When Demands Evolve Larger and Noisier: Learning and Earning in a Growing Environment

Feng Zhu [1] **Zeyu Zheng** [2]

Abstract

We consider a single-product dynamic pricing problem under a specific non-stationary setting, where the underlying demand process grows over time in expectation and also possibly in the level of random fluctuation. The decision maker sequentially sets price in each time period and learns the unknown demand model, with the goal of maximizing expected cumulative revenue over a time horizon T. We prove matching upper and lower bounds on regret and provide near-optimal pricing policies. We show how the growth rate of random fluctuation over time affects the best achievable regret order and the near-optimal policy design. In the analysis, we show that whether the seller knows the length of time horizon T in advance or not surprisingly render different optimal regret orders. We then extend the demand model such that the optimal price may vary with time and present a novel and near-optimal policy for the extended model. Finally, we consider an analogous non-stationary setting in the canonical multi-armed bandit problem, and points out that knowing or not knowing the length of time horizon T render the same optimal regret order, in contrast to the non-stationary dynamic pricing problem.

1. Introduction

In a standard dynamic pricing problem, a merchant, not knowing the demand model at the beginning, dynamically chooses at each time period a price, observes a demand, and collects revenue. Specifically, the price elasticity, that the ratio of change in mean demand in response to a change in price, is not known a priori. The merchant dynamically sets price to learn the price elasticity and to trigger revenue collection, with the goal of maximizing the expected cumulative revenue over a time horizon of T. Such a problem has been extensively studied under convenient yet often violated assumptions on stationarity or stable scale. That is, if given a fixed price over time, the random demand in each period remains the same in distribution, or varies from time to time within a fixed range of distributions in an adversarial sense (stable scale). However, in many business settings, the underlying demand for a product can be *growing* over time, possibly due to growth of business scale, increasing popularity, stronger advertisement, network effects, among other reasons. For example, the demand typically presents a growing trend for successful start-up companies offering new services or products. The demand growth can also be found for certain products sold on e-commerce platforms; see (Glynn & Zheng, 2019). When the demand distribution does not remain stationary or stable in scale over time, the optimal dynamic pricing strategy and best achievable performance analysis in the stationary or stable scale setting may not be valid.

This paper considers a specific non-stationary growing setting where the demand's expectation and its level of random fluctuation grow over time t at rates of t^γ and t^α respectively for $\alpha, \gamma \geq 0$. We assume that the base-line demand model is unknown, while the the growth rates γ and α are known. By defining revenue *regret* as the expected optimal cumulative revenue if the demand model is known a prior subtracted by the expected cumulative revenue achieved by a feasible policy, we prove matching upper and lower bounds on revenue regret, and provide new dynamic pricing policies that are near-optimal (off by a logarithm order). We show how different growth rates α in random fluctuation render different designs of near-optimal pricing policies as well as different best achievable regret orders. Furthermore, in our analysis, we distinguish the cases on whether the merchant knows the length of time horizon T in advance (*fixed-time* scheme) or not (*any-time* scheme), and show that it leads to different optimal regret orders. Lastly, we contrast the dynamic pricing setting with a classical bandit setting in a growing environment and illustrate the differences.

[1]School of Mathematical Sciences, Peking University, Beijing, China [2]Department of Industrial Engineering and Operations Research, University of California, Berkeley, USA. Correspondence to: Feng Zhu <lingxiyuan@pku.edu.cn>, Zeyu Zheng <zyzheng@berkeley.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

1.1. Related Work

When the scale of the demand does not grow in the expectation or in the level of random fluctuation (corresponding to $\gamma = 0$ and $\alpha = 0$), our problem becomes the standard single-product dynamic pricing problem with an unknown stationary demand model, in which the expected demand depends linearly in price. Dynamic pricing with online learning has aroused great interest in revenue management and resource allocation problems. For a comprehensive review, readers may refer to den Boer (2015). Many works focus on the trade-off between *exploration* (conducting price experiments to acquire more information about the demand model) and *exploitation* (leveraging current information to improve revenue in each time period) in the basic or extended settings of the dynamic pricing problem. A part of the representative works include Besbes & Zeevi (2009), Besbes & Zeevi (2011), Broder & Rusmevichientong (2012), Harrison et al. (2012), den Boer & Zwart (2014), Keskin & Zeevi (2014), Qiang & Bayati (2016), Keskin & Zeevi (2017), Chen & Gallego (2018), Nambiar et al. (2019), Bu et al. (2019), (Li & Zheng, 2019), Ban & Keskin (2020) etc. Various demand models, parametric or non-parametric, have been considered in the literature. One particular relevant works to ours is Keskin & Zeevi (2014). In Keskin & Zeevi (2014), the authors give sufficient conditions for a policy to achieve optimal regret order in the dynamic pricing problem under the stationarity assumption, and extend it to the case of multiple products. One of our pricing policies generalizes theirs, but we go beyond by demonstrating that this policy is optimal only in the *any-time* scheme, but not in the *fixed-time* scheme. In this paper, we adopt a simple parametric demand model, and focus on the new features centered around the *growing* environment, where both the expectation of the demand and the level of random fluctuation can grow, at possibly different rates.

Our work is also closely related to multi-armed bandit (MAB) problems. Readers could refer to Lattimore & Szepesvári (2019) and Slivkins et al. (2019) for comprehensive discussion about this field. Contextual bandit problem is a direct generalization of MAB, where the reward is parameterized by a linear function with unknown parameters and the action is characterized by contexts. A part of the representative works include Auer et al. (2002), Dani et al. (2008), Rusmevichientong & Tsitsiklis (2010), Filippi et al. (2010), Abbasi-Yadkori et al. (2011), etc. Among the vast amount of literature, Abbasi-Yadkori et al. (2011) is particularly relevant to ours, where they give a strong high probability inequality for least square estimation with general contexts. However, their goal is to improve the confidence region for standard contextual bandit problem where the reward is bounded over the time, while in our case the contexts are typically increasing and unbounded. Further, we consider the effect of changing random fluctuation on policy design, while standard literature does not.

Regarding online learning under non-stationarity, many works consider bandit or pricing problems in a non-stationary environment where the parameters change over time within a compact set, see, e.g., Besbes & Zeevi (2011), Besbes et al. (2014), Keskin & Zeevi (2017), Cheung et al. (2018), Chen et al. (2019), Auer et al. (2019), etc. Their non-stationarity describes turbulence of model parameters within a given range, and the expected revenue in each period is on the same scale over the entire time horizon. In reality, their assumption means that the scale and range of demand does not change. However, the demand of product over time may present a structured growing pattern, rather than changes with a fixed scale. We consider the non-stationarity that stems from the growth of demand and therefore the expected revenue in each period may be growing over time without a pre-fixed range.

An important topic in online learning is transforming a *fixed-time* policy to an *any-time* one such that the total time periods T is not necessary to know in advance. Previous literature typically focus on the *fixed-time* situation, and the *any-time* case is either naturally included (see, e.g., Keskin & Zeevi, 2014; Cheung et al., 2017) or can be obtained by using methods such as *Doubling Trick*. *Doubling Trick* is a well known technique in online learning, which can be traced back to Auer et al. (1995). Besson & Kaufmann (2018) gives a comprehensive discussion on how to apply this trick for MAB problems to maintain the same regret order. However, in this work we show that in certain cases of our dynamic pricing problem, it is *impossible* to transform an optimal *fixed-time* policy into an *any-time* one because the optimal regret orders in the two situations do not match. Therefore, *doubling trick* is of no use.

1.2. Our Contribution

Our contributions in this work are three-folds:

Structured Model: We study the dynamic pricing problem under a demand model that flexibly capture the rates of growth in the expected demand and the associated random fluctuation respectively. The literature typically consider stationary or stable scale scenarios, in the sense that the demand scale is limited to a compact set without growth. Our work accommodates the scenarios when the demand range grows without a pre-selected bound, and to our knowledge is the first to develop performance analysis and pricing policy design facing growth both in the demand expectation and the level of random fluctuation. The novel modeling of the growing environment renders the discovery of two somewhat surprising findings on how the best achievable regret order depends on the growth rate of the random fluctuation and a critical difference on whether the decision maker knows the length of time horizon T in advance or not.

New Analysis: We provide analysis on best achievable performances by proving best achievable regret orders. The analysis and associated best achievable regret orders are different between the *fixed-time* and *any-time* schemes. In the non-stationary growing environment, we find that knowing the length of the time horizon T in advance or not may lead to a significant difference in the best achievable regret orders, and the difference only mitigates when the growth rate of random fluctuation is small compared to the growth rate of the expectation. In contrast, a growing environment does not create such a difference for the MAB problems, where the best achievable regret orders do not differentiate between the *fixed-time* and *any-time* schemes.

Policy Design: We develop near-optimal (off by logarithm order) pricing policies that match the best achievable performances in terms of regret order. We demonstrate how the optimal policy designs change with respect to the growth rate in demand expectation and the level of random fluctuation. Intuitively, a larger growth rate of the random fluctuation imposes more difficulties in learning and therefore impedes the best achievable revenue. Our optimal policy designs address two issues uniquely aligned with the growing environment. One is that there is no uniform upper bound on the range of random demand. Second is that the convergence rates for the estimators on different parameters can be different.

1.3. Structures and Notation

The paper is organized as follows. In Section 2, we formally introduce our dynamic pricing problem, along with pricing policies and regret definition. We state our main results in Section 3, including matching lower and upper bounds in both *any-time* and *fixed-time* cases, plus a generalization for *fixed-time* case when the growth rate of random fluctuation is relatively large. We compare in Section 4 the MAB problem in a growing environment with the dynamic pricing problem in Section 2. In Section 5, we conclude.

Throughout the paper, all the vectors are column vectors unless otherwise specified. For each $m \in \mathbb{Z}_+$, we use $[m]$ to denote the set $\{1, ..., m\}$. For any $x \in \mathbb{R}^d$ and any positive semi-definite matrix A, we use $\|x\|_A = (x^\top A x)^{\frac{1}{2}}$ to denote the A-norm of x. When A is the identity matrix, we denote $\|x\| = (x^\top x)^{\frac{1}{2}}$ as the l_2-norm. For any compact set Γ, we denote $\|\Gamma\| = \sup_{x\in\Gamma} \|x\|$. The notation $O(\cdot)$, $\Omega(\cdot)$ and $\Theta(\cdot)$ are used by hiding the constant factors.

2. Dynamic Pricing Problem

Model Description We consider a demand model in which the demand in the t-th time period depends on the price p through the following relation

$$d_t(p) = t^\gamma(a - bp) + \epsilon_t. \tag{1}$$

Here $\gamma \geqslant 0$ and the random fluctuation sequence $\{\epsilon_t : t \geqslant 1\}$ are independent and identically distributed (i.i.d.) $t^\alpha\sigma$-sub-Gaussian random variables with mean zero. The scale t^γ and t^α means that the mean demand grows with time at a polynomial rate γ, while the scale of the demand fluctuation grows at rate α. We set $\theta = (a, b) \in \Theta$ and assumes that the range $\Theta = [a_{\min}, a_{\max}] \times [b_{\min}, b_{\max}]$ is a compact set in $\mathbb{R}^2_+$. In the learning problem, we assume that the merchant knows α, γ, σ, and Θ, but does not know the true demand model. Also, we suppose that the feasible price p must live in $[l, u]$ where $u > l > 0$. The assumption of $[l, u]$ is primarily for ease of demonstration and has also appeared in literature, e.g., Keskin & Zeevi (2014), Bu et al. (2019). It in essential assumes that the feasible/optimal price is bounded within a positive range. In fact, as Θ is a positive compact set, the assumption is implicit. We define the revenue function in period t as $r_t(p) = p\mathbb{E}_{\epsilon_t}[d_t(p)] = t^\gamma p(a - bp)$. For each $t \geqslant 0$, we use p_t as the price set in period t and d_t as the realized demand in period t.

Pricing Policies Let H_t be the vector of information available at the end of period t, i.e., $H_t = (p_1, d_1, \cdots, p_t, d_t)$. Now we can formally define two classes of pricing policies: *any-time* and *fixed-time*.

An *any-time* pricing policy is defined as an *infinite* sequence of functions $\pi = (\pi_1, \pi_2, \cdots)$, where $\pi_t : \mathbb{R}^{2t-2} \to [l, u]$ is a measurable function which maps the information vector H_{t-1} to a feasible price in $[l, u]$ for all $t \geqslant 1$.

A *fixed-time* pricing policy for T periods is defined as a *finite* sequence of functions $\pi = (\pi_1, \pi_2, \cdots, \pi_T)$, where $\pi_t : \mathbb{R}^{2t-2} \times \mathbb{Z}_+ \to [l, u]$ is a measurable function which maps the information vector H_{t-1} and the total time periods T to a feasible price in $[l, u]$ for all $1 \leqslant t \leqslant T$.

The above definition of *any-time* pricing policy is consistent with that in Keskin & Zeevi (2014) and Bu et al. (2019). A *fixed-time* policy knows the total time periods T in advance, while an *any-time* policy does not. Evidently, an *any-time* policy generates a price p_t adapted to H_{t-1} in period t. An *any-time* policy can be reduced to a *fixed-time* policy for T periods by choosing its first T components, but extending a *fixed-time* policy to an *any-time* one is non-trivial.

Regret For any pricing policy π, we let $\mathbb{P}^\pi_\theta$ and $\mathbb{E}^\pi_\theta$ be the probability measure and expectation induced by π under the parameter θ respectively. We define the regret $R^\pi_\theta(T)$ as the gap between the optimal revenue and the expected revenue collected by policy π under the parameter θ within T periods, i.e.,

$$R^\pi_\theta(T) = \sum_{t=1}^{T} r_t(p^*_t) - \mathbb{E}^\pi_\theta\Big[\sum_{t=1}^{T} r_t(p_t)\Big], \tag{2}$$

where p_t^* is the optimal price, i.e., $p_t^* = \frac{a}{2b} \triangleq \phi_t(\theta)$. Here we assume that for any $\theta \in \Theta$, the optimal price under θ is an interior point of $[l, u]$. Otherwise, we can choose l and u such that $[\frac{a_{\min}}{2b_{\max}}, \frac{a_{\max}}{2b_{\min}}] \subset [l, u]$. In our basic setting, p_t^* (or $\phi_t(\theta)$) is stationary over the time. Let $\Delta_t = \frac{1}{b}(\phi_t(\theta)(a - b\phi_t(\theta)) - p_t(a - bp_t)) = (\phi_t(\theta) - p_t)^2$. Then $R_\theta^\pi(T)$ can be rewritten as

$$R_\theta^\pi(T) = b\sum_{t=1}^{T} t^\gamma \mathbb{E}_\theta^\pi[\Delta_t].$$

3. Main Results

We start by stating two lemmas that will be frequently referenced in subsequent theorems. We first introduce Lemma 1, a generalized concentration inequality for arbitrary sub-Gaussian random variables.

Lemma 1. *Let $X_1, ..., X_n$ be mean zero independent sub-Gaussian random variables with*

$$\mathbb{E}[\exp(\lambda X_i)] \leqslant \exp\left(\frac{\lambda^2\sigma_i^2}{2}\right), \forall \lambda \in \mathbb{R},$$

i.e., X_i is σ_i-sub-Gaussian for all $i \in [n]$. Then for all $\delta > 0$, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| > \delta\right) \leqslant 2\exp\left(-\frac{\delta^2}{2\sum_{i=1}^{n}\sigma_i^2}\right).$$

We then introduce Lemma 2 that characterizes asymptotic behavior of the sum of s^x.

Lemma 2. *For any $x \geqslant -1$, let $S_{x,t} = \sum_{s=1}^{t} s^x$. We have*

$$0 < S_{x,t} - \frac{t^{x+1} - 1}{x + 1} \leqslant \max\{1, t^x\},$$

where we define $\frac{t^0-1}{0} = \lim_{x\to 0^+}\frac{t^x-1}{x} = \log t$.

From Lemma 2, for any fixed $x \geqslant -1$, we have

$$S_{x,t} = \begin{cases} \Theta(t^{x+1}), & \text{if } x > -1, \\ \Theta(\log t), & \text{if } x = -1. \end{cases}$$

3.1. Lower Bound

In this subsection, we establish the general regret lower bound, i.e., the order of the best achievable regret, for all possible pairs of growth parameters $\gamma \geqslant 0$ and $\alpha \geqslant 0$. Furthermore, we differentiate between *fixed-time* and *any-time* regret lower bound. Typically in the dynamic pricing literature, the proved regret lower bound is for *fixed-time* policies, while we show that the *fixed-time* and *any-time* scenarios may lead to different regret lower bounds. We note that in the growing environment, the optimal expected revenue given by $\sum_{t=1}^{T} r_t(p_t^*)$ is at the order of $T^{\gamma+1}$.

Theorem 1. *Fix $\alpha \geqslant 0$ and $\gamma \geqslant 0$. Then under the demand model (1):*

- ***(Fixed-time regret lower bound)***
  - *If $\alpha \in [0, \gamma + \frac{1}{2})$, then there exists an absolute constant $C > 0$ such that for any pricing policy π, $\sup_{\theta\in\Theta}\{R_\theta^\pi(T)\} \geqslant CT^\beta$, $\forall T \geqslant 1$, where*

$$\beta = \min\left\{\alpha + \frac{1}{2}, \frac{(\gamma+1)^2}{3\gamma - 2\alpha + 2}\right\}.$$

    Specifically, $\beta = \alpha + \frac{1}{2}$ when $\alpha \leqslant \frac{\gamma}{2}$ and $\beta = \frac{(\gamma+1)^2}{3\gamma-2\alpha+2}$ when $\alpha \in (\frac{\gamma}{2}, \gamma + \frac{1}{2})$.
  - *If $\alpha \geqslant \gamma + \frac{1}{2}$, then there exists an absolute constant $C > 0$ such that for any pricing policy π,*

$$\sup_{\theta\in\Theta}\{R_\theta^\pi(T)\} \geqslant \begin{cases} C\frac{T^{\gamma+1}}{\log T}, & \text{if } \alpha = \gamma + \frac{1}{2}, \\ CT^{\gamma+1}, & \text{if } \alpha > \gamma + \frac{1}{2}, \end{cases}$$

    $\forall T \geqslant 2$.
- ***(Any-time regret lower bound)*** *If $\alpha \in [0, \gamma + \frac{1}{2})$, then there exists an absolute constant $C > 0$ such that for any any-time pricing policy π,*

$$\limsup_{T}\left\{\sup_{\theta\in\Theta}\{R_\theta^\pi(T)\}/T^{\alpha+\frac{1}{2}}\right\} \geqslant C.$$

Note that all constants are independent of T, but depend on α, γ, σ, and Θ. Theorem 1 gives a full characterization of the optimal regret order for arbitrary $\alpha \geqslant 0$ and $\gamma \geqslant 0$:

1. When the growth rate of random fluctuation is *relatively small*, i.e., $\alpha \leqslant \frac{\gamma}{2}$, the optimal regret order *only* depends on α. When the growth rate of random fluctuation is *relatively large*, i.e., $\alpha \in (\frac{\gamma}{2}, \gamma + \frac{1}{2})$, the optimal regret order is dependent on both α and γ and is strictly smaller than $\alpha + \frac{1}{2}$. Note that the boundary $\alpha = \frac{\gamma}{2}$ indicates that the mean and the variance of the demand grow over time at the same rate.
2. When the growth rate of random fluctuation is *much larger* than that of mean demand, i.e., $\alpha > \gamma + \frac{1}{2}$, a sub-linear regret is inapproachable for any policy. That $\alpha = \gamma + \frac{1}{2}$ is a critical boundary condition, since our result shows that a sub-linear regret may be achievable, but only by a marginal logarithm order.

Theorem 1 also illustrates that the best achievable regret critically depends on whether the total time periods T is known or not. Here we characterize the *any-time* regret lower bound by stating that the *fixed-time* regret of any *any-time* policy must be $\Omega(T^{\alpha+\frac{1}{2}})$ for infinitely many T, and the constant is universal for all policies. The result on *any-time*

regret lower bound implies that for $\alpha \in [0, \gamma + \frac{1}{2})$, if an *any-time* policy achieves a regret $\sup_{\theta\in\Theta}\{R^{\pi}_{\theta}(T)\} = O(T^{\beta})$ for any T and some β, then we must have $\beta \geqslant \alpha + \frac{1}{2}$. Differently, the optimal *fixed-time* regret can achieve an order smaller than $\alpha + \frac{1}{2}$ when $\alpha > \frac{\gamma}{2}$.

Our proof of Theorem 1 utilizes van Tree inequality to quantify the information retrieval as in Keskin & Zeevi (2014). However, the intrinsic difference between known and unknown T cases requires new proof techniques to differentiate them, and the treatment of lower bound with large random fluctuation in both settings require delicate and novel splitting and estimation techniques. We estimate the sum by splitting $[1, T]$ into $[1, T_0]$ and $(T_0, T]$ (known T) or based on the power of 2 (unknown T) such that the magnitude of errors is carefully controlled. These splitting techniques were not used in the literature and could potentially handle other non-stationary settings.

Theorem 1 indicates that, up to a logarithm factor, a sublinear regret is accessible only if $\alpha \in [0, \gamma + \frac{1}{2})$. Therefore, in the remaining part of this section, we will present pricing policies that match the lower bound (up to a logarithm factor) in Theorem 1 when $\alpha \in [0, \gamma + \frac{1}{2})$.

3.2. Upper Bound: Any-time Pricing Policy

In this subsection, we will illustrate how to design an *any-time* pricing policy that matches the lower bound stated in the second part of Theorem 1. Equivalently, we will present a pricing policy that, without knowing the total time periods T in advance, achieves $O(T^{\alpha+\frac{1}{2}} \log T)$ regret for arbitrary $T \geqslant 3$. The exploration-exploitation trade-off in our policy is enlightened by the optimal stationary pricing policy in Keskin & Zeevi (2014).

Before proceeding to policy design, we introduce new notation and problem pre-processing. Let $\tilde{d}_t(p) = t^{-\alpha} d_t(p)$ and $\tilde{\epsilon}_t = t^{-\alpha}\epsilon_t$. Therefore $\tilde{\epsilon}_t$ becomes a σ-sub-Gaussian noise. We rewrite the demand model using the scaled quantities

$$\tilde{d}_t(p) = t^{\gamma-\alpha}(a - bp) + \tilde{\epsilon}_t. \tag{3}$$

For any $x \geqslant -1$, we define

$$P_{x,t} = \sum_{s=1}^{t} s^x p_s, \quad Q_{x,t} = \sum_{s=1}^{t} s^x p_s^2.$$

We set $\bar{p}_t = \frac{P_{2\gamma-2\alpha,t}}{S_{2\gamma-2\alpha,t}}$ as a weighted average of prices up to time t. Assume $\mathcal{J}_t$ is invertible, then given information H_t, the least square estimator $\hat{\theta}_t$ from (3) is $\mathcal{J}_t^{-1}\mathcal{D}_t$, where

$$\begin{aligned}\mathcal{J}_t &= \sum_{s=1}^{t} [s^{\gamma-\alpha}\ s^{\gamma-\alpha}p_s]^{\top}[s^{\gamma-\alpha}\ s^{\gamma-\alpha}p_s] \\ &= \begin{bmatrix} S_{2\gamma-2\alpha,t} & P_{2\gamma-2\alpha,t} \\ P_{2\gamma-2\alpha,t} & Q_{2\gamma-2\alpha,t} \end{bmatrix}\end{aligned} \tag{4}$$

and

$$\mathcal{D}_t = \sum_{s=1}^{t} [s^{\gamma-\alpha}\ s^{\gamma-\alpha}p_s]^{\top}\tilde{d}_s. \tag{5}$$

Let

$$\vartheta_t = \arg\min_{\vartheta\in\Theta}\{\|\vartheta - \hat{\theta}_t\|\} \tag{6}$$

be the projected estimator. Further, let

$$J_t = \sum_{s=1}^{t} s^{2\gamma-2\alpha}(p_s - \bar{p}_t)^2$$

be the weighted variance of the prices up to time t. Its normalized version $\tilde{J}_t = J_t / S_{2\gamma-2\alpha,t}$ can be regarded as a measure of price deviance. We introduce Lemma 3 that describes the relation between the information matrix and price deviance.

Lemma 3. *Let $\mu_{\min}(t)$ be the smallest eigenvalue of $\mathcal{J}_t$. Then $\mu_{\min}(t) \geqslant \mu J_t$, where $\mu = 2/(1 + 2u - l)^2$.*

An implication of Lemma 3 for designing pricing policy is that by controlling J_t not to be too small, we can lower bound the rate of information acquisition. Building on the observation, we show in Lemma 4 that the least square estimation error decreases exponentially as a function of J_t.

Lemma 4. *There exist finite positive constants ρ and k such that, under any pricing policy π,*

$$\begin{aligned}&\mathbb{P}^{\pi}_{\theta}(\|\hat{\theta}_t - \theta\| > \delta, J_t \geqslant m) \\ &\quad\leqslant k(1 \vee \delta) S_{2\gamma-2\alpha,t} \exp(-\rho(\delta \wedge \delta^2)m)\end{aligned}$$

for all $\delta, m, \alpha, \gamma \geqslant 0$ and $t \geqslant 2$.

Lemma 4 gives a precise characterization on how fast the estimation error may converge as the information accumulates. This leads to the following sufficient conditions for pricing policies to achieve asymptotic near-optimality.

Theorem 2. *Fix γ and $\alpha \in [0, \gamma + \frac{1}{2})$. Let κ_0 and κ_1 be two positive constants and π be a pricing policy such that*

- $J_t \geqslant \kappa_0 \sqrt{S_{2\gamma-2\alpha,t}}$
- $\sum_{s=2}^{t} s^{\gamma}(\phi(\vartheta_s) - p_{s+1})^2 \leqslant \kappa_1 S_{\alpha-\frac{1}{2},t}$

for all $t \geqslant 2$, then there exists a positive constant C such that

$$R^{\pi}_{\theta}(T) \leqslant C T^{\alpha+\frac{1}{2}} \log T \tag{7}$$

for all $T \geqslant 3$ and $\theta \in \Theta$, where C is only concerned with $\alpha, \gamma, l, u, \sigma, \kappa_0, \kappa_1, \Theta$.

The two conditions in Theorem 2 perfectly present the exploration-exploitation trade-off. The first condition states that the price deviance should present a sufficient growth over time, which demonstrates the necessity of exploration. While the second condition imposes that cumulative deviation of the executed price from the myopic optimal should not be too big, enforcing sufficient degree of exploitation.

Theorem 2 points out for near-optimal policies how the magnitude of exploration (reflected by $\tilde{J}_t$) and the magnitude of exploitation (reflected by the cumulative deviance) change with different growth parameters α and γ. When γ increases, $\tilde{J}_t$ should become smaller, while the cumulative deviance should not change. In contrast, when α increases, both $\tilde{J}_t$ and the cumulative deviance should become larger, indicating that a faster growing random fluctuation requires more exploration. Note that when $\alpha = \gamma = 0$, the two conditions in Theorem 2 coincide with the sufficient conditions in Keskin & Zeevi (2014).

We next provide Algorithm 1 that satisfies the conditions in Theorem 2. This policy is an *any-time* pricing policy. In Algorithm 1, for any time period $t \geqslant 2$, we control the price p_{t+1} close to the myopic optimal $\phi(\vartheta_t)$, but meanwhile guarantee a small distance $\kappa t^{\frac{\alpha-\gamma}{2}-\frac{1}{4}}$ between p_{t+1} and $\bar{p}_t$. We have Corollary 1 that establishes the optimality of Algorithm 1.

Algorithm 1 Any-time Pricing Policy

Input: threshold $\kappa > 0$
Initialize estimation $\vartheta_0 \in \Omega$ and price $p = p_0 \in [l, u]$.
for $t = 0, 1, \cdots$ **do**
  if $t = 0$ **then**
    Set $p_{t+1} = \phi(\vartheta_0)$.
  else if $t = 1$ **then**
    Set p_{t+1} such that $p_{t+1} \neq p_t$.
  else
    Let ϑ_t be the estimator in (6).
    Set $\xi_t = \phi(\vartheta_t) - \bar{p}_t$.
    if $|\xi_t| < \kappa t^{\frac{\alpha-\gamma}{2}-\frac{1}{4}}$ **then**
      Set $p_{t+1} = \bar{p}_t + \kappa \cdot \mathrm{sgn}(\xi_t) t^{\frac{\alpha-\gamma}{2}-\frac{1}{4}}$.
    else
      Set $p_{t+1} = \phi(\vartheta_t)$.
    end if
  end if
end for

Corollary 1. *Fix γ and $\alpha \in [0, \gamma + \frac{1}{2})$. Then Algorithm 1 admits a regret $\sup_{\theta\in\Theta}\{R^{\pi}_{\theta}(T)\} = O(T^{\alpha+\frac{1}{2}} \log T)$ for all $T \geqslant 3$.*

3.3. Upper Bound: Fixed-time Pricing Policy

In this subsection, we will illustrate the design of a *fixed-time* pricing policy that matches the lower bound stated in the first part of Theorem 1 when the random fluctuation is relatively large ($\alpha > \frac{\gamma}{2}$). Equivalently, we will present a pricing policy that, *with* knowing the total time periods T in advance, achieves $O(T^{\frac{(\gamma+1)^2}{3\gamma-2\alpha+2}} \log T)$ regret for any fixed $T \geqslant 3$. Recall that when $\alpha \in [0, \frac{\gamma}{2}]$, the optimal *fixed-time* regret order is $\alpha + \frac{1}{2}$, which coincides with the optimal *any-time* regret order. Thus, we only need to focus on large random fluctuation cases, i.e., $\alpha \in (\frac{\gamma}{2}, \gamma + \frac{1}{2})$. We present our fixed-time pricing policy in Algorithm 2.

Algorithm 2 Fixed-time Pricing Policy

Initialize $\lambda = 1 + u^2$, $\eta = \frac{\gamma+1}{3\gamma-2\alpha+2} < 1$ and $l \leqslant l_0 < u_0 \leqslant u$. Set $c \in (0, 1)$ and $T_0 = cT^{\eta} < T$.
for $t = 0, \cdots, T_0 - 1$ **do**
  Set price as $p_{t+1} = l_0 \cdot \mathbb{1}\{t \text{ is even}\} + u_0 \cdot \mathbb{1}\{t \text{ is odd}\}$.
end for
for $t = T_0, \cdots, T - 1$ **do**
  Let $\hat{\theta}_t = (\lambda I + \mathcal{J}_t)^{-1}\mathcal{D}_t$ be the (biased) least square estimator, where $\mathcal{J}_t$ and $\mathcal{D}_t$ are defined in (4) and (5).
  Define confidence set

$$\mathcal{C}_t = \{\theta' : \|\theta' - \hat{\theta}_t\|_{\mathcal{J}_t} \leqslant w_t\},$$

  where w_t is defined as

$$\sigma\sqrt{2\log\left(\frac{1}{2}S_{\gamma,T}\left(1 + S_{2\gamma-2\alpha,t}\right)\right)} + \lambda^{\frac{1}{2}}\|\Theta\|. \quad (8)$$

  if $\mathcal{C}_t \cap \Theta \neq \varnothing$ **then**
    Set $(p_{t+1}, \vartheta_t) = \arg\max_{p\in[l,u],\theta'\in\mathcal{C}_t\cap\Theta} p(a - bp)$.
  else
    Set an arbitrary price $p_{t+1} \in [l, u]$.
  end if
end for

Rather than simultaneous exploration and exploitation throughout the whole time horizon, our policy adopts *pure exploration* at the beginning. In Algorithm 2, we first select a time window $[1, T_0]$ to conduct *pure exploration*. At this stage, we alternately set prices as l_0 and u_0 to collect demand observations as "offline data" for later steps. Note that T_0 is carefully designed such that $T_0 = \Theta(T^{\frac{\gamma+1}{3\gamma-2\alpha+2}})$. After time T_0, we apply a canonical linear-UCB type algorithm that fully captures the information obtained in the pure-exploration stage. The design is built upon Lemma 5 that is initially proposed in Abbasi-Yadkori et al. (2011). A similar idea of design has appeared in Bu et al. (2019), but in their scenario the "offline data" is given and known, so the merchant does not need to tune T_0.

Lemma 5 (Confidence Ellipsoid)**.** *For any $\delta \in (0, 1)$, the*

following event happens w.p. at least $1-\delta$,

$$\|\theta-\hat{\theta}_t\|_{\mathcal{J}_t} \leqslant \sigma\sqrt{2\log\left(\frac{\det(\mathcal{J}_t)^{\frac{1}{2}}\det(\lambda I)^{-\frac{1}{2}}}{\delta}\right)}+ \lambda^{\frac{1}{2}}\|\Theta\|,\ \forall t>0.$$

Theorem 3 establishes the optimality of Algorithm 2.

Theorem 3. *Fix* γ *and* $\alpha\in[0,\gamma+\frac{1}{2})$. *Then Algorithm 2 admits a regret* $\sup_{\theta\in\Theta}\{R^{\pi}_{\theta}(T)\}=O(T^{\frac{(\gamma+1)^2}{3\gamma-2\alpha+2}}\log T)$ *for all* $T\geqslant 3$.

We now add some insight on how we design Algorithm 2. Firstly, traditional UCB fails because it cannot easily control the price deviation, and consequently, the information matrix, at the beginning. In traditional UCB, the reward in each time period is uniformly bounded, and the analysis critically relies on such assumption. In our problem, the analysis becomes ineffective. Second, the exploration phase is enlightened by our proof of lower bound, where we split $[1,T]$ to $[1,T_0]$ and $(T_0,T]$, and only focus on estimating the latter phase. Intuitively, this means that at the beginning the loss is negligible to the regret order and we could seize this opportunity to collect as much information as possible.

3.4. Upper Bound: Extension with Intercept Terms

Before ending this section, we consider extending our demand model by adding an intercept term and investigating the optimal regret order *when variance is relatively large*. Specifically, we extend our model as

$$d_t(p)=a_0+t^{\gamma}(a-bp)+\epsilon_t, \tag{9}$$

where $\gamma>0$ and $\{\epsilon_t: t\geqslant 1\}$ are i.i.d. random variables with ϵ_t be $t^{\alpha}\sigma$-sub-Gaussian. Further, we assume that $\theta=(a_0,a,b)$ lies in a compact set $\Theta=[a_{0\min},a_{0\max}]\times[a_{\min},a_{\max}]\times[b_{\min},b_{\max}]\subset\mathbb{R}^3_+$, and we denote

$$\theta(0)=a_0,\ \theta(1)=a,\ \theta(2)=b.$$

Compared to the original demand model (1), the extended version (9) better calibrates the demand process especially in the beginning. Note that the optimal price of (9) changes over time. We will show that when $\alpha\in(\frac{\gamma}{2},\gamma+\frac{1}{2})$, the optimal regret order in Section 3.3 still holds in this scenario.

Let $\tilde{d}_t(p)=t^{-\alpha}d_t(p)$ and $\tilde{\epsilon}_t=t^{-\alpha}\epsilon_t$, we have

$$\tilde{d}_t(p)=t^{-\alpha}a_0+t^{\gamma-\alpha}(a-bp)+\tilde{\epsilon}_t,$$

where $\tilde{\epsilon}_t$ is σ-sub-Gaussian. Then the information matrix can be written as

$$\begin{aligned}\mathcal{J}_t&=\sum_{s=1}^{t}[s^{-\alpha}\ s^{\gamma-\alpha}\ s^{\gamma-\alpha}p_s]^{\top}[s^{-\alpha}\ s^{\gamma-\alpha}\ s^{\gamma-\alpha}p_s]\\&=\begin{bmatrix}S_{-2\alpha,t}&S_{\gamma-2\alpha,t}&P_{\gamma-2\alpha,t}\\S_{\gamma-2\alpha,t}&S_{2\gamma-2\alpha,t}&P_{2\gamma-2\alpha,t}\\P_{\gamma-2\alpha,t}&P_{2\gamma-2\alpha,t}&Q_{2\gamma-2\alpha,t}\end{bmatrix}.\end{aligned} \tag{10}$$

In this scenario with three model parameters, the convergence rates of estimated parameters are dramatically different from each other. This adds difficulty to the design of a near-optimal policy. We now apply Algorithm 3, a modification of Algorithm 2 and prove the same optimal regret order.

Algorithm 3 Fixed-time Pricing Policy with an Intercept

Initialize $\lambda=1+u^2$, $\eta=\frac{\gamma+1}{3\gamma-2\alpha+2}<1$ and $l\leqslant l_0<u_0\leqslant u$. Set $c\in(0,1)$ and $T_0=cT^{\eta}<T$.
for $t=0,\cdots,T_0-1$ **do**
  Set $p_{t+1}=l_0\cdot\mathbb{1}\{t \text{ is even}\}+u_0\cdot\mathbb{1}\{t \text{ is odd}\}$
end for
for $t=T_0,\cdots,T-1$ **do**
  Let $\hat{\theta}_t=(\lambda I+\mathcal{J}_t)^{-1}\mathcal{D}_t$ be the (biased) least square estimator, where $\mathcal{J}_t$ is defined in (10), and

$$\mathcal{D}_t=\sum_{s=1}^{t}[s^{-\alpha}\ s^{\gamma-\alpha}\ s^{\gamma-\alpha}p_s]^{\top}d_s.$$

  Define confidence set

$$\mathcal{C}_t=\left\{\theta':\left|\theta'(i)-\hat{\theta}_t(i)\right|\leqslant\|e_i\|_{\mathcal{J}_t^{-1}}w_t,\forall i\right\},$$

  where w_t is defined as

$$\sigma\sqrt{2\log\left(\frac{1}{2}S_{\gamma,T}(1+\frac{S_{-2\alpha,t}}{\lambda})^{\frac{1}{2}}(1+S_{2\gamma-2\alpha,t})\right)}+\lambda^{\frac{1}{2}}\|\Theta\|. \tag{11}$$

  if $\mathcal{C}_t\cap\Theta\neq\varnothing$ **then**
    Set $(p_{t+1},\vartheta_t)=\arg\max_{p\in[l,u],\theta'\in\mathcal{C}_t\cap\Theta}p(a-bp)$.
  else
    Set an arbitrary price $p_{t+1}\in[l,u]$.
  end if
end for

To demonstrate the idea of our modification, we introduce the following lemma on confidence region, which is more delicate and is initially obtained in Abbasi-Yadkori et al. (2011).

Lemma 6 (Confidence Interval for any direction)**.** *For any*

$\delta \in (0,1)$, the following event happens w.p. at least $1-\delta$:

$$\forall t > 0, \forall x \in \mathbb{R}^3,$$
$$|x^\top \theta - x^\top \hat{\theta}_t| \leqslant \|x\|_{\mathcal{J}_t^{-1}} \cdot \left(\sigma \sqrt{2 \log \left(\frac{\det(\mathcal{J}_t)^{\frac{1}{2}} \det(\lambda I)^{-\frac{1}{2}}}{\delta} \right)} + \lambda^{\frac{1}{2}} \|\Theta\| \right).$$

In Algorithm 3, we modify the typical ellipsoid confidence region to a rectangle one such that rather than constructing a confidence bound for all parameters *jointly*, each parameter in θ is controlled *separately*. We have Theorem 4 that characterizes the near-optimality of Algorithm 3.

Theorem 4. *Fix γ and $\alpha \in [0, \gamma + \frac{1}{2})$. Then Algorithm 3 admits a regret $\sup_{\theta \in \Theta}\{R_\theta^\pi(T)\} = O(T^{\frac{(\gamma+1)^2}{3\gamma-2\alpha+2}} \log T)$ for all $T \geqslant 3$.*

The critical step in the proof of Theorem 4 is to establish that $\det(\mathcal{J}_{T_0}) = \Theta(S_{-2\alpha,t} S^2_{2\gamma-2\alpha,t})$ when $T_0 = cT^\eta$ and $T \to +\infty$. In other words, we asymptotically bound the growth rate of the information that is obtained at the end of the *pure exploration* phase.

We note that our policy can be further extended to a more complicated demand model that includes a steady sub-model:

$$d_t(p) = a_0 - b_0 p + t^\gamma (a - bp) + \epsilon_t,$$

where $\gamma > 0$, ϵ_t is $t^\alpha \sigma$-sub-Gaussian, and $\theta = (a_0, b_0, a, b)$ lies in a compact set $\Theta \in \mathbb{R}^4_+$. In this case, the optimal regret order in Section 3.3 still holds. The construction of the confidence region is almost the same with Algorithm 3. The only difference is that $\mathcal{J}_t$ becomes 4-dimension, and $\det(\mathcal{J}_t)$ is asymptotically approximated by $\Theta(S^2_{-2\alpha,t} S^2_{2\gamma-2\alpha,t})$, an additional $S_{-2\alpha,t}$ over $\Theta(S_{-2\alpha,t} S^2_{2\gamma-2\alpha,t})$. The proof procedure strictly follows our proof for Theorem 4, so we omit it here.

As a final remark, we have not completed the *any-time* policy design for (9), where the main difficulty is to control the information acquisition. In this situation, Lemma 3 becomes invalid, and the proof for Theorem 2 cannot be followed. Also, though *pure exploration* is useful when the growth rate of random fluctuation is relatively large, it may become useless for a near-optimal *any-time* policy. Results in previous literature on contextual bandits cannot be directly applied either because the reward in our problem is unbounded over the time. We leave it for future work.

4. Comparison with MAB Setting

For a dynamic pricing problem in a growing environment, we have shown that the form of optimal regret order is different depending on whether the growth rate of the random fluctuation is relatively large compared to the growth of the mean. We also have identified the difference in best achievable regret between *any-time* and *fixed-time* cases. In this section, we show that these differences do not emerge in a multi-armed bandit problem for which the mean and variance of the single-period reward grows over time.

Suppose that there are K arms $\{1, ..., K\}$, and the mean reward of arm i grows with time, i.e.,

$$r_t(i) = t^\gamma r(i) + \epsilon_t, \forall i \in [K]. \tag{12}$$

Here $\gamma \geqslant 0$, $\{\epsilon_t : t \geqslant 1\}$ are i.i.d. random variables with ϵ_t being $t^\alpha \sigma$-sub-Gaussian, and $r(i) \in [0,1]$ for all $i \in [K]$. Let $\theta = (r(1), ..., r(K))$, $r^* = \max_{i \in [K]} r(i)$ and $\Delta_t = r^* - r_t$. Then for any policy π, the T-period regret under policy π with true parameter θ can be naturally defined as

$$R_\theta^\pi(T) = \sum_{t=1}^{T} t^\gamma \mathbb{E}_\theta^\pi[\Delta_t]. \tag{13}$$

We first show that the lower bound of the MAB problem above is $\Omega(\sqrt{K} T^{\alpha+\frac{1}{2}})$ for general parameters α and γ. This is very different from Theorem 1.

Theorem 5. *Fix γ and $\alpha \in [0, \gamma + \frac{1}{2})$. Then there exists a constant C such that for any policy π, we have $\sup_{\theta \in [0,1]^K}\{R_\theta^\pi(T)\} \geqslant C\sqrt{K} T^{\alpha+\frac{1}{2}}$ for any T.*

The proof of Theorem 5 is similar to the proof for canonical MAB problem, where given any policy π, we construct two "close" instances that cannot be told apart by π.

Now we show that the lower bound in Theorem 5 can be achieved through Algorithm 4, an *any-time* pricing policy, which is a variant of Successive Elimination Algorithm, see, e.g., Even-Dar et al. (2006).

In Algorithm 4, we impose an order on the pulling of arms in each round. This is because the reward increases with time for general γ, and combining such constraint with action elimination will help bound the total reward of the pulled arm up to any given period easily. We note that the confidence bound constructed in time period t does not depend on the total periods T. Theorem 6 gives the near-optimality of Algorithm 4. Note that the constant does not depend on T or K. For simplicity, we leave how large T should be to our proof in the supplementary material.

Theorem 6. *Fix γ and $\alpha \in [0, \gamma + \frac{1}{2})$. Then Algorithm 4 admits a regret $\sup_{\theta \in [0,1]^K}\{R_\theta^\pi(T)\} = O(\sqrt{K} T^{\alpha+\frac{1}{2}} \log T)$ for all sufficiently large T.*

Now we discuss on some explanations on the differences presented in this paper. In dynamic pricing, why there is difference between *fixed-time* and *any-time* situation? In

Algorithm 4 Successive Elimination Algorithm

Let $\mathcal{A} = \{1, ..., K\}$ be the initial active set. Set $t = 1$.
repeat
  for each arm $i \in \mathcal{A}$ **do**
    Pull arm i and get reward r_t. The order of pulling are based on the index.
    Construct a confidence interval $\mathcal{C}(i) = [x_i, y_i]$ as

$$\left\{ r : \left| \frac{\sum_{s \leqslant t, A_s = i} s^{\gamma - 2\alpha}(r_s - s^{\gamma} r)}{\sqrt{\sum_{s \leqslant t, A_s = i} s^{2\gamma - 2\alpha}}} \right| \leqslant w_t \right\}, \tag{14}$$

    where $w_t = \sigma \sqrt{2 \left(1 + \frac{(\gamma+1-\alpha)(\gamma+1)}{\alpha + \frac{1}{2}}\right) \log t}$.
    $t \leftarrow t + 1$.
  end for
  for each arm $i \in \mathcal{A}$ **do**
    if $\exists j \in \mathcal{A} : y_i \leqslant x_j$ **then**
      $\mathcal{A} \leftarrow \mathcal{A} \setminus \{i\}$.
    end if
  end for
until $t > T$

general, why there is difference between dynamic pricing and MAB? One point is that regret in dynamic pricing is in the form of L_2 loss, which is a higher order compared to L_1 loss, so the regret order can be potentially smaller. This is exactly what happens in the case with large random fluctuation. Another point is as follows. When random fluctuation is large, it may be better to learn the model than simply conducting learning-while-doing at early stages. In early stages, the random fluctuation has not grown too large and the reward loss is relatively small, so one can acquire more information with less cost. However, to achieve this, a delicate tuning of the length of the learning phase is required, and this requires the knowledge of T in advance.

5. Conclusion and Future Work

In this work, we consider the problem of dynamic pricing in a specific non-stationary growing environment, where the mean and the level of random fluctuation in the demand process increase over time at possibly different rates. We construct new frameworks to prove best achievable performance in terms of regret and design new near-optimal policies. We show that the magnitudes of mean growth and random fluctuation growth, as well as their relative growth rate, significantly impact the best achievable performance and policy design. In our analysis, we demonstrate the intrinsic gap in optimal regret orders between the *fixed-time* scheme and the *any-time* scheme, differentiated by whether the decision maker knows the length of time horizon T in advance or not.

There is future theory work in line, that generalizes the demand model to accommodate more features that improve applicability, including multiple products and a broader range for price elasticity modeling. A further consideration is to include settings when the growth rates are also unknown and need to be learned from historical observations. Another future work is to integrate the structured non-stationarity and generic non-stationarity depending on various application needs.

Acknowledgement

The authors would like to express sincere gratitude to Yufan Chen, Peter W. Glynn and Xiaocheng Li for helpful discussions, and to the reviewers as well as the mata-reviewer for their enlightening comments and suggestions that have greatly improved this paper.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pp. 322–331. IEEE, 1995.

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

Auer, P., Gajane, P., and Ortner, R. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pp. 138–158, 2019.

Ban, G.-Y. and Keskin, N. B. Personalized dynamic pricing with machine learning: High dimensional features and heterogeneous elasticity. *Available at SSRN 2972985, accepted by Management Science*, 2020.

Besbes, O. and Zeevi, A. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

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

Besbes, O., Gur, Y., and Zeevi, A. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances*

in neural information processing systems, pp. 199–207, 2014.

Besson, L. and Kaufmann, E. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

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

Bu, J., Simchi-Levi, D., and Xu, Y. Online pricing with offline data: Phase transition and inverse square law. *arXiv preprint arXiv:1910.08693*, 2019.

Chen, N. and Gallego, G. A primal-dual learning algorithm for personalized dynamic pricing with an inventory constraint. *Available at SSRN 3301153*, 2018.

Chen, Y., Lee, C.-W., Luo, H., and Wei, C.-Y. A new algorithm for non-stationary contextual bandits: Efficient, optimal, and parameter-free. *arXiv preprint arXiv:1902.00980*, 2019.

Cheung, W. C., Simchi-Levi, D., and Wang, H. Dynamic pricing and demand learning with limited price experimentation. *Operations Research*, 65(6):1722–1731, 2017.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Hedging the drift: Learning to optimize under non-stationarity. *Available at SSRN 3261050*, 2018.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. *Proceedings of the 21st Conference on Learning Theory*, 2008.

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

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

Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105, 2006.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pp. 586–594, 2010.

Glynn, P. W. and Zheng, Z. Estimation and inference for non-stationary arrival models with a linear trend. In *2019 Winter Simulation Conference (WSC)*, pp. 3764–3773, 2019.

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

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

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

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press (preprint), 2019.

Li, X. and Zheng, Z. Dynamic pricing with external information and inventory constraint. *Available at SSRN*, 2019.

Nambiar, M., Simchi-Levi, D., and Wang, H. Dynamic learning and pricing with model misspecification. *Management Science*, 65(11):4980–5000, 2019.

Qiang, S. and Bayati, M. Dynamic pricing with demand covariates. *Available at SSRN 2765257*, 2016.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35 (2):395–411, 2010.

Slivkins, A. et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.