POMS PRODUCTION AND OPERATIONS MANAGEMENT

PRODUCTION AND OPERATIONS MANAGEMENT

Vol. 30, No. 8, August 2021, pp. 2379–2394
ISSN 1059-1478 | EISSN 1937-5956 | 21 | 3008 | 2379
DOI 10.1111/poms.13380

Earning and Learning with Varying Cost

Ying Zhong
School of Management and Economics, University of Electronic Science and Technology of China, Chengdu, 611731, China, yzhong4@uestc.edu.cn

L. Jeff Hong*
School of Management and School of Data Science, Fudan University, Shanghai, 200433, China, hong_liu@fudan.edu.cn

Guangwu Liu
Department of Management Sciences, College of Business, City University of Hong Kong, Kowloon, Hong Kong China, guanliu@cityu.edu.hk

We study a dynamic pricing problem where the observed cost in each selling period varies from period to period, and the demand function is unknown and only depends on the price. The decision maker needs to select a price from a menu of K prices in each period to maximize the expected cumulative profit. Motivated by the classical upper confidence bound (UCB) algorithm for the multi-armed bandit problem, we propose a UCB-Like policy to select the price. When the cost is a continuous random variable, as the cost varies, the profit of the optimal price can be arbitrarily close to that of the second-best price, making it very difficult to make the correct decision. In this situation, we show that the expected cumulative regret of our policy grows in the order of $(\log T)^2$, where T is the number of selling periods. When the cost takes discrete values from a finite set and all prices are optimal for some costs, we show that the expected cumulative regret is upper bounded by a constant for any T. This result suggests that in this situation, the suboptimal price will only be selected in a finite number of periods, and the trade-off between earning and learning vanishes and learning is no longer necessary beyond a certain period.

Key words: earning and learning; varing cost; multi-armed bandit; upper confidence bound; regret analysis
History: Received: September 2018; Accepted: January 2021 by Dan Zhang, after 3 revisions.
*Corresponding author.

1. Introduction

Pricing is often a very important decision for firms that sell products or provide services. A major difficulty in making pricing decisions is that the demand changes with respect to the price, and the demand function (of price) is often unknown. Therefore, besides *earning* short-term profits, pricing decisions often involve some type of *learning* mechanism that changes price over different periods (or for different customers) to learn the uncertain customer demands in order to increase long-term profits. This type of dynamic pricing problem is also known as the earning-and-learning problem to emphasize the key trade-off between earning and learning. In the classical literature on this problem, one often assumes that the cost of the product or service is constant and mostly, without loss of generality, zero. Therefore, there is often a single optimal price that one needs to learn, if the demand function does not change over time.

In many practical situations, however, the cost may change from period to period or customer to customer. We consider three examples.

- The first example is the sales of fresh products, for example, fruits and vegetables, in supermarkets. While customer demands may depend solely on the price, a supermarket's profit clearly depends on the costs of these products which may change from day to day. As reported by McLaughlin and Perosio (1994), while the wholesaler may guarantee that the wholesale prices charged will not exceed certain levels (referred to as "lid" prices) perhaps 2–4 weeks before the scheduled date of delivery, the exact wholesale prices (and thus the costs to the supermarket) will not be specified until the date of shipment depending on the market changes and prevailing prices in the market at the time. Therefore, the supermarket may have the incentive to change the price from day to day, not only to learn the demands, but also to react to the changes of the costs. In this example, the costs of the products may vary continuously.
- The second example is the sales of foreign products, for example, French wine in American wine stores. In this example, the costs of the products are significantly tied to the volatile exchange rates and thus causing the prices to fluctuate as well.

- The third example is the sales of health insurance products. In practice, an insurance company may use the results from health examinations to classify customers into different risk groups, and the customers from one risk group may share the same demand function. However, the cost of the product within the group may not be a constant. In the United States, the cost of an insurance product can vary from one state to another due to different regulations or because the same regulation may impact differently in different states. For instance, the Affordable Care Act in the United States reduces the costs of insurance products in some states but raises the costs in others because of technological issues or political reasons (Keskin and Birge 2019). Therefore, even for the same customer, the cost of an insurance health product can be different as s/he buys the product in different states. The insurance company may offer the customer different prices due to the difference in cost. In this example, the cost of the product is not entirely known. However, additional information (i.e., the results of health examinations and geographic locations) allows the company to have a more accurate estimate of the conditional expected cost of a customer.

In all three examples, the cost of the product varies from period to period or customer to customer. The dynamic pricing problem becomes more challenging because the optimal price is no longer a constant and it becomes a function of the cost (or the conditional expected cost). One has to learn the optimal price function of the cost in order to maximize the long-term expected profits.

To analyze how varying costs affect the earning-and-learning trade-off, we consider a firm that sells a single product over T periods or customers, where T is unknown. In each period, it first observes the realized cost (or the conditional expected cost) of the product and then selects the price from a menu of K prices1 based on all of the accumulated information up to the period. The demand of each period is a random variable whose mean is unknown and depends solely on the price of the product, and the demands of different periods are independent of each other. The firm's objective is to maximize its expected cumulative profit over all T periods, or equivalently, to minimize its expected cumulative regret from the true optimal expected profit.

It is easy to see that, if one sets the cost as a constant (e.g., zero, without loss of generality), our problem becomes the classical formulation of the earning-and-learning problem considered by Rothschild (1974). It is an example of a multi-armed bandit (MAB) problem, which has been studied extensively in the literature of statistics and machine learning since the classical paper of Robbins (1952). A very important theoretical result of the MAB problem, developed by Lai and Robbins (1985), is that the expected cumulative regret increases at least in the order of $\log T$ no matter what algorithm is used. This asymptotic lower bound is achieved by the upper confidence bound (UCB) algorithm of Auer et al. (2002) among others. Notice that, when the cost is a constant, the identity of the optimal price is fixed. The UCB algorithm spends only $\log T$ order of periods exploring non-optimal prices and the rest on the optimal price. When the cost varies over periods, however, the classical UCB algorithm is not applicable, mainly due to the fact that the identity of the optimal price becomes a function of the cost. In this case, in order to choose the optimal price, one needs to learn the profit functions of different prices with respect to the cost. Earning and learning with varying costs is thus much more challenging, compared to the standard MAB problem.

However, as a distinct feature of our problem, we find that without making additional assumptions, there exists a linear relationship between the profit and the cost, that is,

$$\text{Profit} = \text{Demand} \times (\text{Price} - \text{Cost}) = \text{Revenue} \times \left(1 - \frac{\text{Cost}}{\text{Price}}\right). \tag{1}$$

This structure suggests that learning the profit function (of the cost) boils down to learning the uncertain revenue, which depends only on the price choice and not on the cost. Therefore, the revenue learned at one cost is also meaningful for other costs. With this special structure, we propose a new algorithm based on the UCB algorithm. In particularly, in each period, we compute the UCBs of all the revenues, and use them to construct the UCBs of all the profit functions.

To analyze the expected cumulative regret of our proposed algorithm, we consider two cases of the cost, continuous and discrete. When the cost is a continuous random variable, for example, in the example of selling fresh products or foreign products discussed earlier in this section, as the cost varies, the profit of the optimal price can be arbitrarily close to that of the second-best price, making it very difficult to make the correct decision. We show that the regret of our algorithm is of the order $(\log T)^2$. When the cost is a discrete random variable, for example, in the example of selling health insurance products where customers are in different states, we prove that the expected cumulative regret of our algorithm is *bounded above by a constant*, if all prices are optimal for

some realized costs. From a theoretical point of view, a constant bound is the lowest order of the regret that can be achieved. From a managerial point of view, this result indicates that the suboptimal price will only be selected in a finite number of periods, and that learning is no longer necessary beyond a certain period. To understand why we have such an interesting result, we want to notice that there is a non-zero minimal gap Δ between the best and all others for all possible costs when the cost is discrete. This is similar to the standard MAB problem, and this minimal gap makes the learning of optimal prices easier and leads to a bounded regret.

The rest of the article is organized as follows. We conduct a thorough literature review and summarize our contributions in section 2. We then formulate the problem and propose a UCB-Like policy in section 3, followed by its regret analysis in section 4. Numerical study is provided in section 5, followed by concluding remarks in section 6. Lengthy proofs are postponed to the appendix.

2. Literature Review and Contributions

Our study is related to two lines of literature. One is dynamic pricing with demand learning and the other is the MAB problem.

2.1. Dynamic Pricing with Demand Learning

Dynamic pricing is a very important issue in revenue management. Gallego and Van Ryzin (1994) first consider a dynamic pricing problem with inventory constraint when demand is price sensitive and stochastic. Since then, there has been a stream of research work focusing on dynamic pricing with unknown demand function that needs to be learned. For such problems, the critical issue is to balance the trade-off between earning and learning. A myopic pricing policy without learning mechanism may permanently get stuck at an uninformative choice which provides no quality improvement on the underlying reward functions, leading to poor performance. This is also called the *incomplete learning* phenomenon, see Keskin and Zeevi (2018) for a comprehensive summary about the situations where the phenomenon may appear and the situations where the myopic algorithm can be directly applied without learning.

Most of the work in this area assumes the price is a continuous decision variable, and there is a parametric model of the demand function, often linear in price. In such contexts, Lobo and Boyd (2003) show by numerical experiments that for a linear demand model with Gaussian noise, one would benefit from *price-dithering* in a myopic pricing policy and a convex approximation has been used in that study to get a better solution. Later, both den Boer and Zwart (2014) and Keskin and Zeevi (2014) prove theoretically that myopic pricing policy can lead to incomplete learning and converge to sub-optimal prices.

To solve this problem, Broder and Rusmevichientong (2012) propose the MLE-CYCLE algorithm and den Boer and Zwart (2014) consider a taboo interval, respectively. They show that the expected cumulative regrets of their algorithms are upper bounded by $\mathcal{O}(\sqrt{T})$ and $\mathcal{O}(\sqrt{T}\log T)$, respectively. Keskin and Zeevi (2014) examine a special case where an *incumbent price* at which the decision maker knows the exact expected demand exists. The incumbent price reduces the model's number of unknown parameters. As long as the price does not converge to the incumbent price the incomplete learning in myopic pricing can be fixed and the resulted regret is upper bounded by $\mathcal{O}(\log T)$. Qiang and Bayati (2016) and Ban and Keskin (2017) extend the results to a data-rich environment where the price and other factors interactively affect the demand function and thus the optimal price. Notice that being different from these two studies, we consider a situation where there are idiosyncratic shocks driving the cost to change from period to period and thus the optimal price. In our setting, the cost does not affect the expected demands at different price levels. Nonparametric models of demand functions have also been considered in the literature. For instance, Besbes and Zeevi (2015) show that, by using a linear approximation of the demand function, the nonparametric approach can be almost as good as a parametric one, obtaining a regret bound of $\mathcal{O}(\sqrt{T}(\log T)^2)$. Under the Bayesian framework, the incomplete learning phenomenon still exists. Particularly, Harrison et al. (2012) demonstrate several instances that a myopic Bayesian policy may lead to incomplete learning and propose modifications on the myopic Bayesian policy to avoid such phenomenon.

Another stream of research related to our work is dynamic pricing in non-stationary environments. Besbes and Zeevi (2011) and Keskin and Zeevi (2016) study a setting with the parametric model for the demand function that changes over time. Therefore, besides learning the unknown demand function, one also needs to detect the time points where it changes. Our setting differs from theirs in that the change of the optimal price is caused by the observable varying cost that does not affect the demand function and no parametric demand model is assumed in our work. In the context of non-stationary cost, Keskin and Birge (2019) consider a problem where the product has different quality levels, and the cost is stochastic in nature at each quality level. To maximize the expected profit, the firm needs to learn the unknown functional form between the expected cost and quality level and then determine the best product offering scheme that

assigns different price–quality pairs to customers with different quality sensitivities. In their work, the cost is realized after the decision is made in each period and mainly used to learn the unknown cost function (of the quality), and thus the optimal price is invariant with respect to the cost and time. In our setting, however, the cost is observed before the pricing decision is made and the optimal price is changing with the observed cost.

Dynamic pricing with discrete price choices has also been studied in the literature, see for instance, Gallego and Van Ryzin (1994) and Feng and Xiao (2000). Rothschild (1974) was the pioneer of dynamic pricing with demand learning. The author considers two price choices and models the problem as a two-armed bandit problem. By employing a discount factor on the rewards, the author carries out the analysis over an infinite time horizon and shows that a decision maker pursuing optimal policies cannot be guaranteed to obtain the full information about the demand. In this study, we exclude the discount factor in the problem formulation and focus on minimizing the expected cumulative regret.

2.2. The MAB Problem

Since the seminar study of Robbins (1952) rigorously formulates the problem, the MAB problem has long attracted attentions from multiple disciplines, for example, operations research, computer science, economics, and statistics. A very important theoretical result of the MAB problem, developed by Lai and Robbins (1985), is that the expected cumulative regret increases at least in the order of $\log T$ no matter what algorithm is used, when the expected rewards of all choices are bounded. Numerous algorithms have been proposed to solve the problem, see Bubeck and Cesa-Bianchi (2012) for a recent review. Among these algorithms, the UCB algorithm of Auer et al. (2002) is probably the most famous one. In each period, the algorithm selects the choice with the highest UCB of the mean reward, constructed based on Hoeffding's Inequality. Auer et al. (2002) prove that the expected cumulative regret increases in the order of $\log T$, making it optimal in terms of the asymptotic order.

In many situations, decision makers may have some additional information to help them make choices in each period. The MAB with covariates is a way to incorporate this additional information (i.e., covariates). It is also called bandits with side observations or contextual bandits. In such a context, the expected reward of an arm depends on the observable covariates and can change with the covariates. This type of problems was first studied by Woodroofe (1979) and Sarkar (1991) for the so-called one-armed bandit problem. To simplify the analysis and improve the growth rate of the expected cumulative regret, many papers assume linear relationships between the expected rewards and the covariates of the arms, that is, linear bandits. Then, it is necessary to learn the unknown coefficients of the linear functions. Several notable works studying linear bandit algorithms include Mersereau et al. (2009), Rusmevichientong and Tsitsiklis (2010), Goldenshluger and Zeevi (2013), Bastani and Bayati (2020), and Bastani et al. (2020). The expected cumulative regrets of some of these algorithms can grow in the order of $\log T$. However, to attain the $\log T$ growth rate, many existing linear bandit algorithms adopt a forced sampling mechanism. They predetermine a sequence of time points and forced-sample arms regardless of the observed covariates at these points to guarantee enough learning on each arm. The sequence of times points is determined by input parameters (also known as tuning parameters, see for example, Gabillon et al. (2011) and Chen et al. (2019)), which assume one's foreknowledge of the linear functions. In practice, correct values for the input parameters are often unknown. Setting inadequate values for the input parameters may cause the algorithms to either spend too much sampling effort on learning each arm which can lead to poor finite-time performances or run without any statistical guarantees. Even though the recent work of Bastani et al. (2020) shows that, with some modifications, a forced sampling algorithm may achieve better performance and is more robust to the change of input parameters when there are no uniformly inferior arms, this improvement may not be extended to other cases.

As previously mentioned, in Equation (1), there exists a linear relationship between the profit and the cost. It suggests that our problem may be alternatively viewed as a linear bandit problem with one covariate, that is, the cost. Therefore, some existing linear bandit algorithms may be applied to solve our problem. Compared to these linear bandit algorithms, because of the special linear structure that appears in our problem, it allows us to develop an algorithm without using the forced sampling mechanism. Therefore, as a distinct feature of our algorithm, it does not require tuning parameters as many linear bandit algorithms do.

2.3. Our Contributions

Compared to the literature, our paper makes three contributions on the problem of dynamic pricing with unknown demand and varying costs.

- First, based on our problem formulation, we identify that the problem has an inherent linear structure with a unique feature that only the unknown revenue needs to be learned, which does not depend on the cost covariate. We want to emphasize that this important problem structure is derived solely from the dynamic

pricing problem and it is quite unique compared to typical linear bandit problems.

- Second, based on the unique problem structure, we develop a UCB-Like pricing policy that is different from typical linear bandit algorithms in two ways: (1) it does not need a tuning parameter that is often required in general-purpose linear bandit algorithms such as forced sampling algorithms and (2) its regret increases in a different order compared to forced sampling algorithms. Numerical results show that the UCB-Like policy may significantly outperform the forced sampling algorithms.
- Third, when the cost covariate is discrete, the regret of the UCB-Like pricing policy is upper bounded by a constant, under certain conditions. This is an interesting result because it shows that the traditional trade-off between earning and learning vanishes in this problem. While similar results have been shown in one-arm bandit literature (Goldenshluger and Zeevi 2009), it is new and has significant implications in dynamic pricing literature.

3. Problem Formulation and Pricing Policy

Consider a firm that sells a single product over a time horizon of T periods which is unknown to the decision maker. At the beginning of each period $t \in \{1, 2, \ldots, T\}$, the decision maker of the firm first observes a cost of the product $c_t \geq 0$, and the cost does not change within the period. Let $\mathscr{F}_{t-1}$ denote the filtration generated by random price choices, demands, and costs up to the time point immediately before observing the cost c_t, that is, $\mathscr{F}_{t-1} = \sigma(c_1, p^1, D^1, \ldots, c_{t-1}, p^{t-1}, D^{t-1})$, where p^s and D^s are the price choice and the corresponding demand in period s respectively. Then, we make the following assumption on the cost.

ASSUMPTION 1. *We assume that c_t has a common known support $\mathbb{C}$ for all $t = 1, 2, \ldots, T$. Furthermore, depending on whether $\mathbb{C}$ is continuous or discrete, there exists a positive constant $\underline{F} > 0$ or $\underline{F}^* > 0$ such that*

$$f(c|\mathscr{F}_{t-1}) \geq \underline{F} \text{ or } \mathbb{P}(c_t = c|\mathscr{F}_{t-1}) \geq \underline{F}^* \quad \forall c \in \mathbb{C}, \quad t = 1, 2, \ldots, T,$$

respectively, where $f(c|\mathscr{F}_{t-1})$ is the probability density function for $c_t|\mathscr{F}_{t-1}$.

REMARK 1. Assumption 1 is a fairly weak assumption. Besides the common known support, it only requires c_t to have a lower bound on the conditional density if c_t is continuous and a lower bound on the conditional probability if c_t is discrete. It allows c_t, for $t = 1, 2, \ldots, T$, to be dependent. For the three examples considered in the introduction, it is easy to see that the costs are dependent in the examples of fresh and foreign products and may be independent in the example of health insurance. We assume that Assumption 1 always holds in this study.

Upon observing the cost c_t, the decision maker then chooses a selling price from a menu of possible prices, denoted by $\{p_1, p_2, \ldots, p_K\}$, where $0 < p_1 < p_2 < \cdots < p_K < \infty$. We use $\mathbb{K} = \{1, 2, \ldots, K\}$ to index all the possible selling prices. After choosing a price p_k, where $k \in \mathbb{K}$, in period t, the firm observes the random demand $D_{k,t}$, which is given by

$$D_{k,t} = d(p_k) + \varepsilon_{k,t},$$

where $d(p_k)$ is the expected demand of price p_k, and $\epsilon_{k,t}$'s are unobserved independent and identically distributed (i.i.d.) noise terms with mean zero and a finite variance. In our paper, $d(p_k)$ is allowed to take a general form as a function of p_k, that is, no parametric relationship is assumed. In our setting, expected demand at each price is learned independently, where we ignore the possibility that adjacent prices may have similar expected demands. This setting is appropriate for contexts where the number of selling prices, that is, K, is not too large. This is often the case in many practical pricing applications with discrete prices, as a firm often changes prices by effecting markdowns and there is often a limited number of markdown prices. For instance, Caro and Gallien (2012) show that during the clearance selling, the fashion retailer Zara only chooses prices from a limited number of prices because it is easy to control and implement.

In this study, we assume that $D_{k,t}$ has a bounded support $[l, u]$ for all $k \in \mathbb{K}$ and $t = 1, 2, \ldots, T$, and the observed demand can always be fulfilled by the firm, that is, there is no supply shortage. The revenue of a period depends on both the selling price p_k and the corresponding demand, defined by $\Pi_{k,t} = p_{kDk,t}$. Without loss of generality, we assume throughout the paper that $\Pi_{k,t}$ has a bounded support in [0, 1] for all $k \in \mathbb{K}$ and $t = 1, 2, \ldots, T$. This assumption is equivalent to working with a normalized version of the revenue $[\Pi_{k,t} - lp_1]/(up_K - lp_1)$ that will not affect the analysis.

We let $\pi_k = \mathbb{E}[\Pi_{k,t}]$ denote the expected revenue associated with price p_k. Notice that $\mathbb{E}[\Pi_{k,t}] = p_k d(p_k)$. Then, the expected profit function $\mu_k(c)$ associated with the price p_k for a given cost c can be written as

$$\mu_k(c) = d(p_k)(p_k - c) = \pi_k\left(1 - \frac{c}{p_k}\right), \quad \forall k \in \mathbb{K}.$$

Notice that the profit for different prices depend on the observed cost c in each period. As a consequence, the optimal price to be set may vary from period to period, depending on the observed outcome of c in each period. The objective of the decision maker is to find a pricing policy such that the expected cumulative profit over the course of selling can be maximized. Ideally, if the index of the optimal price

$$i^*(c) = \arg\max_{k\in\mathbb{K}} \mu_k(c),$$

is known for different values of c, the optimal price in each period can then be identified, and the expected profit can be maximized. Compared to the classical MAB setting where the cost c is assumed to be a fixed constant (usually zero), taking into account of the effect of varying costs may lead to a better decision. Specifically, in the classical MAB setting, the cost in each period is assumed to be a constant, that is, the average cost $\mathbb{E}[c_t]$, the maximum achievable expected profit in each period is $\max_{k\in\mathbb{K}} \mu_k(\mathbb{E}[c_t])$, while the maximum achievable expected profit in the setting of varying cost is $\mathbb{E}[\max_{k\in\mathbb{K}} \mu_k(c_t)]$. As $\max_{k\in\mathbb{K}} \mu_k(c)$ is convex in c, Jensen's inequality implies that

$$\max_{k\in\mathbb{K}} \mu_k(\mathbb{E}[c_t]) \le \mathbb{E}[\max_{k\in\mathbb{K}} \mu_k(c_t)].$$

Therefore, accounting for the varying cost may lead to higher expected profit.

To achieve the maximum expected profit, the decision maker seeks to find an optimal pricing policy and learn the optimal price choice $i^*(c)$. Following the standard notion in the MAB literature, maximizing expected cumulative profit is equivalent to minimizing expected cumulative regret that is defined by

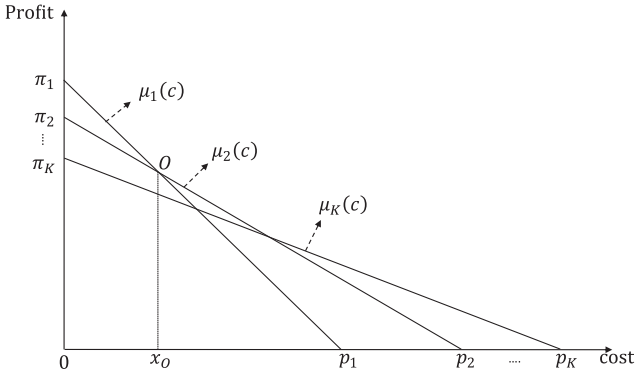
$$\mathbf{R}(T) = \sum_{t=1}^{T} \mathbb{E}[\mu_{i^*(c_t)}(c_t) - \mu_{I_t(c_t)}(c_t)], \tag{2}$$

where the function $I_t(c)$ denotes the index of the price chosen by a pricing policy upon observing a cost c in period t, and the expectation is taken with respective to the randomness in both c_t and $I_t(c_t)$.

To devise a good pricing policy to minimize the expected cumulative regret, a major difficulty is that profit functions $\{\mu_k(c), k\in\mathbb{K}\}$ are unknown and have to be learned by conducting price experiments. On the other hand, price experimentations on suboptimal prices lead to positive expected regrets. Therefore, a good pricing policy needs to balance the trade-off between earning and learning.

Before proposing a pricing policy, let us first investigate in greater detail the structure of the profit functions. As an example, we consider a possible configuration of the profit functions as in Figure 1. It could be observed that, for every $k\in\mathbb{K}$, the profit function $\mu_k(c)$ is a straight line that always passes through a fixed point $(p_k, 0)$ and intersects with y-axis at point $(0, \pi_k)$. As $(p_k, 0)$ is known, learning the profit function is equivalent to learning the expected revenue π_k that is independent of the cost c. This is a very critical observation, which basically says that accurate estimation of the *scalar* π_k leads to accurate estimation of the profit function $\mu_k(c)$.

Figure 1 Configuration of Profit Functions



Based on this observation, we focus on learning of the expected revenue π_k, which is similar to the objective of the classical MAB algorithms. This motivates us to develop a pricing policy based on the well-known UCB algorithm for the classical MAB problem. We briefly review the UCB algorithm as follows. In the classical MAB problem, the cost c is assumed to be a constant (and set to be zero without loss of generality). Then, the objective of the decision maker is to choose in each period the price that gives the maximum revenue. In the UCB algorithm, let $T_k(t-1)$ denote the number of times price p_k has been chosen up to period $t-1$. Then, at the beginning of period t, the expected revenue π_k can be estimated by

$$\bar{\pi}_{k,t} = \frac{1}{T_k(t-1)} \sum_{i=1}^{T_k(t-1)} \Pi_k^i,$$

where Π_k^i denotes the realization of the revenue for the period in which price p_k is chosen for the ith time.

Algorithm 1 UCB Algorithm for The Classical MAB

Initialization: In the first K periods, choose each price p_k once for $k \in \mathbb{K}$.

for $t > K$ **do**

Upon observing c_t, choose the price with the index $I_t = \arg\max_{k\in\mathbb{K}} \bar{\pi}_{k,t} + \sqrt{\frac{2\log t}{T_k(t-1)}}$.

endfor

The classical UCB algorithm (i.e., Algorithm 1) proposes to balance the earning and learning via choosing a price p_k either with a small $T_k(t-1)$ or with a large estimated revenue $\bar{\pi}_{k,t}$. When $T_k(t-1)$ is small for some price p_k, the decision maker has incentive to conduct price experimentations on p_k to learn more accurately the associated expected revenue. When the estimated revenue $\bar{\pi}_{k,t}$ is large, p_k may lead to large revenue, and thus the decision maker tends to choose p_k. The trade-off between price experimentation and revenue maximization is balanced via the incorporation of an UCB of the revenue estimates. Essentially, the UCB algorithm ensures that the suboptimal prices will be chosen with a sufficiently large amount of times (in a logarithmic order) so that the optimal revenue and the suboptimal ones are distinguishable based on the resulting revenue estimates.

In our problem setting, the introduction of varying cost leads to more complicated forms of profit functions. However, the expected revenue π_k remains the same as that in the classical MAB setting. Therefore, we borrow the idea of the classical UCB algorithm to devise a pricing policy for our setting. More specifically, in period t, we consider a UCB of the expected revenue π_k for each k:

$$\bar{\pi}_{k,t} + \sqrt{\frac{2\log t}{T_k(t-1)}}.$$

Then, we obtain an estimate of the profit function $\mu_k(c)$:

$$\hat{\mu}_{k,t}(c) = \left(\bar{\pi}_{k,t} + \sqrt{\frac{2\log t}{T_k(t-1)}}\right)\left(1 - \frac{c}{p_k}\right), \quad k \in \mathbb{K}.$$

Intuitively, $\hat{\mu}_{k,t}(c)$ can be viewed as an UCB for profit function $\mu_k(c)$, for any values of c. Following the spirit of the UCB algorithm, we propose a pricing policy that chooses at the beginning of period t the price with the index $I_t(c_t) = \arg\max_{k\in\mathbb{K}} \hat{\mu}_{k,t}(c_t)$ upon observing the cost c_t. The pseudo-code for this policy is provided in Algorithm 2.

Algorithm 2 UCB-Like Policy

Initialization: In the first K periods, choose each price p_k once $k \in \mathbb{K}$.

for $t > K$ **do**

Upon observing c_t, choose the price with the index $I_t(c_t) = \arg\max_{k\in\mathbb{K}} \hat{\mu}_{k,t}(c_t)$.

endfor

In a nutshell, the pricing policy proposed in Algorithm 2 chooses in each period t the price that leads to the maximum UCB among all profit functions upon observing the outcome of the cost c_t. The chosen price depends on the observed realizations of c_t and varies from period to period. It should be emphasized that this pricing policy is a natural extension of the UCB algorithm to the setting with varying cost. In a special case where the cost in each period is a constant (and set to be zero without loss of generality), it can be easily checked that Algorithm 2 is exactly the same as Algorithm 1.

By counting the number of times the suboptimal prices can be chosen, one can derive that the expected cumulative regret of the UCB algorithm for the classical MAB problem is of an order of $\log T$, where T is the total number of selling periods. The regret analysis for the pricing policy in Algorithm 2 is, however, much more challenging, due to the introduction of the varying cost c_t. In the following section, we analyze the expected cumulative regret of the proposed pricing policy and provide insights on how the policy works.

4. Regret Analysis

In this section, we establish an upper bound for the expected cumulative regret $\mathbf{R}(T)$, defined in Equation (2), of the UCB-Like pricing policy. In particular, we are interested in developing an upper bound of the growth rate of $\mathbf{R}(T)$ as T increases. In the pricing literature, these upper bounds are often used to compare policies, and policies with slower growth rates are often more preferred.

Notice that $\mathbf{R}(T)$ can be divided into regrets incurred by choosing different prices, that is,

$$\mathbf{R}(T) = \sum_{k=1}^{K} \mathbf{R}_k(T),$$

where

$$\mathbf{R}_k(T) = \sum_{t=1}^{T} \mathbb{E}[(\mu_{i^*(c_t)}(c_t) - \mu_k(c_t))\mathbb{I}(I_t(c_t) = k)],$$

with $\mathbb{I}(A)$ being an indicator function, which is equal to 1 if event A occurs and 0 otherwise. In the regret analysis, we often focus on analyzing $\mathbf{R}_k(T)$ for all $k = 1, \ldots, K$ and sum them up to obtain the overall regret $\mathbf{R}(T)$.

To start the regret analysis, we define a pair of complementary events as follows.

Definition 4.1. *Let* $\mathbf{A}_t = \{\bar{\pi}_{k,t} \in CI_{k,t}, \forall k \in \mathbb{K}\}$, *where* $CI_{k,t}$ *is an interval defined by*

$$CI_{k,t} = \left[\pi_k - \sqrt{\frac{2\log t}{T_k(t-1)}}, \pi_k + \sqrt{\frac{2\log t}{T_k(t-1)}}\right].$$

The event $\mathbf{A}_t$ basically states that in period t, the estimated revenue for every price p_k falls into the interval $CI_{k,t}$. Then, we have the following lemma, which shows that $\hat{\mu}_{k,t}(c)$ is an upper bound of $\mu_k(c)$ for all $k \in \mathbb{K}$ and all $c \in \mathbb{C}$ if $\mathbf{A}_t$ occurs.

LEMMA 1. *If $\mathbf{A}_t$ occurs in period t, then $\hat{\mu}_{k,t}(c) \geq \mu_k(c)$ for all $k \in \mathbb{K}$ and $c \in \mathbb{C}$.*

Let $\mathbf{A}_t^c$ denotes the complement of $\mathbf{A}_t$. Notice that $\mathbf{A}_t^c$ means that at least one of the estimated revenues are not in the interval $CI_{k,t}$ in period t. In the next lemma we show that $\mathbf{A}_t^c$ occurs with a very small probability. Its proof is based on Hoeffding's inequality and is provided in the appendix.

LEMMA 2. (Probability bound for $\mathbf{A}_t^c$). $\sum_{t=1}^{T} \mathbb{P}(\mathbf{A}_t^c) \leq 3.6K$.

Lemma 2 implies that the expected number of occurrences of events $\mathbf{A}_t^c$ during the T periods is upper bounded by a constant. Combined with the fact that regret in a single period is at most 1, the total expected cumulative regret incurred due to the occurrences of $\mathbf{A}_t^c$ is bounded by a constant. Therefore, in what follows we mainly analyze the expected cumulative regret incurred when $\mathbf{A}_t$ occurs.

4.1. Inferior Prices and their Regrets

In our setting, different costs may correspond to different optimal prices. Among all possible prices $\{p_1, \ldots, p_K\}$, some of them may be optimal for some costs in $\mathbb{C}$, while others may never be optimal for any cost in $\mathbb{C}$. We call the latter *inferior prices* and define them as follows.

DEFINITION 4.2. *A price p_j is called an **inferior price** if there does not exist a nonempty set $\mathbb{C}_j \subseteq \mathbb{C}$ such that $\mu_j(c) > \max_{i \in \mathbb{K} \setminus \{j\}} \mu_i(c)$ for all $c \in \mathbb{C}_j$.*

To facilitate analysis, we make the following regularity condition on inferior prices.

ASSUMPTION 2. (Arm optimality). *If price p_j is inferior, there exists a constant $r > 0$ such that $\max_{i \neq j} \mu_i(c) > \mu_j(c) + r$ for all $c \in \mathbb{C}$.*

Assumption 2 rules out the case where $\mu_j(c)$ may be arbitrarily close to $\max_{i \neq j} \mu_i(c)$, which makes it difficult to identify its inferiority. This assumption is also commonly used in the literature of the MAB with covariate, for example, Assumption 3 in Goldenshluger and Zeevi (2013) and Assumption 3 in Bastani and Bayati (2020).

To analyze the expected cumulative regret, we first consider the regret incurred by choosing an inferior price. In the classical MAB problem, the total amount of times that an inferior price is chosen grows linearly in $\log T$. It turns out that, with the results in Lemmas 1 and 2, we can prove that our UCB-Like pricing policy also inherits this important property. We further note that when an inferior price is chosen, the resulting regret is bounded by 1, because $\sup_{i \neq j} |\mu_i(c) - \mu_j(c)| \leq 1$ for all $c \in \mathbb{C}$ due to the assumption that π_k takes values in [0, 1]. As a result, the total regret incurred by an inferior price is of the order $\log T$. Since this analysis is a direct application of the regret analysis of the UCB algorithm, we summarize this result in the following proposition and the proof is provided in the appendix.

PROPOSITION 1. *Suppose that p_j is an inferior price and Assumption 2 holds. Then,*

$$\mathbf{R}_j(T) \leq 3.6K + 1 + \frac{8\log T}{r^2}$$

for any $T = 1, 2, \ldots$, where r is defined in Assumption 2.

Proposition 1 shows that the expected cumulative regret incurred by choosing an inferior price grows linearly in $\log T$. Note that, in our pricing policy, we do not explore the smoothness of the demand function. In practice, the demand function may satisfy a certain smoothness condition. For example, if the demand is Lipschitz continuous, the revenue information we collect at one price may also reveal some revenue information on other prices. Then it is possible that as long as we have good estimations on some non-inferior prices, we do not need to choose an inferior price many times to conclude that it is inferior. Incorporating such information in the policy design may help us avoid unnecessary learning on some inferior prices and reduce the expected cumulative regret incurred by choosing these inferior prices. In the rest of this section, the analysis is focused on non-inferior prices.

4.2. Continuous Cost

In this subsection, we assume that the cost of the product c_t is a continuous random variable in $\mathbb{C}$. We make the following assumption on the distribution of $c_t | \mathscr{F}_{t-1}$.

ASSUMPTION 3. (Finite density). *There exists a constant $\xi > 0$ such that the probability density function $f(c|\mathscr{F}_{t-1}) \leq \xi$, for all $c \in \mathbb{C}$ and $t = 1, 2, \ldots, T$.*

One of the most crucial parts of the regret analysis for our pricing policy is that, when the cost (or covariates) falls in the vicinity area of the decision boundary, the pricing policy (or algorithms) is very likely to choose a suboptimal price. In the meantime, the amount of regret incurred is relatively small in this situation. Thus, simply counting the number of times the policy (or algorithms) chooses suboptimal prices without considering the amount of regret incurred in each time may overestimate the upper bound on the expected cumulative regret. In what follows, we show how to derive a tight upper bound on the expected cumulative regret of our pricing policy.

4.2.1. The Case with $K = 2$. We first consider the case with two prices, that is, $K = 2$, to illustrate the main idea of the regret analysis. Without loss of generality, we assume that both prices are non-inferior. For the case where one of them is inferior, the result in Proposition 1 implies that the expected cumulative regret incurred by choosing the inferior price grows linearly in $\log T$, and thus the total regret $\mathbf{R}(T)$ also grows linearly in $\log T$.

In what follows, we provide insights into the analysis of our main result (i.e., Theorem 1), while postponing its rigorous proof to the appendix. We mainly discuss the expected cumulative regret incurred by choosing p_1 (i.e., $\mathbf{R}_1(T)$), while analysis of that incurred by choosing p_2 follows in the same manner. Notice that by Lemma 2, the regret incurred due to the occurrences of $\mathbf{A}_t^c$ is negligible. We focus on analyzing $\mathbf{R}_1(T)$ when $\mathbf{A}_t$ holds. Thus, we can write $\mathbf{R}_1(T)$ as,

$$\begin{aligned}\mathbf{R}_1(T) &\approx \sum_{t=1}^{T}\mathbb{E}\Big[\Big(\mu_{i^*(c_t)}(c_t)-\mu_1(c_t)\Big)\mathbb{I}(I_t(c_t)=1,\mathbf{A}_t)\Big]\\ &= \sum_{s=1}^{T}\sum_{t=1}^{T}\mathbb{E}\Big[\Big(\mu_{i^*(c_t)}(c_t)-\mu_1(c_t)\Big)\mathbb{I}(I_t(c_t)=1,\mathbf{A}_t,T_1(t-1)=s)\Big]\\ &= \sum_{s=1}^{T}\sum_{t=1}^{T}\mathbb{E}\big[(\mu_2(c_t)-\mu_1(c_t))\mathbb{I}\big(\text{falsely choose } p_1\big)|I_t(c_t)=1,\mathbf{A}_t,T_1(t-1)=s\big]\cdot\\ &\quad\ \mathbb{P}(I_t(c_t)=1,\mathbf{A}_t,T_1(t-1)=s),\end{aligned} \tag{3}$$

Therefore, it is critical to analyze

$$r_{1,t}(s)=\mathbb{E}\big[(\mu_2(c_t)-\mu_1(c_t))\mathbb{I}\big(\text{falsely choose } p_1\big)|I_t(c_t)=1,\mathbf{A}_t,T_1(t-1)=s\big].$$

Figure 2 shows a possible configuration of the true profit functions and their UCB functions for both p_1 and p_2 conditioned on $\mathbf{A}_t$ and $T_1(t-1) = s$. By Lemma 1, the straight lines $\hat{\mu}_{1,t}(c)$ and $\hat{\mu}_{2,t}(c)$ lie above $\mu_1(c)$ and $\mu_2(c)$, respectively for all $c\in\mathbb{C}$. Furthermore, since $\mathbf{A}_t$ holds, one can deduce that $\hat{\mu}_{1,t}(c)$ lies below the dotted line. Therefore, both the points O and $\bar{O}$ are on the left hand side of the line segment EE', that is, $x_O \le x_E$ and $x_{\bar{O}} \le x_E$. Then, it is clear that the UCB-Like policy chooses p_1 if the cost $c_t \le x_{\bar{O}}$ and p_2 otherwise. A positive regret incurs only when $x_O \le c_t \le x_{\bar{O}}$ (i.e., falsely choosing p_1), and the regret is $\mu_2(c_t)-\mu_1(c_t)$. Notice that in our analysis, we assume that $x_{\bar{O}} > x_O$ and ignore the case where $x_{\bar{O}} \le x_O$ because no regret would be incurred by choosing p_1 in this case.

In the circle of Figure 2, we zoom in the area (marked as grey) where the regret incurs. From the figure, the first result we can derive is that, when $\mathbf{A}_t$ holds and $T_1(t-1) = s$,

$$\mu_2(c_t)-\mu_1(c_t)\le E\bar{E}',$$

where $E\bar{E}'$ denotes the length of the line segment EE'. Hence,

$$r_{1,t}(s)\le E\bar{E}'\cdot\mathbb{P}\big(\text{falsely choose } p_1|I_t(c_t)=1,\mathbf{A}_t,T_1(t-1)=s\big). \tag{4}$$

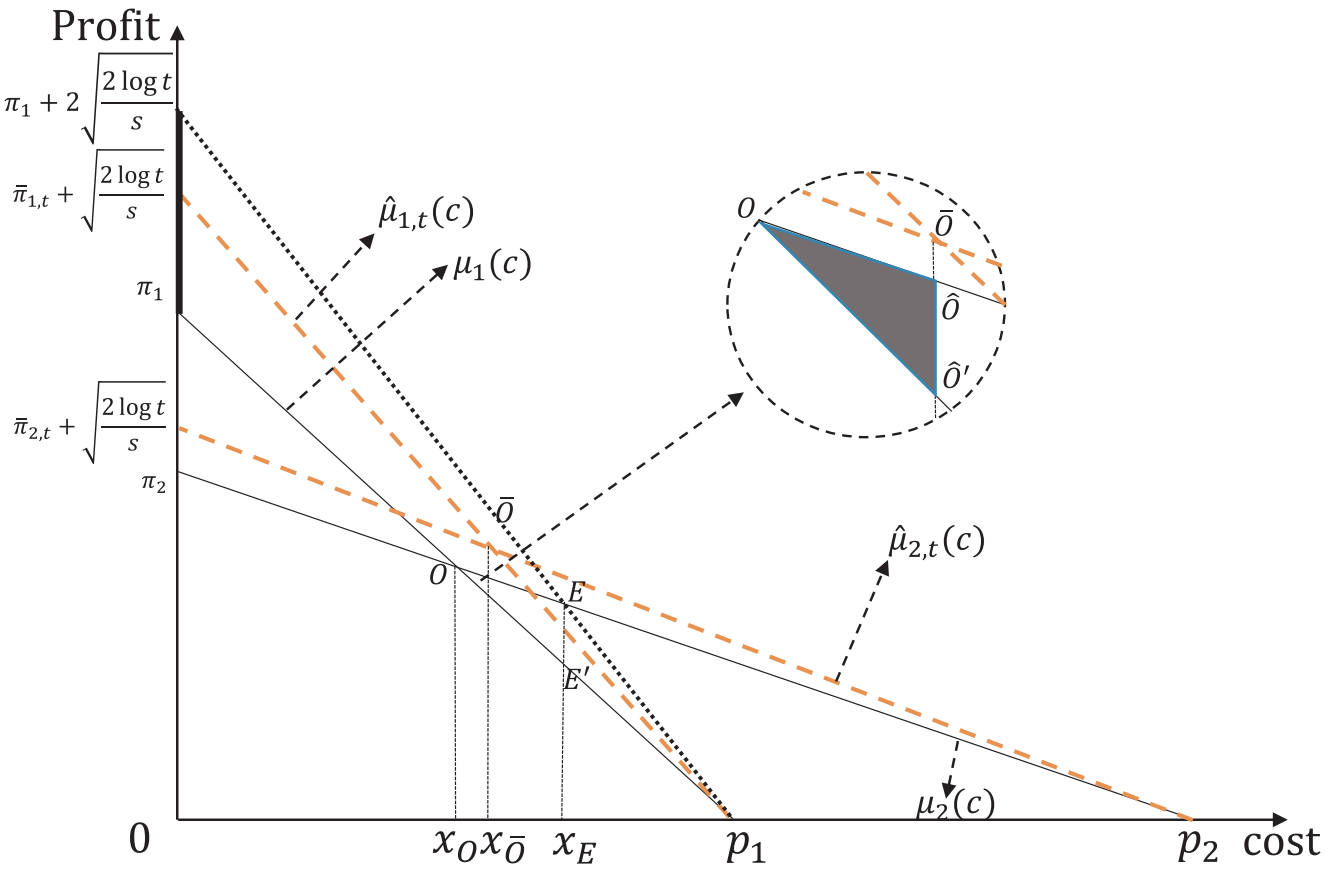
To analyze the probability in Equation (4), we notice that falsely choosing p_1 happens when c_t falls in the interval $[x_O, x_{\bar{O}}]$. By Assumptions 1 and 3, the probability that c_t falls in an interval is proportional to the length of the interval. We use the fact that the length of $[x_O, x_{\bar{O}}]$ is upper bounded by $x_E - x_O$ and prove the following result,

$$\mathbb{P}\big(\text{falsely choose } p_1|I_t(c_t)=1,\mathbf{A}_t,T_1(t-1)=s\big)\le(x_E-x_O)\frac{\xi}{\kappa_1}, \tag{5}$$

where κ_1 is a constant and ξ is defined in Assumption 3. Then, by elementary geometry,

$$x_E-x_O=\frac{x_O}{\pi_1-\pi_2}E\bar{E}'. \tag{6}$$

Figure 2 Configuration of $\mu_1(c)$ and $\mu_2(c)$ [Color figure can be viewed at wileyonlinelibrary.com]



Combining Equations (4)–(6) yields

$$r_{1,t}(s) \le \frac{x_O\xi}{\kappa_1(\pi_1-\pi_2)}E\bar{E}'^2 \le \frac{x_O\xi}{\kappa_1(\pi_1-\pi_2)}\frac{8\log t}{s} \le \frac{8x_O\xi\log T}{\kappa_1(\pi_1-\pi_2)s}, \tag{7}$$

where the second inequality holds because $E\bar{E}'$ is less than the length of the bold line segment on y-axis. Therefore, we have,

$$\begin{aligned}\mathbf{R}_1(T) &\approx \sum_{s=1}^{T}\sum_{t=1}^{T} r_{1,t}(s)\mathbb{P}(I_t(c_t)=1,\mathbf{A}_t,T_1(t-1)=s)\\ &\le \frac{8x_O\xi\log T}{\kappa_1(\pi_1-\pi_2)}\sum_{s=1}^{T}\frac{1}{s}\sum_{t=1}^{T}\mathbb{P}(I_t(c_t)=1,\mathbf{A}_t,T_1(t-1)=s)\\ &\le \frac{8x_O\xi}{\kappa_1(\pi_1-\pi_2)}(\log T)^2,\end{aligned}$$

where the last inequality follows from the fact that the event $\{I_t(c_t)=1,T_1(t-1)=s\}$ can occur at most once from $t=1$ to T.

To summarize, the expected cumulative regret incurred by choosing p_1 grows linearly in $(\log T)^2$. The analysis for the expected regret incurred by choosing p_2 follows in the same manner. We summarize this result in the following theorem. The detailed proof of the theorem is provided in the appendix.

THEOREM 1. *Suppose that Assumption 3 holds, $K = 2$ and both prices are non-inferior. Then,*

$$\mathbf{R}(T) \le \frac{8\xi x_O}{\pi_1-\pi_2}\left(\frac{1}{\kappa_1}+\frac{1}{\kappa_2}\right)(\log T)^2 + \left\lceil\frac{8\log T}{(\pi_1-\pi_2)^2}\right\rceil + 14.4$$

for all $T = 1, 2, \ldots$, where $\kappa_1 > 0$ and $\kappa_2 > 0$ are two constants.

4.2.2. The Case with General $K \ge 2$. The regret analysis for the case with $K = 2$ may be extended to cases with $K \ge 2$. In what follows we provide a general discussion on how the analysis may be extended and give a theorem to characterize the bound of the expected cumulative regret.

In order to characterize the growth rates of functions, we first introduce two notations: 1) $h(n) = \mathcal{O}(g(n))$ means that $|h(n)|$ is bounded above by $g(n)$ asymptotically, that is, there exist two constants $\beta > 0$ and $n_0 > 0$ such that

$$|h(n)| \le \beta \cdot g(n),$$

for any $n \geq n_0$; 2) $h(n) = \Theta(g(n))$ stands for the fact that $h(n)$ grows in the same order with $g(n)$, that is, there exist constants $\beta_1 > 0$, $\beta_1' > 0$ and $n_1 > 0$ such that

$$\beta_1 \cdot g(n) \leq h(n) \leq \beta_1' \cdot g(n),$$

for any $n \geq n_1$.

One of the key insights we can derive from the analysis of the $K = 2$ case is that the expected number of times a non-inferior price is selected by period t by the UCB-Like policy grows linearly in t. Otherwise, the sub-linear growth rate on the expected cumulative regret cannot be attained. We prove in the appendix that this is also true for the case with $K \geq 2$, and the result is summarized in the following lemma.

LEMMA 3. (Linear growth rate). *If price p_k is not an inferior price, then* $\mathbb{E}[T_k(t-1)] = \Theta(t)$.

Lemma 3 implies that the expected number of times that any non-inferior price is chosen grows linearly with time t. However, only the result on $\mathbb{E}[T_k(t-1)]$ is not enough, we also need to ensure that, at every time t, $T_k(t-1)$ does not deviate significantly from its expectation. For that, we need the following concentration inequality (Lemma 4). To prove it, we construct a martingale process that enables us to use Azuma's Inequality (Azuma 1967). The detailed proof is provided in the appendix.

LEMMA 4. (Concentration inequality). *For any $\lambda > 0$,*

$$\mathbb{P}(T_k(t-1) \geq \mathbb{E}[T_k(t-1)] + \lambda) \leq \exp\left(-\frac{\lambda^2}{2t}\right),$$

$$\mathbb{P}(T_k(t-1) \leq \mathbb{E}[T_k(t-1)] - \lambda) \leq \exp\left(-\frac{\lambda^2}{2t}\right).$$

With Lemmas 3 and 4, the growth rate of the expected cumulative regret can be established, which is summarized in the following theorem. The proof of the theorem is provided in the appendix. Unlike the result of $K = 2$ in Theorem 1, we are unable to establish a finite-time bound for the expected cumulative regret. The result in Theorem 2 is only asymptotic.

THEOREM 2. *Suppose that Assumptions* 2 *and* 3 *hold. Then,* $\mathbf{R}(T) = \mathcal{O}((\log T)^2)$

Compared to the case where the cost is a constant, when the cost is a continuous random variable, as the cost varies, the profit of the optimal price can be arbitrarily close to that of the second-best price, making it very difficult to make the correct decision. Theorems 1 and 2 show that, our UCB-Like pricing policy has an expected cumulative regret growing in the order of $(\log T)^2$ in this situation.

We are aware of that as there exists a linear relationship between the profit and the cost, our problem can be alternatively viewed as a linear bandit problem with one covariate, that is, the cost, and existing linear bandit algorithms may be applied to solve our problem. Some of them have expected cumulative regrets that grow linearly in $\log T$. Therefore, in terms of the growth rate of the expected cumulative regret, our UCB-Like policy performs slightly worse than these linear bandit algorithms do. However, to attain the $\log T$ growth rate, these algorithms often require one's prior knowledge on the profit functions. For example, in the classical work of Goldenshluger and Zeevi (2013), the authors develop an ordinary least square (OLS) algorithm. In contrast to selecting prices based on the UCBs of different profit functions used by our UCB-Like policy, it directly uses the OLS method to estimate the profit functions and select prices. Because OLS estimators are unbiased estimators and better converge to the true profit functions, it may lead to lower regret. However, choosing the prices only based on OLS estimators may lead to insufficient learning of some prices and cause the incomplete learning phenomenon. Therefore, some forced sampling is added to the algorithm to ensure sufficient learning. A sequence of time points, where certain prices are forced selected regardless of the observed costs, is predetermined at the beginning of the algorithm. To determine these time points, one must know some structure information about the profit functions, for example, the maximum distance between the profit of a price and the profits of the others in the region where the price is optimal. As we shall see in section 5, due to the lack of such information, the performance of the algorithm may not be satisfactory in practice. Moreover, because positive regret is always incurred at forced sampling time points, the growth rate of the expected cumulative regret in the OLS algorithm remains the same when the cost is discrete. In the next subsection, we show that when the cost is discrete, the expected cumulative regret of our UCB-Like policy may be upper bounded by a constant.

We close this subsection with a remark that, in the appendix, motivated by the OLS algorithm of Goldenshluger and Zeevi (2013), we propose a forced sampling (FS) policy of ours. We show that the expected cumulative regret of our FS policy also grows linearly in $\log T$. Compared to the OLS algorithm in which the forced sampling is static and predetermined at the beginning of the algorithm, our FS policy dynamically conducts forced sampling when insufficient learning is detected. We add the performance of our FS policy while comparing the numerical performances of our UCB-Like policy and some linear bandit algorithms in section 5.

4.3. Discrete Cost

We now consider the situations where the cost has a finite discrete distribution. We make the following assumption.

ASSUMPTION 4. (Finite cost set). *The support of the cost c_t is a finite set, that is, $|\mathbb{C}|<\infty$.*

Notice that Assumption 4 often holds in practical applications. For instance, customers buying the same insurance product in different states can lead to different expected costs. The cost is discrete in this situation. In the example of selling fresh products, the cost of the product may also be quoted from a menu of a finite number of wholesale prices. Then, the cost is also discrete. Indeed, one may even argue that *most practical applications have a discrete cost*, because costs are typically set with a minimal unit, for example, dollar or cent. Therefore, the discrete cost may be arguably more realistic than the continuous one.

There is a fundamental difference between discrete and continuous costs. Considering the case of $K = 2$ in Figure 2, we see that a positive regret occurs only if the cost $c_t \in [x_O, x_{\bar{O}}]$. When the profit functions $\hat{\mu}_{1,t}(c)$ and $\hat{\mu}_{2,t}(c)$ are estimated accurate enough in period t, the interval $[x_O, x_{\bar{O}}]$ may be so small that it does not include any point in $\mathbb{C}$. Then, there will be no positive expected cumulative regret in this period. This intuition allows us to prove the following theorem and the proof is included in the appendix.

THEOREM 3. *Suppose that there are no inferior prices and Assumption 4 holds. Then, there exists a constant $C_1 > 0$ so that $\mathbf{R}(T) \le C_1$ for all $T = 1, 2, \ldots$.*

Theorem 3 shows that the dynamic pricing problem with varying cost may be easier to solve than that with constant cost, even though the optimal is now a function of the cost. It also reveals that the earning and learning trade-off no longer exists in the limit. To explain this result, notice that when the cost is discrete, we can always find a smallest positive profit gap between the optimal price(s) and the suboptimal ones for any $c \in \mathbb{C}$. In the presence of the gap, the UCB-Like policy only conducts learning when a price is chosen less than a $\log t$ order of times. By Lemma 3, all non-inferior prices are chosen with a linear order of times as they are optimal for some costs, and thus, sufficient to meet the demand for learning. Therefore, no learning is necessary beyond a certain time point and this leads to a constant upper bound on the expected cumulative regret.

In a more general case where inferior prices may exist, it has been shown in Proposition 1 that the expected cumulative regret incurred by choosing the inferior price grows linearly in $\log T$ as T increases. Adding this result to Theorem 3, we have the following corollary.

COROLLARY 1. *Suppose that there are inferior prices and Assumptions 2 and 4 hold. Then, there exist constants $C_2 > 0$ and $C_2' > 0$ so that $\mathbf{R}(T) \le C_2 + C_2' \log T$ for all $T = 1, 2, \ldots$.*

5. Numerical Results

In this section, we use several numerical examples to examine the performances of the proposed policy under different cost settings. The objective of the numerical studies is to demonstrate that the performances of the UCB-Like policy indeed match the theoretical results in the previous sections and compare the policy with the FS policy provided in the appendix and some existing linear bandit algorithms numerically, aiming to shed lights on which policy is more preferred in practical settings.

5.1. Discrete Costs

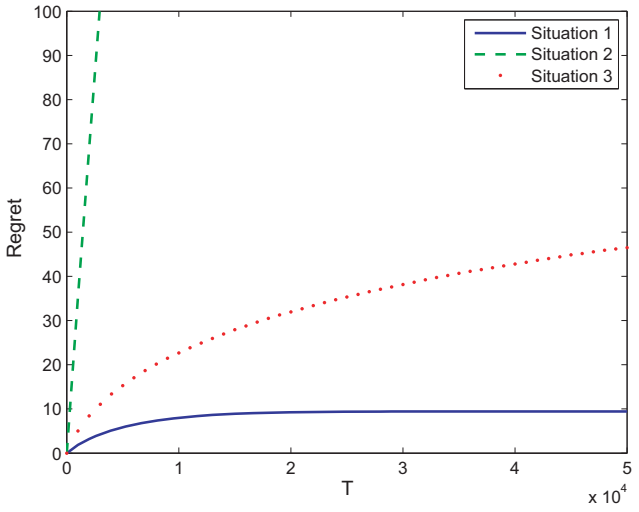
In this subsection, we consider an example of two prices. These two prices are non-inferior, and the cost takes discrete values from a finite set. More specifically, suppose $p_1 = 4$ and $p_2 = 7$, and the corresponding revenues following Bernoulli distributions with mean $\pi_1 = 0.6$ and $\pi_2 = 0.4$ after normalization, respectively. We let the cost observed in each period be independent and identically distributed, taking values $\{1, 2, 3, 4\}$ with equal probabilities. For this example, the optimal price is $p_1 = 4$ if the observed cost is equal to 1 or 2, and p_2 otherwise.

To understand the benefit gained from considering a varying cost and the performance of the UCB-Like policy, we consider three situations: 1) the cost is random, and we treat it as random; 2) the cost is random, but we treat it as deterministic and use its mean value 2.5; and 3) the cost is deterministic, and it is 2.5. We use the UCB-Like policy in the first situation, and the UCB policy in the rest two. To compare the performance, we estimated the expected cumulative regrets of all three situations, based on 1000 independent replications. The results are summarized in Figure 3.

We highlight the main findings from this set of numerical studies as follows.

- Treating a random cost as deterministic results in a fast growing expected cumulative regret and the growth rate is linear in T (Situation 2, dash line in Figure 3).
- Taking into consideration of the varying cost leads to much smaller expected cumulative regret (Situation 1, solid line in Figure 3). The gap between Situations 1 and 2 can be very

Figure 3 Benefit of Considering Varying Cost [Color figure can be viewed at wileyonlinelibrary.com]



large, showing that considering varying cost may bring a significant amount of profit for firms.

- Comparing Situation 3 (dotted line in Figure 3) to Situation 1, we find that, while introducing the varying cost makes the problem more challenging, it also makes it possible to significantly reduce the expected cumulative regret and thus increase the profit.
- As T becomes large (e.g., larger than 3×10^4 in Figure 3), the expected cumulative regret of the UCB-Like policy (the solid line) tends to become flat, and appears to be bounded above as T increases. This observation matches the theoretical result in Theorem 3.

5.2. Continuous Cost

In this subsection, we provide a comparison of the UCB-Like policy, the FS policy proposed in the Appendix, and two linear bandit algorithms, namely, the classical ordinary least squares (OLS) algorithm proposed by Goldenshluger and Zeevi (2013) and a state-of-the-art algorithm, the Greedy-First algorithm proposed by Bastani et al. (2020) for the continuous cost setting.

The Greedy-First algorithm is recently proposed by Bastani et al. (2020). It is an improvement of the traditional forced sampling algorithms for linear bandit problems. The algorithm first tries a greedy algorithm to conduct the selection. When the algorithm detects that the learning speed of the underlying profit functions is below a threshold, it will switch to a traditional forced sampling algorithm, for example, the OLS algorithm. By doing so, the algorithm conducts learning only when necessary and thus greatly reduces the effort spent on forced sampling when there is no inferior price in the system. Meanwhile, in the presence of an inferior price, this algorithm is equivalent to the traditional forced sampling algorithm as the switch occurs with probability one.

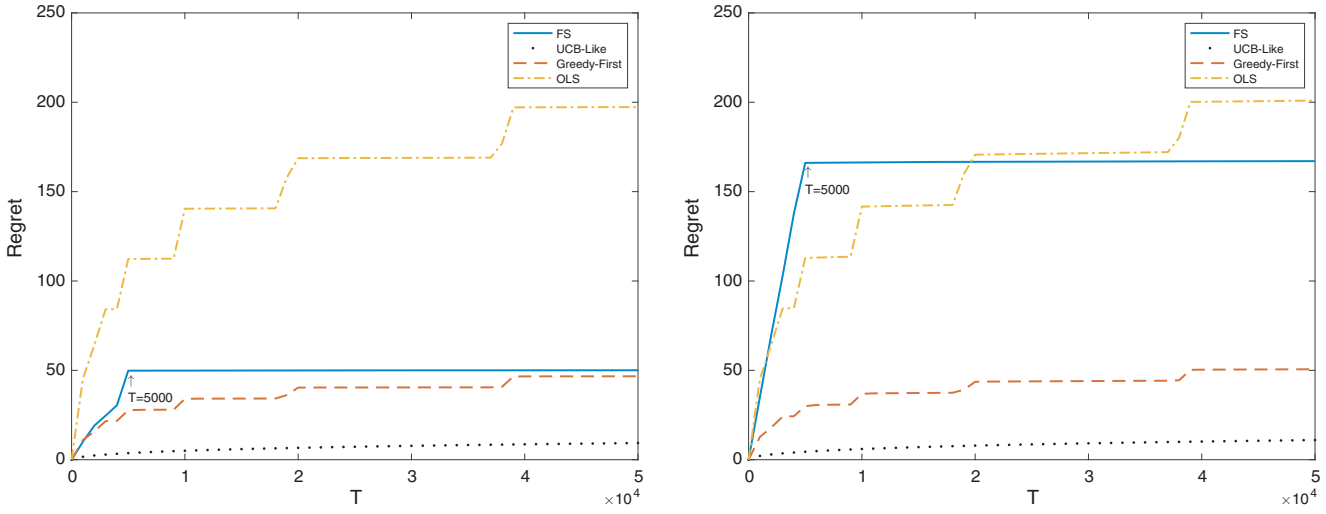
In this example, all the settings are the same as those in the previous example except that the cost observed in each period is sampled from a distribution with the support [0, 4], and we consider two types of distributions where the revenues are drawn: Bernoulli distributions with means π_1 and π_2 and uniformly distributions with supports $[\pi_1-0.4, \pi_1+0.4]$ and $[\pi_2-0.4, \pi_2+0.4]$. Moreover, we assume that the probability density function of c_t depends on the cost observed in the previous period, that is,

$$f(c|c_{t-1}\le 2)=\begin{pmatrix}0.3, & \text{if } 0\le c\le 2\\ 0.2, & \text{if } 2<c\le 4\end{pmatrix} \quad \text{and } f(c|c_{t-1}>2)$$
$$=\begin{pmatrix}0.2, & \text{if } 0\le c\le 2\\ 0.3, & \text{if } 2<c\le 4\end{pmatrix}$$

It suggests that a large cost in the current period is more likely to associate with a large cost in the following period. Initially, we let $c_0 = 2$. For the FS policy and the two linear bandit algorithms, they require the users to input a parameter h to determine when the forced sampling should be conducted. For the FS policy, h should be set strictly less than 0.171 to guarantee statistical validity in this case. We tried several different values for h and found that in general, the larger h is, the better performance the FS policy has. Thus, we set the parameter $h = 0.17$ for the FS policy in this experiment. For consistency, in this numerical study, we implement the OLS algorithm and the Greedy-First algorithm based on the pseudo-codes provided in Bastani et al. (2020). Note that besides h, their performances additionally depend on an input parameter q. We set $h = 0.1$ and $q = 300$, at which both algorithms perform relatively well. Moreover, for the Greedy-First algorithm, the number of warm-up periods[2] t_0 is set to be 20.

We apply all four policies in this example, and the results are summarized in Figure 4. From the figure, we observe that the UCB-Like policy has the smallest expected cumulative regret and outperforms the other three policies in both settings. The OLS algorithm performs the worst in the setting of uniformly distributed revenues for all $T \ge 1$ and performs worse than the other three policies when $T \ge 20{,}000$ in the setting where the revenues are drawn from Bernoulli distributions. In terms of robustness, the change of revenue distributions affects more on the FS policy than the other three policies. Notice that the expected cumulative regret of the FS policy comes mostly from the

Figure 4 **A Comparison of the Upper Confidence Bound-like, the Forced Sampling, the Ordinary Least Square, and the Greedy-First Policies: No Inferior Price [Color figure can be viewed at wileyonlinelibrary.com]**



(a) Revenues: Uniform Distribution

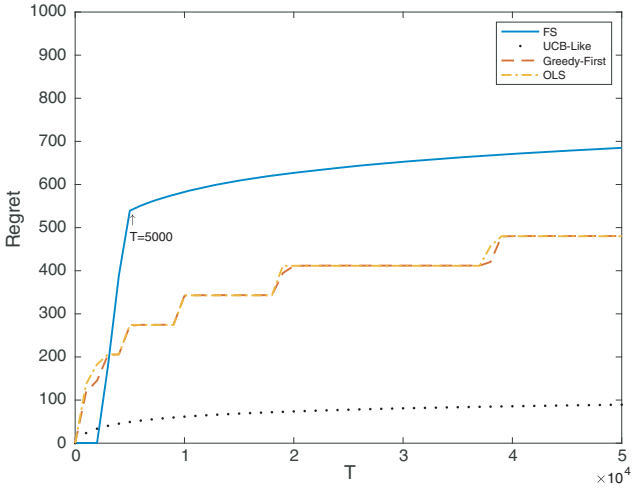
(b) Revenues: Bernoulli Distribution

early stages in both settings, particularly $T \leq 5000$. This phenomenon can be explained as follows. When $T = 5000$ and $h = 0.17$, $8 \log T/h^2 \approx T/2$. Because the FS policy conducts forced sampling whenever it detects $T_k(t-1) \leq \frac{8\log t}{h^2}$ for some k, it suggests that, when $t \leq 5000$, the FS policy mainly conducts forced sampling so that each price gets roughly the same amount of samples to ensure sufficient learning. Therefore, the regret incurred in the early stages is considerably large. From this example, we may conclude that, even though, the growth rates of the FS policy, the OLS algorithm, and the Greedy-First algorithm are theoretically smaller than that of the UCB-Like policy, in this particular example, we do not observe that the theoretically superior performances are translated into practical advantages for these three policies for reasonable sizes of T. Even when we resume the experiment and push T up to 10^6, the expected cumulative regret of the UCB-Like policy is still much smaller than those of the other three policies.

Then, we test the performances of all four policies when one of the prices is inferior. We interchange the expected revenues of the two prices in the previous example, that is, we let $\pi_1 = 0.4$ and $\pi_2 = 0.6$. Therefore, $p_1 = 4$ is an inferior price. Since except for the FS policy, the performances of the other three policies do not change much as we consider different types of distributions for the revenues, in this example, we only consider the situation where the revenues are uniformly distributed. We run all four policies to solve this problem and plot the results in Figure 5.

From Figure 5, we can draw the following conclusions. First, in this example, we still observe that the

Figure 5 **A Comparison of the Upper Confidence Bound-like, the Forced Sampling, the Ordinary Least Square, and the Greedy-First Policies: p_1 is an Inferior Price [Color figure can be viewed at wileyonlinelibrary.com]**



expected cumulative regret of the UCB-Like policy is much smaller than those of the other three policies for all $T \geq 1$. Second, because one of the prices is inferior, while solving this problem, the Greedy-First algorithm would switch to the OLS algorithm with probability one. Therefore, the performances of the Greedy-First algorithm and the OLS algorithm are almost the same in this example. Third, in this example, because the FS policy always needs to spend some sampling effort on the inferior price which can incur positive regret, we can observe that the expected cumulative regret of the FS policy keeps increasing after a sharp increase in the early stages.

To summarize, our numerical results suggest that, the UCB-Like policy may have a better practical performance among all four policies. Therefore, we suggest using the UCB policy even when the cost is continuous.

6. Concluding Remarks

In this article, we study a dynamic pricing problem where the demand function is unknown and the cost varies from period to period. We develop a UCB-Like policy to balance the earning and learning for this dynamic pricing problem, and show that the expected cumulative regret of the policy grows linearly in $(\log T)^2$ when the cost is continuous, where T denotes the number of selling periods. In a special case when the cost takes a finite number of discrete values and there are no inferior prices, we show that the expected cumulative regret is bounded above by a constant as T increases. Compared to some existing linear bandit algorithms which can also be used to solve this problem and have expected cumulative regrets that grow linearly in $\log T$, the expected cumulative regret of our policy grows slightly faster for the case of continuous cost and slower for the case of discrete cost. However, for these linear bandit algorithms, better asymptotic results do not always translate into better practical performances because to attain the $\log T$ growth rate, they typically require foreknowledge of the profit functions that is unknown in practice. Our numerical studies reveal that the UCB-Like policy is very competitive with existing linear bandit algorithms for up to a reasonably large T. The numerical results also confirm the theoretical growth rates developed in this paper appear correct for the UCB-Like policy.

There are two potential research directions of this work. First, in this study, we consider the setting where the cost of the product is the same at different price levels in each period. In practice, however, a firm may sell a product with different qualities at different price levels. For example, one may observe that the higher the price, the lower its deductible (i.e., higher cost) for an insurance product. Under this setting, the cost additionally depends on the chosen price, and one may observe different costs at different price levels. Even though the UCB-Like policy can still be applied to this case by incorporating different cost information at different price levels, the current regret analysis for the proposed policy is no longer applicable. It is of both practical and theoretical interests to develop efficient policies with sound theoretical guarantees for this setting, which is left as a topic for future research. Second, in this paper, we conduct the upper bound analysis on the expected cumulative regret of our policy under different parameter settings. In many similar works, lower bound analysis is also an important part of regret analysis. It would be equally valuable to conduct lower bound analysis under these parameter settings in the future.

Acknowledgments

The authors would like to thank the Department Editor, Prof. Dan Zhang, the Senior Editor, and the anonymous reviewers for their insightful and detailed comments that have significantly improved this paper. L. Jeff Hong is the corresponding author of this paper. This research was supported in part by the National Natural Science Foundation of China [NSFC 72091211, NSFC 72071148] and the Research Grants Council of Hong Kong [GRF 11508620].

Notes

[1]Sellers only select prices from a finite list of admissible prices, maybe due to marketing purposes, industrial consensus, or managerial considerations. The same finite-choice setting is used commonly in the revenue management literature, see for instance, Gallego and Van Ryzin (1994) and Feng and Xiao (2000).

[2]For the Greedy-First algorithm, when revenues are drawn from Bernoulli distributions, directly using a greedy algorithm in warm-up periods can lead the algorithm to switch to the OLS algorithm with high probability. To avoid this issue, in this paper, we select the two prices equal number of times in the warm-up periods for the case where revenues are sampled from Bernoulli distributions.

References

Auer, P., N. Cesa-Bianchi, P. Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **47**(2–3): 235–256.

Azuma, K. 1967. Weighted sums of certain dependent random variables. *Tohoku Math. J.* **19**(3): 357–367.

Ban, G. Y., N. B. Keskin. 2017. Personalized Dynamic Pricing with Machine Learning. Working paper, University of Maryland, College Park, MD, SSRN: Available at: https://ssrn.com/abstract=2972985 (accessed date Feb 23, 2021).

Bastani, H., M. Bayati. 2020. Online decision making with high-dimensional covariates. *Oper. Res.* **68**(1): 276–294.

Bastani, H., M. Bayati, K. Khosravi. 2020. Mostly exploration-free algorithms for contextual bandits *Management Sci.* Forthcoming.

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

Besbes, O., A. Zeevi. 2015. On the (surprising) sufficiency of linear models for dynamic pricing with demand learning *Management Sci.* **61**(4): 723–739.

den Boer, A. V., B. Zwart. 2014. Simultaneously learning and optimizing using controlled variance pricing. *Management Sci.* **60** (3): 770–783.

Broder, J., P. Rusmevichientong. 2012. Dynamic pricing under a general parametric choice model. *Oper. Res.* **60**(4): 965–980.

Bubeck, S., N. Cesa-Bianchi. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.* **5**(1): 1–122.

Caro, F., J. Gallien. 2012. Clearance pricing optimization for a fast-fashion retailer. *Oper. Res.* **60**(6): 1404–1422.

Chen, Y., C.-W. Lee, H. Luo, C.-Y. Wei. 2019. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. A. Beygelzimer, D. Hsu, eds. *Proceedings of the Thirty-Second Conference on Learning Theory*. PMLR, Phoenix, USA, pp. 696–726.

Feng, Y., B. Xiao. 2000. A continuous-time yield management model with multiple prices and reversible price changes. *Management Sci.* **46**(5): 644–657.

Gabillon, V., M. Ghavamzadeh, A. Lazaric, S. Bubeck. 2011. Multi-bandit best arm identification. J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, K. Q. Weinberger, eds. *Advances in Neural Information Processing Systems 24*. Curran Associates, Inc., Granada, Spain, 2222–2230.

Gallego, G., G. Van Ryzin. 1994. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Sci.* **40**(8): 999–1020.

Goldenshluger, A., A. Zeevi. 2009. Woodroofe's one-armed bandit problem revisited. *Ann. Appl. Probab.* **19**(4): 1603–1633.

Goldenshluger, A., A. Zeevi. 2013. A linear response bandit problem. *Stoch. Syst.* **3**(1): 230–261.

Harrison, J. M., N. B. Keskin, A. Zeevi. 2012. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Sci.* **58**(3): 570–586.

Keskin, N. B., J. R. Birge. 2019. Dynamic selling mechanisms for product differentiation and learning. *Oper. Res.* **67**(4): 1069–1089.

Keskin, N. B., A. Zeevi. 2014. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Oper. Res.* **62**(5): 1142–1167.

Keskin, N. B., A. Zeevi. 2016. Chasing demand: Learning and earning in a changing environment. *Math. Oper. Res.* **42**(2): 277–307.

Keskin, N. B., A. Zeevi. 2018. On incomplete learning and certainty-equivalence control. *Oper. Res.* **66**(4): 1136–1167.

Lai, T. L., H. Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **6**(1): 4–22.

Lobo, M. S., S. Boyd. 2003. Pricing and Learning with Uncertain Demand. Working paper, Duke University, Durham, NC, Available at https://web.stanford.edu/~boyd/papers/pdf/pric_learn_unc_dem.pdf (accessed date Feb 23, 2021).

McLaughlin, E. W., D. J. Perosio. 1994. Fresh fruit and vegetable procurement dynamics: The role of the supermarket buyer. Research Bulletin, Cornell University, Ithaca, NY, Available at http://ageconsearch.umn.edu/record/123000 (accessed date Feb 23, 2021).

Mersereau, A. J., P. Rusmevichientong, J. N. Tsitsiklis. 2009. A structured multiarmed bandit problem and the greedy policy. *IEEE Trans. Automat. Contr.* **54**(12): 2787–2802.

Qiang, S., M. Bayati. 2016. Dynamic pricing with demand covariates. Working paper, Stanford University, Stanford, CA, arXiv: Available at https://arxiv.org/abs/1604.07463 (accessed date Feb 23, 2021).

Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.* **58**(5): 527–535.

Rothschild, M. 1974. A two-armed bandit theory of market pricing. *J. Econ. Theory* **9**(2): 185–202.

Rusmevichientong, P., J. N. Tsitsiklis. 2010. Linearly parameterized bandits. *Math. Oper. Res.* **35**(2): 395–411.

Sarkar, J. 1991. One-armed bandit problems with covariates. *Ann. Stat.* **19**(4): 1978–2002.

Woodroofe, M. 1979. A one-armed bandit problem with a concomitant variable. *J. Am. Stat. Assoc.* **74**(368): 799–806.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix S1: Appendix A: A Forced Sampling Policy.
Appendix B: Useful Known Technical Results without Proofs.
Appendix C: Proofs of Lemmas.
Appendix D: Proofs of Propositions.
Appendix E: Proofs of Theorems.