seen in Figures A.1, A.2, A.3, and A.4, our policy consistently performs better when ρ is sufficient close to 1.

Table A.2: The number of exploration periods as a function of T under our policy, policy MLE-CYCLE in BR, and policy ILS-d in KZ, for $\rho = 0.999999$.

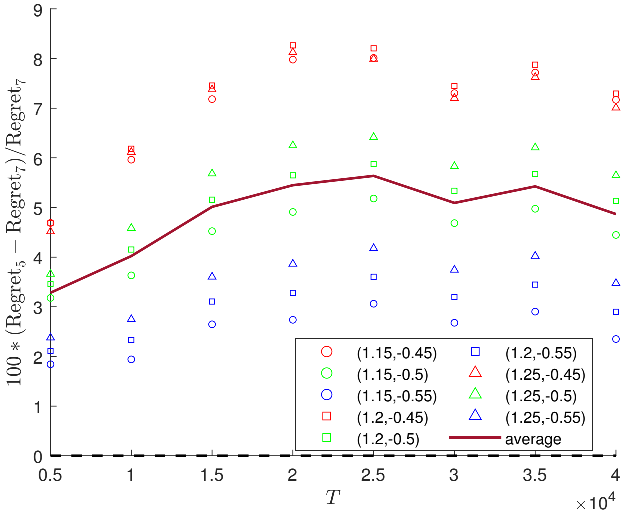
T	our policy	MLE-CYCLE	ILS-d
5000	142	196	140
10000	200	278	199
15000	244	342	244
20000	282	396	282
25000	314	444	316
30000	344	486	346
35000	370	526	374
40000	396	562	399

Figure A.2: The relative percentage difference between the average regret under policy MLE-CYCLE and that under our policy for a logit model ($\rho = 0.999999$).



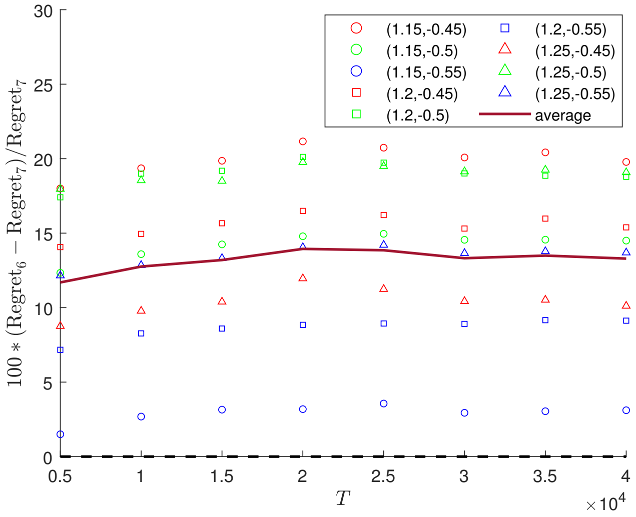
Note: Regret$_3$ is the average regret under MLE-CYCLE and Regret$_4$ is the average regret under our policy.

Figure A.3: The relative percentage difference between the average regret under policy ILS-d and that under our policy ($\rho = 0.999999$).



Note: Regret$_5$ is the average regret under ILS-d and Regret$_7$ is the average regret under our policy.

Figure A.4: The relative percentage difference between the average regret under policy CILS and that under our policy ($\rho = 0.999999$).



Note: Regret_6 is the average regret under CILS and Regret_7 is the average regret under our policy.