Online Pricing with Offline Data: Phase Transition and Inverse Square Law

Jinzhi Bu [1] David Simchi-Levi [1] Yunzong Xu [1]

Abstract

This paper investigates the impact of pre-existing offline data on online learning, in the context of dynamic pricing. We study a single-product dynamic pricing problem over a selling horizon of T periods. The demand in each period is determined by the price of the product according to a linear demand model with unknown parameters. We assume that the seller already has some pre-existing offline data before the start of the selling horizon. The seller wants to utilize both the pre-existing offline data and the sequential online data to minimize the regret of the online learning process. We characterize the joint effect of the size, location and dispersion of the offline data on the optimal regret of the online learning process. Our results reveal surprising transformations of the optimal regret rate with respect to the size of the offline data, which we refer to as phase transitions. In addition, our results demonstrate that the location and dispersion of the offline data also have an intrinsic effect on the optimal regret, and we quantify this effect via the inverse-square law.

1. Introduction

Classical statistical learning theory distinguishes between offline and online learning. Offline learning deals with a setting where the entire training data set is directly available before the algorithm is applied, while online learning deals with a setting where data become available in a sequential manner that may depend on the actions taken by the algorithm. While offline learning assumes access to offline data (but not online data) and online learning assumes access to online data (but not offline data), in reality, a broad class of real-world problems incorporate both aspects: there is an offline historical data set (based on historical actions) at the time that the learner starts an online learning process.

Currently, there is no standard framework for the above type of learning problems, as classical offline learning theory and online learning theory have different settings and goals. While establishing a framework that bridges all aspects of offline and online learning is generally a very complicated task, in this paper, we propose a framework that bridges the gap between offline and online learning in a specific problem setting, which, however, already captures the essence of many dynamic pricing problems that sellers face in practice.

The problem: online pricing with offline data. We study the "Online Pricing with Offline Data" (OPOD) problem as follows. Consider a seller offering a single product with an infinite amount of inventory over a selling horizon of T periods. Customer demand is determined by the price charged by the seller according to an underlying linear demand model. The seller knows neither the true demand parameters nor the distribution of random noise. However, we assume that before the selling horizon starts, she has used some *historical prices* in the past and collected market sales data. In other words, the seller has a pre-existing offline data set before the start of the online learning process. We assume that the pre-existing offline data set contains n samples: $\{(\hat{p}_1, \hat{D}_1), \ldots, (\hat{p}_n, \hat{D}_n)\}$, where each sample $(\hat{p}_i, \hat{D}_i)$ is an input-output pair consisting of a historical price $\hat{p}_i$ and an associated demand observation $\hat{D}_i$ generated from the same linear demand model ($i = 1, \ldots, n$). The seller's objective is to design a learning algorithm that utilizes both the offline data and the data collected on-the-fly to learn the unknown demand model while concurrently maximizing total revenue over T periods.

Following the convention of online learning, we measure the performance of an algorithm for the aforementioned OPOD problem by *regret*, which is the difference between the optimal expected revenue and the total expected revenue generated by the algorithm over T periods in the online stage. A notable difference between the regret defined here and the classical one is that the regret of the OPOD problem depends on the pre-existing offline data. We refer to the best achievable regret (which depends on offline data) as the optimal regret of the OPOD problem.

At a high level, we seek to answer the following question: *How do offline data affect the statistical complexity of on-*

[1] Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139. Correspondence to: Jinzhi Bu <jzbu@mit.edu>, David Simchi-Levi <dslevi@mit.edu>, Yunzong Xu <yxu@mit.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

line learning? We investigate this fundamental question in the OPOD framework, where the statistical complexity is measured by the optimal regret of the problem.

Main contributions. We identify that the size, location and dispersion of the offline data have an intrinsic effect on the optimal regret, where the size is measured by the number of offline samples n, the location is measured by the distance between the average historical price and the true optimal price δ, and the dispersion is measured by the the standard deviation of the historical prices σ. Note that δ is an instance-dependent quantity that **uniquely** appears when offline learning and online learning are combined, as it quantifies the relationship between the offline decisions and the optimal online decision. We prove that the optimal regret is $\widetilde{\Theta}(\sqrt{T} \wedge \frac{T}{(n \wedge T)\delta^2 + n\sigma^2})$, except for a corner case of $\delta^2 \lesssim \frac{1}{n\sigma^2} \lesssim \frac{1}{\sqrt{T}}$, where the optimal regret becomes $\widetilde{\Theta}(T\delta^2)$. We summarize the order of the optimal regret under different combinations of (n, σ, δ) in Table 1, where the formal definitions of "$\widetilde{\Theta}(\cdot)$", "$\lesssim$", "$\gtrsim$" are provided at the end of this section. We highlight that all the regret bounds derived in this paper are *non-asymptotic* finite-time bounds that hold uniformly over T. We emphasize the following technical highlights.

(1) We prove a instance-dependent lower bound on the optimal regret by reducing the OPOD problem to a hybrid of estimation and hypothesis testing problems. Our lower bound is stronger and harder to prove than the traditional "minimax" lower bound. See §4.

(2) We propose a *parameter-free* and *anytime* algorithm that achieves the optimal regret without knowing δ and T, based on the Optimism in the Face of Uncertainty (OFU) principle. While the OFU principle is well-studied in the linear bandit literature, previous analysis fails to provide an instance-dependent upper bound in our setting as our revenue function is quadratic and our optimal price is an interior point. We overcome this methodological challenge by conducting a period-by-period trajectory analysis. We present novel inductive arguments to show that under a good initialization, the random trajectory of the distance between the optimistic price and the average historical price admits a high-probability global lower bound. See §5.

(3) Combining (1) and (2) we obtain a tight instance-dependent bound on the optimal regret. To the best of our knowledge, this is the first "beyond the worst-case" instance-dependent regret bound obtained in (i) the dynamic learning and pricing literature, and (ii) a continuous-armed bandit problem where the optimal action may not be an extremal point (in contrast to the extremal-point requirement in Dani et al. 2008 and Abbasi-Yadkori et al. 2011).

Our tight characterization of the optimal regret leads to important practical implications on the value of offline data. Our results reveal significant transitions between the regret-decaying patterns when the size of the offline data changes, which we refer to as *phase transitions*. In addition, our results demonstrate that the optimal regret is inversely proportional to the square of the quantities δ and σ, which is referred to as the *inverse-square law*. See §6.

We also conduct computational experiments to validate our theoretical results. See §7.

Table 1. Optimal regret as a function of (n, σ, δ)

$\delta \gtrsim T^{-\frac{1}{4}}$ and $\sigma \lesssim \delta$				
size	$0 \le n \lesssim \frac{\sqrt{T}}{\delta^2}$	$\frac{\sqrt{T}}{\delta^2} \lesssim n \lesssim T$	$T \lesssim n \lesssim \frac{T\delta^2}{\sigma^2}$	$n \gtrsim \frac{T\delta^2}{\sigma^2}$
opt. reg.	$\widetilde{\Theta}(\sqrt{T})$	$\widetilde{\Theta}(\frac{T}{n\delta^2})$	$\widetilde{\Theta}(\frac{1}{\delta^2})$	$\widetilde{\Theta}(\frac{T}{n\sigma^2})$
$\delta \gtrsim T^{-\frac{1}{4}}$ and $\sigma \gtrsim \delta$				
size	$0 \le n \lesssim \frac{\sqrt{T}}{\sigma^2}$		$n \gtrsim \frac{\sqrt{T}}{\sigma^2}$	
opt. reg.	$\widetilde{\Theta}(\sqrt{T})$		$\widetilde{\Theta}(\frac{T}{n\sigma^2})$	
$\delta \lesssim T^{-\frac{1}{4}}$				
size	$0 \le n \lesssim \frac{\sqrt{T}}{\sigma^2}$		$\frac{\sqrt{T}}{\sigma^2} \lesssim n \lesssim \frac{1}{\delta^2\sigma^2}$	$n \gtrsim \frac{1}{\delta^2\sigma^2}$
opt. reg.	$\widetilde{\Theta}(\sqrt{T})$		$\widetilde{\Theta}(T\delta^2)$	$\widetilde{\Theta}(\frac{T}{n\sigma^2})$

Notations and remarks. Throughout the paper, all the vectors are column vectors unless otherwise specified. For each $m \in \mathbb{N}$, we use $[m]$ to denote the set $\{1, 2, \dots, m\}$. For any $x = (x_1, \dots, x_n) \in \mathbb{R}^n$, we use $|| \cdot ||$ to denote the l^2 norm, i.e., $||x|| = (\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}}$. We use $a \wedge b$ to denote $\min\{a, b\}$ and $a \vee b$ to denote $\max\{a, b\}$. The notation $f(T) = O(g(T))$ represents that there exists a known constant C independent of the problem instance such that $f(T) \le Cg(T)$ for all $T > 0$. Similarly, $\Omega(\cdot)$ and $\Theta(\cdot)$ are used by hiding constant factors, and $\widetilde{O}(\cdot)$, $\widetilde{\Omega}(\cdot)$ and $\widetilde{\Theta}(\cdot)$ are used by hiding constant factors and logarithmic factors. The notations $A \lesssim B$ and $A \gtrsim B$ represent $A = \widetilde{O}(B)$ and $A = \widetilde{\Omega}(B)$ respectively.

Finally, we remark that the full version of this paper (containing additional theoretical results, computational experiments, and missing proofs) is available at https://arxiv.org/abs/1910.08693.

2. Related Literature

Dynamic pricing with online learning. Dynamic pricing with online learning has generated great interest in recent years, see den Boer (2015) for a comprehensive survey. Among all the models studied in this area, the problem of "dynamic pricing with an unknown linear demand model" (DP-ULD) is possibly the most classical and basic one. In this problem, the demand is determined by the price according to a linear demand model with unknown parameters and random noise. The DP−ULD problem is first analyzed in Broder & Rusmevichientong (2012), den Boer & Zwart (2013), Keskin & Zeevi (2014), with $\widetilde{\Theta}(\sqrt{T})$ minimax regret identified and various algorithms proposed. It then becomes a "building block" for a vast dynamic pricing literature studying more complicated generalizations, e.g., den Boer (2014), Besbes & Zeevi (2015), Keskin & Zeevi (2016), Qiang & Bayati (2016), den Boer & Keskin (2017), Nambiar et al. (2019), Ban & Keskin (2019), Bastani et al. (2019). All of the existing papers focus on the pure online learning setting. In this paper, we take the classical DP−ULD problem as our baseline, but significantly extend it by incorporating offline data into online pricing.

The paper by Keskin & Zeevi (2014) is the most relevant to ours. Starting from the classical DP−ULD model, they study whether knowing the *true* expected demand under a single price in advance helps reduce the regret. Depending on whether the seller has this knowledge or not, they prove that the best achievable regret is $\Theta(\log T)$ and $\widetilde{\Theta}(\sqrt{T})$ respectively. Compared with their work, the OPOD problem studied in our paper seems more relevant to practice, and is more general in theory. Practically, while firms will never know the true expected demand under a given price exactly (which requires infinitely many demand observations), they usually have some pre-existing historical data (which are finitely many) before the online learning starts. Theoretically, the results in Keskin & Zeevi (2014) can be viewed as a special case of our results when $\sigma = 0, n = 0$ or $\sigma = 0, n = \infty$. In addition, for the case of $\sigma = 0, n = \infty$, they make a strong assumption that δ is lower bounded by a *known* constant, and use this constant in their algorithm. Since δ is completely unknown and can be arbitrarily small in reality (and in our problem), their algorithms and analysis cannot be extended to our setting.

Multi-armed bandits. Our paper is also related to the literature of multi-armed bandits (MAB). In most of the literature in MAB, the decision maker is assumed to start with no data available before she sequentially pulls the arms. By contrast, a few papers study bandit problems in settings where the algorithms may utilize different types of historical information, see, e.g., Shivaswamy & Joachims (2012), Bouneffouf et al. (2019), Hsu et al. (2019), Gur & Momeni (2019), Ye et al. (2020), of which Shivaswamy & Joachims (2012) is the most relevant to this paper.

Shivaswamy & Joachims (2012) study the MAB problem with offline observations of rewards collected before the online learning algorithm starts. While they share similar spirits with us in incorporating offline data and quantifying how they affect the regret, there are significant differences between the two papers in model settings, main results and analytical techniques. First, Shivaswamy & Joachims (2012) study the MAB problem with discrete and finitely many arms. All their results rely on the finite-armed property and cannot be extended to our setting where there are continuous and infinitely many prices. Second, under the so-called *well-separated condition*, Shivaswamy & Joachims (2012) prove some regret upper bounds that change from $O(\log T)$ to $O(1)$ when the amount of offline observations of rewards for *each* arm exceeds $\Omega(\log T)$. Unfortunately, the authors provide no regret lower bound, and hence no tight characterization on the value of offline data. In comparison, we characterize the optimal regret via matching upper and lower bounds, and figure out surprising phase transitions as well as the elegant inverse square law. Third, Shivaswamy & Joachims (2012) use a conventional approach in bandit literature to upper-bound the regret via the reward gap between the best and second best actions, while we are bounding the regret via δ, σ and n. As a result, we present different regret analysis that may be of independent interest.

3. Model Formulation

Basic model. Consider a firm selling a single product with infinite amount of inventory over a selling horizon of T periods. In each period $t \in [T] = \{1, \ldots, T\}$, the seller chooses a price p_t from a given interval $[l, u] \subset [0, \infty)$ to offer its customers, and then observes the random demand D_t for that period t. In this paper, we focus on the canonical linear demand model: the demand in each period is a linear function of the price plus some random noise. Specifically, for each $t \in [T]$,

$$D_t = \alpha^* + \beta^* p_t + \epsilon_t, \tag{1}$$

where α^* and β^* are two unknown demand parameters in the known interval $[\alpha_{\min}, \alpha_{\max}] \subseteq (0, \infty)$ and $[\beta_{\min}, \beta_{\max}] \subseteq (-\infty, 0)$ respectively, and $\{\epsilon_t : t \geq 1\}$ are *i.i.d.* random variables with zero mean and the unknown generic distribution $\epsilon \sim \mathcal{D}$. We assume that ϵ is an R^2-sub-Gaussian random variable, i.e., there exists a constant $R > 0$ such that $\mathbb{E}[e^{x\epsilon}] \leq e^{\frac{x^2 R^2}{2}}$ for any $x \in \mathbb{R}$, and use $\mathcal{E}(R)$ to denote the class of all R^2-sub-Gaussian random variables. To simplify notation, let $\theta^* = (\alpha^*, \beta^*)$, $\Theta^\dagger = [\alpha_{\min}, \alpha_{\max}] \times [\beta_{\min}, \beta_{\max}]$, and $\theta = (\alpha, \beta)$ be any vector in the set $\Theta^\dagger$.

Given the demand parameter θ, the seller's single-period expected revenue is defined as $r(p; \theta) = p(\alpha + \beta p)$, $\forall p \in [l, u]$ Let $\psi(\theta)$ be the price that maximizes the expected revenue $r(p; \theta)$ over the interval $[l, u]$, i.e., $\psi(\theta) = \arg\max\{r(p; \theta) : p \in [l, u]\}$. In particular, we also use p^* to denote the underlying true optimal price, i.e., $p^* = \psi(\theta^*)$. Let $r^*(\theta)$ be the optimal expected revenue under the demand parameter θ, i.e., $r^*(\theta) = \psi(\theta)(\alpha + \beta\psi(\theta))$.

Without loss of generality, we assume that for any $\theta \in \Theta^\dagger$, the optimal price under θ is an interior point of the feasible set $[l, u]$, and $\psi(\theta) = \frac{\alpha}{-2\beta}$. This is because for any $\theta \in \Theta^{\dagger}$, $\frac{\alpha}{-2\beta} \in [\frac{\alpha_{\min}}{-2\beta_{\max}}, \frac{\alpha_{\max}}{-2\beta_{\min}}]$, and we can choose l and u such that the range of $\frac{\alpha}{-2\beta}$ belongs to $[l, u]$, which guarantees that $\frac{\alpha}{-2\beta}$ is an interior point of the interval $[l, u]$.

Offline data. In reality, the seller does not know the exact values of α^* and β^*, but has access to some pre-existing offline data before the start of the online learning process. We assume that the offline data set contains n independent samples: $\{(\hat{p}_1, \hat{D}_1), \ldots, (\hat{p}_n, \hat{D}_n)\}$, where $\hat{p}_1, \ldots, \hat{p}_n$ are n fixed prices, and for each $i \in [n]$, $\hat{D}_i$ is a demand sample under the historical price $\hat{p}_i$, drawn independently according to the underlying linear demand model (1). The seller can use the offline data as well as the data generated on-the-fly to estimate the unknown parameters and maximize the revenue. We refer to this problem as the online pricing with offline data (`OPOD`) problem.

Let $\bar{p}_{1:n}$ be the average historical price, i.e., $\bar{p}_{1:n} = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_i$, and σ be the standard deviation of the historical prices, i.e., $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - \bar{p}_{1:n})^2}$. In addition, let δ be $|\bar{p}_{1:n} - p^*|$. Intuitively, the quantity δ measures how far the offline data set is away from the true (unknown) optimal price p^*, and is referred to as the *generalized distance*.

Pricing policies and performance metrics. For each $t \geq 0$, let H_t be the vector of information available at the end of period t (or at the beginning of period $t+1$), i.e., $H_t = (\hat{p}_1, \hat{D}_1, \ldots, \hat{p}_n, \hat{D}_n, p_1, D_1, \ldots, p_t, D_t)$. A pricing policy is defined as a sequence of functions $\pi = (\pi_1, \pi_2, \ldots)$, where $\pi_t : \mathbb{R}^{2n+2t-2} \to [l, u]$ is a measurable function which maps the information vector H_{t-1} to a feasible price. Therefore, the policy π generates a price sequence $(p_1, p_2, \ldots)$ with each p_t adapted to H_{t-1}.

Let Π be the set of all policies. For any policy $\pi \in \Pi$, let $\rho_{\theta}^{\pi}(T)$ be the T-period expected revenue, i.e., $\rho_{\theta}^{\pi}(T) = \mathbb{E}_{\theta}^{\pi}\left[\sum_{t=1}^{T} r(p_t; \theta)\right]$, where $\mathbb{E}_{\theta}^{\pi}[\cdot]$ is the expectation induced by the policy π when the demand parameter is θ. The *regret* of the policy π is defined as

$$R_{\theta}^{\pi}(T) = Tr^*(\theta) - \rho_{\theta}^{\pi}(T),$$

which is the gap between the expected revenues generated by the clairvoyant policy that knows the exact value of θ and the pricing policy π.

Optimal regret. As shown in the existing literature, even when there is no offline data available, many simple and pure-online policies, e.g., the explore-then-commit policy in Broder & Rusmevichientong (2012) and the semi-myopic policies in Keskin & Zeevi (2014), already guarantee $\tilde{O}(\sqrt{T})$ regret for all $\theta \in \Theta^{\dagger}$. Intuitively, in the presence of the offline data, the performance of a "reasonable" pricing policy should be as good as, if not better than the regret when there were no offline data available, and hence, for any demand parameter $\theta \in \Theta^{\dagger}$, the regret of such pricing policies should not exceed $\tilde{\Theta}(\sqrt{T})$. Therefore, in this paper, we restrict our attention to such "reasonable" policies, and more formally, we define the class of *admissible* policies as:

$$\Pi^{\circ} = \left\{\pi \in \Pi : \ \forall \theta \in \Theta^{\dagger}, \ R_{\theta}^{\pi}(T) \leq K_0\sqrt{T}\log T\right\},$$

where $K_0 > 0$ is an arbitrary constant. For any admissible policy $\pi \in \Pi^{\circ}$, the *instance-dependent regret* is defined as

$$R^{\pi}(T, n, \sigma, \delta) = \sup_{\substack{\mathcal{D} \in \mathcal{E}(R); \\ \theta \in \Theta^{\dagger} : |\psi(\theta) - \bar{p}_{1:n}| \in [(1-\xi)\delta, (1+\xi)\delta]}} R_{\theta}^{\pi}(T),$$

where $\xi \in (0, 1)$ is an arbitrary constant. In other words, we consider the environment class as the set of all possible problem instances such that the distance between the associated optimal price and the average historical price has the same order as the generalized distance δ.[1] The *optimal (instance-dependent) regret* is defined as

$$R^*(T, n, \sigma, \delta) = \inf_{\pi \in \Pi^{\circ}} R^{\pi}(T, n, \sigma, \delta). \tag{2}$$

We use the optimal regret to measure the statistical complexity of the `OPOD` problem. We refer the reader to the full version of this paper for more discussions on the definition of the optimal regret.

4. Lower Bound

In this section, we establish a lower bound on the optimal regret.

Theorem 1. *Suppose* $\frac{u^2 R^2}{256 K_0 \beta_{\min}^2 e} > \frac{3}{2}$*, then for any admissible policy* $\pi \in \Pi^{\circ}$*,* $T \geq 2$*,* $n \geq 1$*,* $\delta \in [0, u-l]$*, and* $\sigma \in [0, u-l]$*,*

$$R^{\pi}(T, n, \sigma, \delta) = \begin{cases} \Omega(T\delta^2), & \textit{if } \delta^2 \lesssim \frac{1}{n\sigma^2} \lesssim \frac{1}{\sqrt{T}}; \\ \Omega\left(\frac{\sqrt{T}}{\log T} \wedge \frac{T}{(n \wedge T)\delta^2 + n\sigma^2}\right), & \textit{otherwise}. \end{cases}$$

Theorem 1 distinguishes the lower bound on the optimal regret under two cases. We call the case $\delta^2 \lesssim \frac{1}{n\sigma^2} \lesssim \frac{1}{\sqrt{T}}$ as the "*corner case*", and its complement as the "*regular case*". We remark that the corner case rarely happens because it requires that the generalized distance δ is very small and the dispersion $n\sigma^2$ is very large, such that there is no need of online learning, see the discussion in §5.2.

When all the historical prices are identical, i.e., $\sigma = 0$, we can actually prove an additional lower bound $\Omega(\frac{\log T}{\delta^2})$.

[1] Since the objective of this paper is to study the impact of the order of the generalized distance δ on the complexity of the online learning process, allowing $|\psi(\theta) - \bar{p}_{1:n}|$ to perturb around δ within a constant factor does not affect this goal.

Therefore, when $n = 0$, or $\sigma = 0$, $n = \infty$ and δ is a constant, our results recover both Theorem 1 and Theorem 3 in Keskin & Zeevi (2014).

We next outline the key idea in the proof of Theorem 1 (the detailed proof can be found in the full version of this paper). Our proof consists of two steps. First, by relating the regret to $\sum_{t=1}^{T} \mathbb{E}_{\theta}^{\pi}[(p_t - \psi(\theta))^2]$, we consider an "auxiliary" estimation problem for the optimal price $\psi(\theta)$, and appeal to the multivariate van Trees inequality, which is a Bayesian version of the Cramér-Rao bound. By choosing a suitable instance-dependent prior distribution of θ, and decomposing the Fisher information into a function of the regret, we prove the regret lower bound $\Omega(\sqrt{T} \wedge \frac{T}{\delta^{-2}+n\sigma^2+(n\wedge T)\delta^2})$ for any pricing policy. When $\delta^2 \gtrsim T^{-\frac{1}{2}}$, or $\delta^2 \lesssim T^{-\frac{1}{2}}$ and $n\sigma^2 \gtrsim \sqrt{T}$, Theorem 1 is implied from this bound. Second, for the remaining case when $\delta^2 \lesssim T^{-\frac{1}{2}} \lesssim \frac{1}{n\sigma^2}$, by defining an instance-dependent hypothesis testing problem, we use the KL divergence arguments to prove the regret lower bound $\Omega(\frac{\sqrt{T}}{\log T})$ for any admissible policy $\pi \in \Pi^{\circ}$.

Although the van Trees inequality and KL divergence arguments are also used in Keskin & Zeevi (2014) and Broder & Rusmevichientong (2012) respectively for a similar DP-ULD problem, our result and analysis are different from theirs in several aspects. First, compared with Keskin & Zeevi (2014), we need to carefully choose an instance-dependent prior distribution of θ that depends on the generalized distance δ, such that its Fisher information is in the order of $\Theta(\delta^{-2})$. By contrast, in Keskin & Zeevi (2014), it is only required to define an instance-independent prior with constant Fisher information. In addition, in the van Trees inequality, we also need to decompose the quadratic form of the Fisher information matrix to incorporate the effects of the offline data into the regret bound, while it is unnecessary in Keskin & Zeevi (2014) since they do not assume the existence of offline data. Second, compared with Broder & Rusmevichientong (2012), our result is stronger in the sense that even with the help of the offline data, when $\delta^2 \lesssim T^{-\frac{1}{2}} \lesssim \frac{1}{n\sigma^2}$, the regret is still lower bounded by $\widetilde{\Omega}(\sqrt{T})$ for any admissible policy. Moreover, since we need to show an instance-dependent lower bound, we have to define an instance-dependent hypothesis set, decompose the KL divergence using the offline data, and also leverage the property of $\pi \in \Pi^{\circ}$.

5. Algorithms and Upper Bounds

In §5.1, we first propose the "Online and Offline Optimism in the Face of Uncertainty" (O3FU) algorithm, and prove that it matches the lower bound on regret for the regular case. In §5.2, we present the Tweaked O3FU (T-O3FU) algorithm, which is built upon the O3FU algorithm and matches the lower bound for both the regular and corner cases.

5.1. O3FU Algorithm: Optimal in the Regular Case

The first algorithm O3FU is constructed based on the principle of "Optimism in the Face of Uncertainty", and the pseudo-code is given in the following Algorithm 1. We define

$$w_t = R\sqrt{2\log\Big(\frac{1}{\epsilon_t}\big(1+(1+u^2)(t+n)/\lambda\big)\Big)} + \sqrt{\lambda(\alpha_{\max}^2+\beta_{\min}^2)}, \tag{3}$$

where the constants ϵ_t and λ are algorithm input.

Algorithm 1 O3FU Algorithm

```
**Input:** offline data $\{(\hat{p}_1, \hat{D}_1), \ldots, (\hat{p}_n, \hat{D}_n)\}$, support of demand parameters $\Theta^{\dagger}$, support of feasible price $[l, u]$, regularization parameter $\lambda = 1 + u^2$, $\{w_t : t \geq 1\}$ defined in (3) with $\epsilon_t = \frac{1}{t^2} \wedge \frac{1}{n\sigma^2}$.
**Initialization:** $V_{0,n} = \lambda I + \sum_{i=1}^{n}[1\ \hat{p}_i]^{\intercal}[1\ \hat{p}_i]$, $Y_{0,n} = \sum_{i=1}^{n} \hat{D}_i[1\ \hat{p}_i]^{\intercal}$.
**for** $t = 1$ **to** $T$ **do**
  **if** $t = 1$, or $t > 1$ and $\mathcal{C}_{t-1} \cap \Theta^{\dagger} = \emptyset$ **then**
    Let $p_t = l\mathbb{I}\{\bar{p}_{1:n} > \frac{u+l}{2}\} + u\mathbb{I}\{\bar{p}_{1:n} \leq \frac{u+l}{2}\}$.
  **end if**
  **if** $t > 1$ and $\mathcal{C}_{t-1} \cap \Theta^{\dagger} \neq \emptyset$ **then**
    Let $(p_t, \tilde{\theta}_t) = \arg\max_{p\in[l,u],\theta\in\mathcal{C}_{t-1}\cap\Theta^{\dagger}} p(\alpha + \beta p)$.
  **end if**
  Charge the price $p_t$ and observe the demand $D_t$;
  Update $V_{t,n} = V_{t-1,n} + [1\ p_t]^{\intercal}[1\ p_t]$, $Y_{t,n} = Y_{t-1,n} + D_t[1\ p_t]^{\intercal}$, $\hat{\theta}_t = V_{t,n}^{-1}Y_{t,n}$;
  Update $\mathcal{C}_t = \big\{\theta \in \mathbb{R}^2 : ||\theta - \hat{\theta}_t||_{V_{t,n}} \leq w_t\big\}$.
**end for**
```

As a result of the OFU principle, the O3FU algorithm achieves the balance between exploration and exploitation by maximizing the "optimistic revenue" $\max_{\theta\in\mathcal{C}_{t-1}\cap\Theta^{\dagger}} p(\alpha + \beta p)$, which can be thought of as the sum of the estimated revenue and a "bonus" of exploration. Th O3FU algorithm is *parameter-free* in the sense that it does not need to use any information about the unknown δ. It is also an *anytime* algorithm since it does not need to know the length of the selling horizon T.

The following theorem provides a per-instance upper bound on the regret of the O3FU algorithm.

Theorem 2. *Let π be the O3FU algorithm. Then for any $T \geq 1$, $n \geq 1$, $\sigma \geq 0$, $\bar{p}_{1:n} \in [l, u]$, and $\theta \in \Theta^{\dagger}$,*

$$R_{\theta}^{\pi}(T) = O\big(\sqrt{T}(\log T) \wedge \frac{T(\log T)^2}{(n \wedge T)(\bar{p}_{1:n} - \psi(\theta))^2 + n\sigma^2}\big).$$

Compared with Theorem 1, the upper bound in Theorem 2 matches with the lower bound up to a logarithmic factor in the regular case. Therefore, in the regular case, our O3FU algorithm is optimal, and the optimal regret in this case is $R^*(T,n,\delta,\sigma)=\widetilde{\Theta}\big(\sqrt{T}\wedge\frac{T}{(n\wedge T)\delta^2+n\sigma^2}\big)$.

To show Theorem 2, we need to prove both an instance-independent upper bound $O(\sqrt{T}\log T)$ and an instance-dependent upper bound $O(\frac{T(\log T)^2}{n\sigma^2+(n\wedge T)\delta^2})$. The first bound can be easily proved by treating our problem as a linear bandit and applying arguments in the regret analysis therein, e.g., Abbasi-Yadkori et al. (2011). To prove the second bound, our analysis relies on the following crucial lemma.

Lemma 1. *Suppose* $\sigma\le|\psi(\theta)-\bar{p}_{1:n}|$, $|\psi(\theta)-\bar{p}_{1:n}|\ge\max\{\frac{\sqrt{2(\alpha_{\max}^2+\beta_{\max}^2)}}{\beta_{\max}^2}\frac{T^{1/4}w_T}{n^{1/2}},\sqrt{C}_1T^{-1/4}\}$, *and* $\theta\in\mathcal{C}_t$ *for each* $t\in[T]$, *then two sequences of events* $\{U_{t,1}:t\ge1\}$ *and* $\{U_{t,2}:t\ge2\}$ *also hold, where*

$$U_{t,1}=\Big\{|p_t-\bar{p}_{1:n}|\ge\min\big\{1-\frac{\sqrt{2}}{2},\frac{C_2}{2}\big\}\cdot|\psi(\theta)-\bar{p}_{1:n}|\Big\},$$
$$U_{t,2}=\Big\{||\theta-\tilde{\theta}_t||^2\le\frac{w_{t-1}^2C_3}{(n\wedge(t-1))(\psi(\theta)-\bar{p}_{1:n})^2+n\sigma^2}\Big\},$$

where C_1, C_2 *and* C_3 *are constants.*

Lemma 1 indicates that under the O3FU algorithm, the algorithm's pricing sequence is uniformly bounded away from the average historical price $\bar{p}_{1:n}$ proportional to the unknown quantity $|\psi(\theta)-\bar{p}_{1:n}|$ (as implied by $\{U_{t,1}\}_{t=1}^T$), and will gradually approach $\phi(\theta)$ in an l^2-rate of $O(\frac{\log t}{(n\wedge t)(\psi(\theta)-\bar{p}_{1:n})^2+n\sigma^2})$ (as implied by $\{U_{t,2}\}_{t=1}^T$), with high probability. This implies that the algorithm can "automatically" explore to a suitable degree, to create an efficient "collaboration" between the online prices and historical prices, while concurrently approaching the unknown optimal price. This property is nontrivial and cannot be implied from the existing analysis of the OFU-type algorithms. To prove this crucial lemma, we conduct a period-by-period trajectory analysis of the random pricing sequence generated by our algorithm. Specifically, we find that the occurrence of $U_{t,2}$ relies on the joint occurrence of $U_{1,1},\ldots,U_{t-1,1}$, while the occurrence of $U_{t,2}$ (combined with the specific structure of the optimistic revenue curve) in turn leads to the occurrence of $U_{t,1}$. We thus introduce novel induction-based arguments to prove Lemma 1, see details in the full version of this paper.

In the linear bandits, Abbasi-Yadkori et al. (2011) proved an instance-dependent upper bound $O(\frac{(\log T)^2}{\Delta})$ by assuming a polytope action set, with Δ defined as the sub-optimality gap between the rewards of the best and second best extremal points. We emphasize that their analysis cannot be applied to our setting. The property that guarantees the instance-dependent bound in Abbasi-Yadkori et al. (2011) comes from the fact that their algorithm only selects actions among the extremal points of the action set, which only holds under their assumptions. Our problem, however, has a quadratic objective function, with the optimal price being an interior point of the interval $[l,u]$. As a result, the sub-optimality gap becomes zero, and standard arguments do not work.

5.2. T-O3FU Algorithm: Optimal in All Cases

While the O3FU algorithm is optimal (up to a logarithm factor) in the regular case, it achieves $\widetilde{O}(\frac{T}{n\sigma^2})$ regret in the corner case, which is slightly looser than the $\widetilde{\Omega}(T\delta^2)$ lower bound in Theorem 1. To close this small gap, we first make several observations on the corner case condition.

(i) When $\frac{1}{n\sigma^2}\lesssim\frac{1}{\sqrt{T}}$ and $\delta^2\lesssim\frac{1}{\sqrt{T}}$, the offline data provides so much information such that there is no need of exploration in the online stage. Indeed, the least-square estimation error from the offline data is within $\widetilde{O}(\frac{1}{n\sigma^2})$. Thus, by simply charging the fixed myopic price calculated based on the offline data in every online period, we achieve $\widetilde{O}(\frac{T}{n\sigma^2})$ regret. In fact, this rate of estimation error cannot be improved in the online process by any admissible policy. Therefore, if the algorithm knows that it is in the corner case, then there is no need of exploration.

(ii) In the extreme case of $\delta^2\lesssim\frac{1}{n\sigma^2}$, if an algorithm knows that it is in the corner case, then by simply charging the average historical price $\bar{p}_{1:n}$ in every online period, it achieves $\widetilde{O}(T\delta^2)$ regret for all instances such that $|\psi(\theta)-\bar{p}_{1:n}|=\Theta(\delta)$, which is even better than $\widetilde{O}(\frac{T}{n\sigma^2})$.

(iii) However, since the algorithm does not know the value of δ and may not know the selling horizon T in advance, it does not know whether it is in the corner case. If the conditions in (i) do not hold, then the algorithm still needs exploration and online learning; if the condition in (ii) does not hold, then the algorithm still needs offline regression.

The above observations explain why the corner case is special, and indicate that an all-case optimal algorithm has to utilize the offline data to learn whether the corner case happens through online decision making. Motivated by these observations, we design the Tweaked O3FU (T-O3FU) algorithm, whose pseudo-code is given in Algorithm 2.

Compared with the O3FU algorithm, the first major difference in the T-O3FU algorithm lies in the preliminary step that tests whether the distance between the average historical price $\bar{p}_{1:n}$ and the interval $\{\psi(\theta):\theta\in\mathcal{C}_0\}$ is smaller than a constant times the length of the interval $\{\psi(\theta):\theta\in\mathcal{C}_0\}$. The goal of this step is to test whether $\delta^2\lesssim\frac{1}{n\sigma^2}$ holds. Specifically, since p^* lies in this confidence interval $\{\psi(\theta):\theta\in\mathcal{C}_0\}$ with high probability, if $\bar{p}_{1:n}$ is "close enough" to this confidence interval, then $\delta^2\lesssim\frac{1}{n\sigma^2}$ with high probability. The second major difference lies in the "re-starting" step after the threshold period $(\lfloor n\sigma^2\rfloor)^2$. The goal of this step is to decide whether $n\sigma^2\lesssim\sqrt{T}$, without knowing T in advance.

Algorithm 2 T-O3FU Algorithm

```
**Input:** offline data $\{(\hat{p}_1, \hat{D}_1), \ldots, (\hat{p}_n, \hat{D}_n)\}$, support of demand parameters $\Theta^\dagger$, support of feasible price $[l, u]$, regularization parameter $\lambda = 1 + u^2$, tuning parameter $K$.
**Initialization:** $V_{0,n} = \lambda I + \sum_{i=1}^n [1\ \hat{p}_i]^\intercal [1\ \hat{p}_i]$, $Y_{0,n} = \sum_{i=1}^n \hat{D}_i [1\ \hat{p}_i]^\intercal$, $\hat{\theta}_0 = V_{0,n}^{-1} Y_{0,n}$, $\mathcal{C}_0 = \{\theta \in \mathbb{R}^2 : ||\theta - \hat{\theta}_0||_{V_{0,n}} \le w_0\}$.
**if** $\frac{\min_{\theta \in \mathcal{C}_0} |\psi(\theta) - \bar{p}_{1:n}|}{|\{\psi(\theta) : \theta \in \mathcal{C}_0\}|} \le K$ **then**
  **for** $t = 1$ **to** $(\lfloor n\sigma^2 \rfloor)^2$ **do**
    Charge the price $p_t = \bar{p}_{1:n}$;
    Observe demand realization $D_t$.
    **if** the algorithm hasn't stopped **then**
      Treat all the data available up to period $(\lfloor n\sigma^2 \rfloor)^2$ as the offline data and run the O3FU Algorithm.
    **end if**
  **end for**
**else**
  Run the O3FU Algorithm.
**end if**
```

The following theorem provides a regret upper bound for the T-O3FU algorithm, see a proof in the full version of this paper.

Theorem 3. *Let π be the T-O3FU algorithm. Then for any $T \ge 1$, $n \ge 1$, $\sigma \ge 0$, $\bar{p}_{1:n} \in [l, u]$, and $\theta \in \Theta^\dagger$, under the regular case, the regret $R_\theta^\pi(T) = O((\sqrt{T} \log T) \wedge \frac{T(\log T)^2}{(n \wedge T)(\bar{p}_{1:n} - \psi(\theta))^2 + n\sigma^2})$, and under the corner case, $R_\theta^\pi(T) = O(T(\bar{p}_{1:n} - \psi(\theta))^2)$.*

The T-O3FU algorithm is optimal for both the regular and the corner cases: the upper bound in Theorem 3 matches the lower bound in Theorem 1 up to a logarithmic factor. We are thus able to completely characterize the optimal regret in each case.

Corollary 1. *The optimal regret for the* OPOD *problem is*

$$R^*(T, n, \sigma, \delta) = \begin{cases} \widetilde{\Theta}\left(\sqrt{T} \wedge \frac{T}{(n \wedge T)\delta^2 + n\sigma^2}\right), & \textit{for the regular case;} \\ \widetilde{\Theta}\left(T\delta^2\right), & \textit{for the corner case.} \end{cases}$$

6. Phase Transitions and Inverse-Square Law

The characterization of the optimal regret for the OPOD problem in Corollary 1 leads to important implications.

First, the decaying patterns of the optimal regret are quite different when the offline sample size n belongs to different ranges. Consider a simple case when $\delta = \Theta(1)$ and $\sigma = o(1)$, which happens when the average historical price is *well-separated* from the optimal price, and the offline data are not so dispersive. In this case, the optimal regret remains at the level of $\widetilde{\Theta}(\sqrt{T})$ when the offline sample size n is within $O(\sqrt{T})$, then decays according to $\widetilde{\Theta}(\frac{T}{n})$ when n is between $\Theta(\sqrt{T})$ and $\Theta(T)$. After that, the optimal regret remains at $\widetilde{\Theta}(\log T)$ when n is between $\Theta(T)$ and $\Theta(\frac{T}{\sigma^2})$, and finally decays according to $\widetilde{\Theta}(\frac{T}{n\sigma^2})$. In this example, there are four ranges of the offline sample size, referred to as the first, second, third and fourth *phase* respectively, and the optimal regret exhibits different decaying patterns in different phases. We refer to the significant transitions between the regret-decaying patterns of different phases as *phase transitions*. For the above example, the result also demonstrates that changing the order of the offline sample size within the first or third phase does not help to reduce the optimal regret, while increasing the order of the offline sample size in the second and fourth phases leads to a considerably fast decay rate of the optimal regret. The above phase transitions hold more generally when $\delta = \Omega(T^{-\frac{1}{4}})$ and $\sigma = O(\delta)$, as depicted in Figure 1. Other values of δ and σ lead to different phase transitions, see the full version of this paper for detailed discussion.

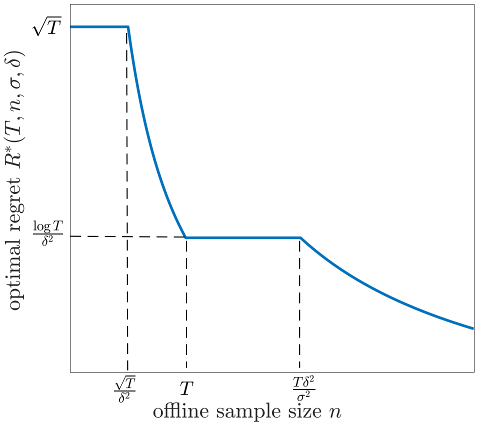

Figure 1. Phase transitions when $\delta = \Omega(T^{-\frac{1}{4}})$ and $\sigma = O(\delta)$

Second, Corollary 1 also characterizes the impacts of the location and dispersion of the offline data on the optimal regret. For the (more interesting) regular case, such impacts can be stated in the following *inverse-square law*: the optimal regret is inversely proportional to the generalized distance δ and standard deviation σ of the offline data. Therefore, the factors δ and σ are intrinsic in the regret bound. The dependency of the optimal regret on σ is consistent with our intuition. Indeed, as the historical prices become more dispersive, i.e., σ increases, the seller gains more information about the unknown demand parameters before she starts online learning, and hence will incur smaller regret loss. The dependency of the optimal regret on δ is more intriguing, which actually suggests that the closer the historical prices are to the optimal price, the worse the optimal regret will be. Seemingly counterintuitive, this is a consequence of the exploration-exploitation tradeoff. Specifically, whenever an algorithm tries to learn the true demand model, it has to make substantial efforts on exploration by charging different prices from the existing average historical price. Therefore, when δ is very small, such a deviation will also lead to a significant gap with the optimal price, leading to greater revenue loss for the seller.

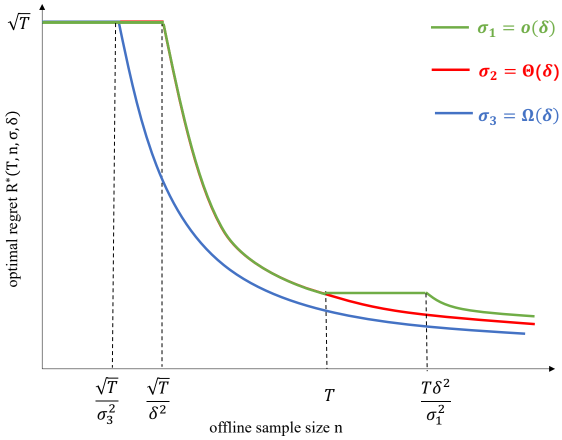

Figure 2. Optimal regret when $\delta = \Omega(T^{-\frac{1}{4}})$ and σ changes

It is also interesting to investigate the joint effect of δ and σ on the optimal regret. When $\delta \gtrsim T^{-\frac{1}{4}}$, there is an important trichotomy on the behavior of the optimal regret, depending on whether σ is less than, equal to or greater than δ. This is illustrated in Figure 2, where the green, red and blue curves depict the above three cases respectively. If $\sigma = o(\delta)$ as shown in the green curve, as discussed before, the optimal regret exhibits four decaying patterns as n shifts between different ranges. If σ and δ are of the same order, as shown in the red curve, the optimal regret exhibits two decaying patterns with changing n: when n is within $O(\frac{\sqrt{T}}{\delta^2})$, the optimal regret remains at the level of $\widetilde{\Theta}(\sqrt{T})$, and when n exceeds $\Omega(\frac{\sqrt{T}}{\delta^2})$, the optimal regret decays according to $\widetilde{\Theta}(\frac{T}{n\delta^2})$. Finally, if $\delta = o(\sigma)$ as shown in the blue curve, the optimal regret exhibits two decaying patterns: it remains at the level of $\widetilde{\Theta}(\sqrt{T})$ when n is within $O(\frac{\sqrt{T}}{\sigma^2})$, and decays according to $\widetilde{\Theta}(\frac{T}{n\sigma^2})$ when n exceeds $\Omega(\frac{\sqrt{T}}{\sigma^2})$. Therefore, as σ gradually increases, depending on its magnitude relative to δ, the number of phases of the optimal regret changes from four phases to two phases, and the entire patterns of the phase transitions of the optimal regret also change accordingly.

When the generalized distance $\delta = O(T^{-\frac{1}{4}})$, Corollary 1 indicates that there are three phases of the optimal regret as n changes. When $n = O(\frac{\sqrt{T}}{\delta^2})$, the optimal regret stays at $\widetilde{\Theta}(\sqrt{T})$. When n is between $\Theta(\frac{\sqrt{T}}{\delta^2})$ and $\Theta(\frac{1}{\delta^2\sigma^2})$, the optimal regret experiences a *sudden* drop from $\widetilde{\Theta}(\sqrt{T})$ to $\widetilde{\Theta}(T\delta^2)$. When n increases to $\Theta(\frac{1}{\delta^2\sigma^2})$, the optimal regret decays according to $\widetilde{\Theta}(\frac{T}{n\sigma^2})$. In particular, the second phase corresponds to the *corner case* defined in §4. In this case, a smaller δ leads to a lower optimal regret, which is in contrast to the inverse-square law for the regular case. This is because when the offline data are extremely dispersive, and the average historical price happens to be very close to the true optimal price, the policy that always charges $\bar{p}_{1:n}$ incurs small regret. In this case, there is no need for exploration and the exploration-exploitation tradeoff does not exist. By contrast, the inverse-square law in the regular case is a consequence of the exploration-exploitation tradeoff.

7. Numerical Experiments

We conduct a numerical study on a synthetic data set to test the performance of our algorithm. Specifically, we compare our O3FU algorithm with the *modified* constrained iterated least squares (CILS) algorithm in Keskin & Zeevi (2014), for both settings without and with offline data. We also investigate the effects of the offline data on our algorithm's empirical regret. Our results demonstrate that O3FU performs comparably with CILS when there is no offline data, and significantly outperforms CILS when there exist offline data. Moreover, O3FU is more stable than CILS (i.e., the variance of the empirical regret of O3FU is much smaller than CILS). We thus believe O3FU should be preferable in many real-life scenarios. The numerical results also provide empirical evidence for phase transitions and the inverse-square law. All the details of the numerical experiments can be found in the full version of this paper at https://arxiv.org/abs/1910.08693.

Acknowledgements

This work was supported in part by MIT Data Science Lab. The authors thank the anonymous referees for many useful suggestions and valuable feedback.

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.

Ban, G.-Y. and Keskin, N. B. Personalized dynamic pricing with machine learning: High dimensional features and heterogeneous elasticity. *Available at SSRN 2972985*, 2019.

Bastani, H., Simchi-Levi, D., and Zhu, R. Meta dynamic pricing: Learning across experiments. *arXiv preprint arXiv:1902.10918*, 2019.

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

Bouneffouf, D., Parthasarathy, S., Samulowitz, H., and Wistub, M. Optimal exploitation of clustering and history information in multi-armed bandit. *arXiv preprint arXiv:1906.03979*, 2019.

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

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

den Boer, A. and Keskin, N. B. Dynamic pricing with demand learning and reference effects. *Available at SSRN 3092745*, 2017.

den Boer, A. V. Dynamic pricing with multiple products and partially specified demand distribution. *Mathematics of operations research*, 39(3):863–888, 2014.

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, 2013.

Gur, Y. and Momeni, A. Adaptive sequential experiments with unknown information flows. *arXiv preprint arXiv:1907.00107*, 2019.

Hsu, C.-W., Kveton, B., Meshi, O., Martin, M., and Szepesvari, C. Empirical bayes regret minimization. *arXiv preprint arXiv:1904.02664*, 2019.

Keskin, N. 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, 2016.

Nambiar, M., Simchi-Levi, D., and Wang, H. Dynamic learning and pricing with model misspecification. *Management Science*, 2019.

Qiang, S. and Bayati, M. Dynamic pricing with demand covariates. *arXiv preprint arXiv:1604.07463*, 2016.

Shivaswamy, P. and Joachims, T. Multi-armed bandit problems with history. In *Artificial Intelligence and Statistics*, pp. 1046–1054, 2012.

Ye, L., Lin, Y., Xie, H., and Lui, J. Combining offline causal inference and online bandit learning for data driven decisions. *arXiv preprint arXiv:2001.05699*, 2020.