Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Robust Dynamic Pricing with Demand Learning in the Presence of Outlier Customers

Xi Chen, Yining Wang

To cite this article:
Xi Chen, Yining Wang (2022) Robust Dynamic Pricing with Demand Learning in the Presence of Outlier Customers. Operations Research

Published online in Articles in Advance 14 Apr 2022

. https://doi.org/10.1287/opre.2022.2280 http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–25
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Robust Dynamic Pricing with Demand Learning in the Presence of Outlier Customers

Xi Chen,[a,*] Yining Wang[b]

[a] Leonard N. Stern School of Business, New York University, New York, New York 10012; [b] Naveen Jindal School of Management, University of Texas at Dallas, Richardson, Texas 75080
*Corresponding author
Contact: xc13@stern.nyu.edu, https://orcid.org/0000-0002-9049-9452 (XC); yxw220006@utdallas.edu, https://orcid.org/0000-0001-9410-0392 (YW)

Received: May 8, 2020
Revised: March 11, 2021; July 31, 2021; October 28, 2021
Accepted: January 31, 2022
Published Online in Articles in Advance: April 14, 2022

Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2022.2280

Abstract. This paper studies a dynamic pricing problem under model misspecification. To characterize model misspecification, we adopt the ε-contamination model—the most fundamental model in robust statistics and machine learning. In particular, for a selling horizon of length T, the online ε-contamination model assumes that demands are realized according to a typical unknown demand function only for $(1-\varepsilon)T$ periods. For the rest of εT periods, an outlier purchase can happen with arbitrary demand functions. The challenges brought by the presence of outlier customers are mainly due to the fact that arrivals of outliers and their exhibited demand behaviors are completely arbitrary, therefore calling for robust estimation and exploration strategies that can handle any outlier arrival and demand patterns. We first consider unconstrained dynamic pricing without any inventory constraint. In this case, we adopt the Follow-the-Regularized-Leader algorithm to hedge against outlier purchase behavior. Then, we introduce inventory constraints. When the inventory is insufficient, we study a robust bisection-search algorithm to identify the clearance price—that is, the price at which the initial inventory is expected to clear at the end of T periods. Finally, we study the general dynamic pricing case, where a retailer has no clue whether the inventory is sufficient or not. In this case, we design a meta-algorithm that combines the previous two policies. All algorithms are fully adaptive, without requiring prior knowledge of the outlier proportion parameter ε. Simulation study shows that our policy outperforms existing policies in the literature.

Funding: X. Chen is supported by the National Science Foundation (NSF) [Grant IIS-1845444].
Supplemental Material: The supplementary material is available at https://doi.org/10.1287/opre.2022.2280.

Keywords: dynamic pricing • regret analysis • robustness • ε-contamination model

1. Introduction

Many operations problems, such as dynamic pricing, assortment optimization, and supply chain management, are built on an underlying probabilistic model. For example, in dynamic pricing, literature often assumes that the realized demand at each time period follows a nonincreasing function of the offered price (which is known as demand function or demand curve) plus a stochastic noise. However, for a multiperiod decision problem with T periods, it is never the case that every time period follows exactly the same underlying probabilistic model. Indeed, nonstandard or outlier purchasing behaviors happen from time to time in reality. In other words, probabilistic models are inherently misspecified, especially in a multiperiod decision problem with a large time horizon T. Therefore, some natural questions arise. First, what should be an appropriate robust model to capture outlier purchasing behavior? Second, how should one design robust online policies to hedge against this outlier behavior?

This paper addresses model misspecification for dynamic pricing, which has been a central problem in revenue management. Let us consider a typical dynamic pricing problem, with T selling periods and an initial inventory $x(T)$. At each selling period t, the retailer offers a price p_t and observes the realized demand d_t based on the customer's purchase:

$$d_t = \min\{f_t(p_t) + \xi_t, x(t)\}, \tag{1}$$

where f_t is an *unknown* demand function at selling period t, which needs to be learned over time; ξ_t is the stochastic noise; and $x(t)$ is the remaining inventory level at the beginning of time t. This paper considers *nonparametric* demand functions f_t, without assuming f_t belongs to any particular parametric families, such as linear or generalized linear demand models. After realizing the demand, the retailer collects the revenue $r_t := p_t d_t$ and updates the inventory level $x(t+1) = x(t) - d_t$.

To model outlier purchasing behaviors, we adopt the ε-contamination model for the online setting. The ε-contamination model, which dates back to the 1960s (Huber 1964), is perhaps the most widely used model in robust statistics. In a standard setup of the ε-contamination model, we are given *n independent and identically distributed* (*i.i.d.*) samples drawn from a distribution $(1-\varepsilon)P + \varepsilon Q$, where P denotes the distribution of interest, and Q is an arbitrary outlier distribution. The parameter $\varepsilon > 0$, which is usually very small, reflects the level at which contamination occurs. In a static robust estimation problem, the goal is to learn the distribution P of interest, in the presence of outlier observations from Q. Recent literature in machine learning has proposed models to extend the ε-contamination model to multiarmed bandit (MAB) settings (Lykouris et al. 2018, Gupta et al. 2019, Zimmert and Seldin 2021), which incorporate adversarial corruptions into MAB. In this paper, we adopt this model for the dynamic pricing problem. In particular, for T selling periods, we assume that there are at most εT periods of outlier demand functions. Moreover, we allow the following two features in the dynamic ε-contamination model:

1. Instead of assuming a fixed outlier demand function g, we model potentially different outlier demand functions g_t for different time periods t;
2. We assume that the εT outlier time periods and corresponding demands can be *arbitrary* and even *adaptive* to historical information (e.g., pricing decisions and realized demands in prior time periods). The outlier time periods and associated demand functions are unknown to the retailer.

This adversarial outlier setting is more practically favorable than "random arrival." For example, in a holiday season, consecutive periods might contain excessively large demand realizations, which cannot be captured by the "random arrival of outliers" in the original ε-contamination model. It is also worth noting that the "outlier proportion" ε is unknown to the retailer, and, thus, the designed online policy needs to be *adaptive* to the unknown ε. The adversarial outlier setting has been recently explored in the machine learning literature for a wide range of problems, including multiarmed bandit (Lykouris et al. 2018, Gupta et al. 2019, Bogunovic et al. 2020, Agarwal et al. 2021, Zimmert and Seldin 2021), reinforcement learning (Lykouris et al. 2019), and contextual pricing (Krishnamurthy et al. 2021). The main technical challenge in our problem arises from learning a *nonparametric* demand function and incorporating inventory constraints simultaneously. Please refer to the related work in Section 2 for more detailed discussions.

To better understand this problem, we first consider a simple setting without any inventory constraint. Now, the main challenge lies in how to learn the unknown demand model under this adversarial corruption model. It is worth noting that existing upper-confidence-bound-type algorithms cannot be directly applied because lengths of the confidence intervals depend on the outlier proportion ε, which is not known a priori (see more discussions in Section 4). Thus, we adopt the Follow-the-Regularized Leader (FTRL) algorithm with the regularizer in Equation (6) (see Algorithm 1).

The regularizer, which is a variant of the α-Tsallis-Inf regularizer (Audibert and Bubeck 2009, Zimmert and Seldin 2021), prevents the sampling probability weight parameters from being too close to either zero or one. This construction of regularizer is essential in establishing a "self-bounding" property in Lemma 1. We also note that such a regularizer was also used by Zimmert and Seldin (2021) for the robustness purpose in multiarmed bandit. The FTRL algorithm is essentially an online mirror descent strategy, which automatically balances exploration and exploitation in the presence of outlier demands. In our analysis, by constructing "shadow" regret terms, we provide a decomposition analysis (see the proof of Lemma 2), which derives regret bounds that are nearly optimal in both the number of arms and the suboptimality gaps of *each* individual arm. This technical result is the key to establishing the optimal regret upper bound in continuous pricing with outliers.

Here, we highlight one technical subtlety: In contrast to the multiarmed bandit, the action space (i.e., price range) is continuous in dynamic pricing. To handle this case, we choose to discretize the action space. However, the number of discretized prices (corresponding to the number of arms) needs to scale polynomially with the time horizon T because the discretization has to be sufficiently dense. Thus, existing Bandit with Knapsacks techniques (see, e.g., Badanidiyuru et al. 2018) will yield suboptimal regret bounds (see more discussions in Section 2). We delve deeper into the structure of the *suboptimality gaps* among the candidate prices. More specifically, the key observation is that prices that are farther away from the revenue-maximization price have significantly smaller expected revenue (and hence a larger suboptimality gap) due to *concavity* of the revenue function. This key structure in dynamic pricing plays a critical role in our analysis.

In the second part of the paper, we introduce the inventory constraint. When the inventory is sufficient, it has little impact on the pricing policy, and one can directly use the FTRL policy. Thus, we focus on the pricing problem when the initial inventory level is insufficient. In particular, denote by p^o the *revenue-*

maximization price—that is, the price that maximizes the revenue $r(p) := pf_0(p)$ without any inventory constraint; denote by p^c the *clearance price*—that is, the price that (in expectation) depletes the inventory at the end of T periods. Gallego and Van Ryzin (1994) showed that the optimal price p^* with respect to the fluid approximation is the maximum of p^o and p^c. Thus, we consider the insufficient-inventory capacity case when $p^c > p^o$. We note that it could be hard to determine whether inventory is sufficient or not in practice due to the unknown demand function. Thus, the algorithm developed in this setting can be used as a subroutine of the general inventory setting described later. Intuitively, when inventory is scarce, the retailer should charge a higher price than the revenue-maximization price. When $p^c > p^o$, it is natural to propose a search-based strategy to find p^c. We first consider the simple case where the outlier proportion ε is known. By leveraging the fact that demand rate function f_0 is strictly monotonically decreasing, we propose a robust bisection-search algorithm in Algorithm 2 to identify the clearance price p^c. To make the proposed algorithm robust to outlier customers, our procedure is quite different from the bisection-search algorithms in the existing dynamic pricing literature (Lei et al. 2014, Wang et al. 2014). In particular, our confidence bounds in the bisection search are carefully designed to incorporate the outlier proportion ε, which adds an additional layer of sophistication in both the algorithm and its regret analysis. When ε is unknown, we apply the robust bisection-search algorithm from Lykouris et al. (2018) and Krishnamurthy et al. (2021) to multiple "threads" of candidate values for the parameter ε, which we term as "ε-candidates," searching for the right ε in parallel. More specifically, the multithread bisection search in Algorithm 3 runs in parallel on a grid of geometrically discretized candidates of ε and, thus, is adaptive to the unknown ε.

Finally, we consider general dynamic pricing with an arbitrary inventory level. Because the underlying demand rate f_0 is unknown, the relationship between p^c and p^o is unknown as well, and, thus, one has to decide whether the inventory level is sufficient or not. To address this challenge, we develop a meta-algorithm that learns the relationship between p^c and p^o and invokes the appropriate algorithm for different cases based on the doubling trick and randomized exploration strategy (see Algorithm 5). The doubling trick refers to a common technical strategy used in bandit learning algorithms that divide the time horizon into epochs with geometrically increasing lengths to facilitate the learning or estimation of an unknown quantity with small incurred regret. It is worth noting that typical exploration phases that perform randomized price experiments during the first T_0 selling periods will *not* work in the presence of robust/outlier customers because it is possible that purchases made throughout the entire exploration phase are outliers, leading to completely erroneous learned information. To address this challenge, we first use the doubling trick to divide the entire T selling periods into epochs with geometrically increasing lengths and then randomly inject exploration periods into each epoch. Such a combination of the doubling trick and a randomized exploration is the key to hedge against outlier purchases. Table 1 summarizes the algorithm choices we made in this paper under different settings of inventory levels.

For each algorithm, we establish the upper bound of its incurred cumulative regret, which measures the gap between the optimal revenue and expected revenue collected from our dynamic pricing policy. All of our established regret upper bounds take the form of $\widetilde{O}(\varepsilon T + \sqrt{T})$, where in the notation $\widetilde{O}$, we drop logarithmic factors in T. In Theorem 5, we also prove that this regret bound is rate-optimal up to logarithmic factors in T. The term εT in the regret bound is the price paid for being robust. When $\varepsilon = 0$, the regret bound automatically reduces to the optimal regret bound of $\sqrt{T}$ for the classical dynamic pricing when there is no outlier demand. We also remark that the regret bound $\widetilde{O}(\varepsilon T + \sqrt{T})$ can be implied by a potential "fully adversarial" dynamic pricing algorithm, with no assumptions imposed on the number of outlier selling periods and $\widetilde{O}(\sqrt{T})$ regret compared against a stationary benchmark in hindsight. Such an algorithm, however, does not exist as far as we know and is likely to be very difficult to design. The closest algorithms are perhaps the ones developed for the *bandit convex optimization* question (Flaxman et al. 2004, Hazan and Levy 2014, Besbes et al. 2015, Bubeck et al. 2017). Although these algorithms indeed work under the fully adversarial setting, there are other significant differences in terms of convexity assumptions and inventory constraints that prevent them from being applicable to our problem. We discuss in further detail significant differences from this line of previous works in the next section.

The rest of the paper is organized as follows. Section 2 discusses the related work and highlights the difference between this paper and existing works. Section 3 provides the model primitives and necessary assumptions. Sections 4 and 5 propose FTRL and multithread robust bisection-search algorithms for the unconstrained-inventory setting and the insufficient-inventory setting, respectively. Section 6 develops the meta-algorithm with a partial exploration scheme, which effectively identifies whether the inventory is insufficient or not. In Section 7, we provide illustrative numerical studies, followed by a conclusion in Section 8. The technical proofs will be relegated to the supplementary material.

Table 1. Summary of the Developed Algorithms in this Paper

Inventory scenarios	Developed algorithms
Unconstrained/sufficient inventory	A Follow-the-Regularized Leader (Algorithm 1)
Insufficient inventory	Multithread robust bisection search (Algorithms 2 and 3)
General inventory	Meta-algorithm (Algorithm 5) with a partial exploration scheme (Algorithm 4)

2. Related Work

In this section, we briefly review literature from three perspectives: dynamic pricing, robust machine learning, and bandit convex optimization. We highlight the main technical challenges of our problem as compared with existing literature.

2.1. Dynamic Pricing Literature

Because of the increasing popularity of online retailing, dynamic pricing has become an active research area in revenue management in the past decade. We only briefly review a few related works on single-product pricing problem and refer the interested readers to Bitran and Caldentey (2003), Elmaghraby and Keskinocak (2003), and den Boer (2015) for a comprehensive literature survey. The seminal work by Gallego and Van Ryzin (1994) lays out the foundation of the problem and shows that the optimal price with respect to the fluid approximation is the larger price between the revenue-maximization price and inventory-clearance price. Earlier work in dynamic pricing (see the surveys Bitran and Caldentey 2003 and Elmaghraby and Keskinocak 2003) assumes that demand information is known to the retailer a priori and either characterizes or computes the optimal pricing decisions.

In many retailing industries, such as fast fashion, the underlying demand function is unknown and cannot be easily estimated from historical data. As a result, a lot of recent research in this area focuses on the joint learning and decision-making problem, which simultaneously learns the underlying demand function and makes the price decision (see, e.g., Araman and Caldentey 2009, Besbes and Zeevi 2009, Farias and Van Roy 2010, Broder and Rusmevichientong 2012, Harrison et al. 2012, den Boer and Zwart 2013, Keskin and Zeevi 2014, Lei et al. 2014, Wang et al. 2014, Chen et al. 2015, Miao et al. 2019, Chen et al. 2021a, Wang et al. 2021, Chen et al. 2022, and references therein). Along this line of research, Besbes and Zeevi (2009) first proposed separate exploration and exploitation strategies, which lead to suboptimal regret of $\widetilde{O}(T^{3/4})$ for nonparametric demands and $\widetilde{O}(T^{2/3})$ for parametric demands. Wang et al. (2014) improved this result by developing joint exploration and exploitation strategies that achieve the optimal regret of $\widetilde{O}(T^{1/2})$ up to a logarithmic factor in T. Lei et al. (2014) further improved the result by removing the logarithmic factor in T. Moreover, den Boer and Zwart (2013) proposed a controlled variance pricing policy, and Keskin and Zeevi (2014) discussed a semi-myopic pricing policy for a class of parametric demand functions without inventory constraints. With inventory constraint, Chen et al. (2015) proposed a linear price-correction policy that performs computationally efficient price reoptimization for nonparametric demand functions. Broder and Rusmevichientong (2012) also proposed a $O(\log T)$-regret policy when demand functions satisfy a "well-separated" condition. In addition, there are a number of works proposing Bayesian strategies for dynamic pricing (Farias and Van Roy 2010, Harrison et al. 2012).

There are many important extensions of single-product dynamic pricing, such as network revenue management with multiple products (see, e.g., Gallego and Van Ryzin 1997, Ferreira et al. 2018, Chen and Shi 2019, and references therein), pricing in a dynamically changing environment (Besbes et al. 2015, Keskin and Zeevi 2016), dynamic pricing with a limited number of price changes (Cheung et al. 2017), and dynamic pricing with potentially high-dimensional covariates (Ban and Keskin 2017, Lobel et al. 2018, Javanmard and Nazerzadeh 2019, Nambiar et al. 2019, Chen et al. 2021b, Chen and Gallego 2021). It is an interesting future direction to investigate robust pricing policies for these more complex dynamic pricing problems.

We also position our paper in the Bandit with Knapsacks literature (Badanidiyuru et al. 2018). A key difference between Badanidiyuru et al. (2018) and Agrawal and Devanur (2014) and our problem setting is that in Badanidiyuru et al. (2018) and Agrawal and Devanur (2014), the action space is finite, and the regret upper bound depends polynomially on the number of arms N_1 (more specifically, $\widetilde{O}(\sqrt{N_1 T})$ when inventory levels scale linearly with T (Badanidiyuru et al. 2018)). In contrast, the action space in our problem (i.e., the price range) is continuous. Although it is possible to discretize the price space, the number of discretized prices needs to scale polynomially with T (e.g., $N_1 \asymp T^{1/4}$), which would lead to suboptimal regret bounds, such as $\widetilde{O}(T^{5/8})$. Moreover, Agrawal and Devanur (2015) studied continuous action spaces, but the demand model is parametric (linear to be more specific) and cannot be easily applied to nonparametric demand-learning problems.

2.2. Literature on Model Misspecification and Robust Statistics

In learning and decision-making settings, a few recent works investigate the impact of model misspecification in revenue management—for example, Cooper et al. (2006) for capacity booking problems, Lei et al. (2014) and Besbes and Zeevi (2015) for dynamic pricing, and Chen et al. (2019) for assortment optimization. In particular, Besbes and Zeevi (2015) shows that a class of pricing policies based on linear demand functions perform well, even when the underlying demand is not linear. Lei et al. (2014) proposed nonparametric dynamic pricing policies that achieve the optimal regret for parametric models. Cooper et al. (2006) also identified some cases where simple decisions are optimal under misspecification. However, our setting is quite different and has not been considered in existing dynamic pricing literature. We study the model misspecification from a model-corruption perspective, which allows arbitrary outlier purchasing behavior in adversarially chosen time periods. We also note that this outlier behavior has been recently studied in a different problem—assortment optimization under multinomial logit models (Chen et al. 2019). However, the assortment optimization is structurally different from the dynamic pricing problem considered in this paper. First, the choice function in assortment optimization is a parametric problem (parameterized by utility parameters), whereas our dynamic pricing problem has nonparametric demand functions. Moreover, the dynamic pricing problem needs to learn the relationship between the revenue-maximizing price and the inventory-clearance price and combine these two cases via a meta-algorithm.

In statistics and machine learning literature, the ε-contamination model, which was proposed by P. J. Huber (Huber 1964), is perhaps the most widely used robust model and has recently attracted much attention from the machine learning community (see, e.g., Chen et al. 2016, Diakonikolas et al. 2017 and 2018, and references therein). Despite this attention, online learning in the ε-contamination model or its generalizations is relatively unexplored. In the online setting, Esfandiari et al. (2018) studied online allocation under a mixing adversarial and stochastic model, but the setting does not require any learning component. For online learning, the works of Lykouris et al. (2018), Gupta et al. (2019), and Zimmert and Seldin (2021) studied the contaminated stochastic multiarmed bandit. Our FTRL algorithm is adopted from Zimmert and Seldin (2021), and the "multilayer active arm race" for MAB (Lykouris et al. 2018) motivates our multithread bisection-search algorithm. We note that existing results for robust multiarmed bandit may *not* be optimal in terms of their regret dependency on the total number of arms. Therefore, directly using either result from Lykouris et al. (2018) or Zimmert and Seldin (2021) leads to regret significantly worse than the optimal rate $\widetilde{O}(\varepsilon T+\sqrt{T})$ in the presence of outliers. Please refer to the discussion below Lemma 2 and Remark 1 for more details.

More recently, the adversarial outlier model has been studied in many settings—for example, Gaussian process bandit optimization (Bogunovic et al. 2020), reinforcement learning (Lykouris et al. 2019), dueling bandits (Agarwal et al. 2021), assortment optimization (Chen et al. 2019), contextual pricing (Krishnamurthy et al. 2021), and product rankings (Golrezaei et al. 2020). The work of Krishnamurthy et al. (2021) is closer to our modeling in the sense that Krishnamurthy et al. (2021) assumes that outlier customers could be completely irrational, with the number of such customers being bounded by a corruption-level parameter. Nevertheless, Krishnamurthy et al. (2021) studied a parametric model, where truthful agents (or agents with bounded rationality) realize their valuations according to a linear model, which is significantly different from the nonparametric modeling carried out in this paper. In addition, comparing the recent work on dynamic assortment optimization under the ε-contamination model (Chen et al. 2019), we remark that assortment optimization has a large, yet finite, action set, whereas in dynamic pricing, the size of the action set (i.e., the number of prices) is infinite. Thus, it requires different analyses into the structure of the underlying problem. Moreover, our work considers the inventory constraint, which has been fully explored in previous works on adversarial outliers. We have also remarked on the important technical differences between our results and the existing results from Lykouris et al. (2018), Gupta et al. (2019), and Zimmert and Seldin (2021) in Sections 4 and 5.

There has been existing work on studying dynamic pricing in the presence of *strategic* customers. In most such work, there is an underlying behavior model that characterizes how a strategic customer realizes his valuations (Golrezaei et al. 2019). In contrast, the ε-contamination modeling allows for *arbitrary* behaviors of outlier customers, who do not necessarily follow any predetermined behavior models.

2.3. Literature on Bandit Convex Optimization

Bandit convex optimization (Flaxman et al. 2004, Hazan and Levy 2014, Besbes et al. 2015, Bubeck et al. 2017) is an active field in machine learning that is closely related to the dynamic pricing problem. In bandit convex optimization, for every time period t, there is an unknown convex function h_t. The algorithm provides an approximate minimizer x_t and receives feedback $h_t(x_t)+\xi_t$ with ξ_t being i.i.d. centered noise. The regret of an algorithm is then measured against a stationary benchmark x in hindsight, or, more specifically, $\sum_t h_t(x_t)-\min_x\sum_t h_t(x)$.

The bandit convex optimization question is related to the dynamic pricing question, by setting $h_t(p)=-pf_t(p)$ and trying to find p that minimizes the negative-expected revenue. There are, however, two major differences between the bandit convex optimization setup and our problem:

1. The bandit convex optimization question assumes every h_t is convex in x, whereas the (negative) revenue function $-r(p) = -pf_t(p)$ is *not* necessarily convex in p. Instead, the conventional assumption is that $-r(d) = -df_t^{-1}(d)$ is convex in the *demand*, not the price (see, for example Assumption 3). This also means mainstream bandit convex optimization algorithms, such as estimating gradient descent (Flaxman et al. 2004, Hazan and Levy 2014, Besbes et al. 2015), *cannot* be applied because one cannot run gradient descent on the demand rate d directly (because the algorithm does not know the precise price p resulting in demand rate d). Running gradient descent on the price variable p, on the other hand, does not work either because $-r(p)$ is *not* convex in p.

2. The bandit convex optimization question essentially assumes no initial inventory constraints. Whereas in the stochastic setting (in which f_t does not change over time), the initial inventory constraint can be handled via a pure-exploration phase, in fully adversarial settings, there is no easy way to address the initial inventory constraint because the demand function f_t evolves over time.

3. Model Primitives

In this section, we formally introduce our online ε-contamination model for dynamic pricing and corresponding model assumptions.

Assume that there are a total of T selling periods. The retailer has an initial inventory level of $x(T) = x_T \in (0, T]$ at the beginning. Without loss of generality, we assume that the total inventory x_T is normalized to be less than T, and we use x_T to make the dependence of the total inventory on the time horizon more explicit. At each selling period $t \in [T]$, the retailer decides a price $p_t \in [\underline{p}, \overline{p}]$, where $\underline{p}$ and $\overline{p}$ are the minimum and maximum prices, respectively. Given the price p_t, the retailer observes a realized demand

$$d_t = \min\{f_t(p_t) + \xi_t, x(t)\}, \tag{2}$$

where $f_t : [\underline{p}, \overline{p}] \to [0,1]$ is an *unknown* demand function, which varies with t and follows our online ε-contamination model defined in the following. The term ξ_t is the stochastic noise, which satisfies $\mathbb{E}[\xi_t \mid p_t] = 0$. Because our total inventory $x(T)$ is normalized to be less than T, we assume that the realized demand is also normalized and bounded with $d_t \in [0,1]$ almost surely. At each time t, $x(t)$ represents the remaining inventory level when there are t time periods remaining. With the realized demand d_t, the retailer collects revenue $r_t = p_t d_t$ and updates his inventory level with $x(t-1) = x(t) - d_t$.

In our ε-contamination model, there are εT selling periods during which the incoming customer is an *outlier*, whose demand curve could be drastically different from that of typical customers. In particular, we let $f_0 : [\underline{p}, \overline{p}] \to [0,1]$ be the unknown demand function of a *typical* customer. For each time period t, we use $\iota_t \in \{0,1\}$ to denote whether the customer at time t is an outlier ($\iota_t = 1$ corresponds to an outlier, and $\iota_t = 0$ corresponds to a typical customer). The demand function f_t at time t is then defined as

$$f_t(p_t) = \begin{cases} f_0(p_t), & \text{if } \iota_t = 0, \\ g_t(p_t), & \text{if } \iota_t = 1; \end{cases}$$

where $g_t : [\underline{p}, \overline{p}] \to [0,1]$ is an *arbitrary* measurable function characterizing the demand of the outlier customer at time t. Note that in this formulation, demands of outlier customers may not be the same across different selling periods.

The occurrences and demand curves of outlier customers in this paper are modeled by using the *adaptive adversary* model in the bandit learning literature. More specifically, let $\mathcal{H}_{t-1} = \{p_{t'}, d_{t'}, f_{t'}, \iota_{t'}\}_{t'<t}$ denote the filtration of the history of all selling periods prior to time t. Recall that x_T denotes the total inventory level. We define $x_0 = x_T/T \in (0,1]$, which is known as the inventory rate. A problem instance is modeled as $\mathcal{E} = \{f_0, x_0, \varphi_1, \varpi_1, \varphi_2, \varpi_2, \cdots, \varphi_T, \varpi_T\}$, such that $\iota_t = \varpi_t(\mathcal{H}_{t-1})$, $f_t = f_0$ if $\iota_t = 0$ and $f_t = g_t = \varphi_t(\mathcal{H}_{t-1})$ if $\iota_t = 1$, where ϖ_t and φ_t are functions of $\mathcal{H}_{t-1}$. It is guaranteed that $\sum_t \iota_t \le \varepsilon T$ almost surely, for any policy.

3.1. Dynamic Pricing with Demand-Learning Policies

In this paper, we are interested in designing efficient dynamic pricing policies with demand learning, meaning that the policy we designed does not have knowledge of either f_0 or g_t a priori and must learn the demand-rate function f_0 over time.

Mathematically, an admissible policy π can be parameterized as $\pi = (\pi_1, \pi_2, \cdots, \pi_T)$, such that each price $p_t \sim \pi_t(\mathcal{F}_{t-1})$, where $\mathcal{F}_{t-1} = \{p_{t'}, d_{t'}\}_{t'<t}$. This means that the price decision p_t at time t must only be based on observations from selling periods *prior* to period t. Note that the filtration $\mathcal{F}_{t-1}$ does *not* contain $f_{t'}$ or $\iota_{t'}$ because the demand-rate function or whether the period t' is an outlier *cannot* be known to the retailer, even after period t'.

3.2. Assumptions and the Asymptotic Regret Regime

We make following assumptions throughout this paper:

Assumption 1 (Strictly Monotonic and Smooth Demand Curve). $f_0 : [\underline{p}, \overline{p}] \to [0,1]$ *is strictly monotonically decreasing, with* $f_0(\underline{p}) = 1$ *and* $f_0(\overline{p}) = 0$. *There exists an inverse function* $f_0^{-1} : [0,1] \to [\underline{p}, \overline{p}]$ *such that* $f_0(f_0^{-1}(d)) = d$ *for all* $d \in [0,1]$. *Furthermore, there exists constants*

Published in Operations Research on April 14, 2022 as DOI: 10.1287/opre.2022.2280.
This article has not been copyedited or formatted. The final version may differ from this version.

$0 < \underline{L}_p \le \overline{L}_p < \infty$ and $0 < \underline{L}_d \le \overline{L}_d < \infty$, with $\underline{L}_d, \underline{L}_p \ge 1$, such that

$$\underline{L}_p|p - p'| \le |f_0(p) - f_0(p')| \le \overline{L}_p|p - p'|, \qquad \forall p, p' \in [\underline{p}, \overline{p}];$$
$$\underline{L}_d|d - d'| \le |f_0^{-1}(d) - f_0^{-1}(d')| \le \overline{L}_d|d - d'|, \qquad \forall d, d' \in [0, 1].$$

Assumption 2 (Strongly Concave and Smooth Revenue Function). *Let $r(d) = df_0^{-1}(d)$ be the revenue function. Then, r is twice continuously differentiable on (0, 1). Furthermore, there exist constants $0 < \sigma^2 \le M^2 < \infty$ such that $\sigma^2 \le -r''(d) \le M^2$ for all $d \in (0, 1)$.*

Assumption 3 (Outlier Frequency). *There exists a small constant $\alpha > 0$ such that $\varepsilon \le T^{-\alpha}$, and, thus, the total number of outliers is bounded by $\varepsilon T \le T^{1-\alpha}$.*

Assumptions 1 and 2 are standard Lipschitz continuity and concavity assumptions adopted in the pricing literature. Notice that because the range of f_0 is $[0, 1]$ and the realized demands $\{d_t\}$ also belong to $[0, 1]$ almost surely, the stochastic noise variables $\{\xi_t\}$ are bounded almost surely. Assumption 3 imposes an additional constraint on the outlier proportion parameter ε. Essentially, we assume that the total number of corrupted periods εT is *sublinear* with respect to the time horizon T. This is a practical and easily justifiable assumption because in most applications, there will not be too many periods or customers who behave like outliers. And if there are excessive amount of outlier customers, it is unlikely for a retailer to learn an effective pricing policy. Also, the constants in the above assumptions $(\underline{L}_d, \underline{L}_p, \overline{L}_p, \overline{L}_d, M, \sigma)$ are for theoretical analysis only, and our proposed algorithms do not need to know these constants to operate.

In this paper, we consider the following asymptotic regime. Let $f_0 : [\underline{p}, \overline{p}] \to [0, 1]$ be a *fixed*, but unknown, demand-rate function (for typical customers) satisfying all the above assumptions. When there is no inventory constraint and the demand is known, it is clear that the revenue-maximization price $p^o = \text{argmax}_{p \in [\underline{p}, \overline{p}]} pf_0(p)$ maximizes the total expected revenue.

For the inventory-constrained case with T time periods, the initial inventory level $x_T = x_0 T \in (0, T]$ is known to the retailer before the first selling period. It is a well-known result in the literature (Gallego and Van Ryzin 1994) that the optimal expected revenue of the fluid approximation (without any outlier customer) takes the following form:

$$T \times (\max\{p^o, p^c\}) f_0(\max\{p^o, p^c\}), \tag{3}$$

where $p^c = f_0^{-1}(x_0)$ is the *clearance price* at which the initial inventory $x_T = x_0 T$ is expected to clear at the end of T time periods. Alternatively, we write the first term in Equation (3) as

$$T \times r^* := T \times \max_{p \in [\underline{p}, \overline{p}]} r(p; f_0, x_0), \tag{4}$$

where $r(p; f_0, x_0) := p\min\{f_0(p), x_0\}$.

Let $r^* = \max_{p \in [\underline{p}, \overline{p}]} r(p; f_0, x_0)$. By theorem 2 from Gallego and Van Ryzin (1994), Tr^* is an upper bound of the expected revenue for any dynamic pricing policy. In other words, Tr^* is an upper bound of the total optimal expected revenue (with respect to the Markov Decision Process formulation). Thus, for an admissible policy π, its *regret* over T time periods with at most E outlier selling periods is defined as

$$\mathfrak{R}_{T,E}(\pi; f_0, x_0) := Tr^* - \mathbb{E}^{\pi}\left[\sum_{t=1}^{T} r_t\right], \tag{5}$$

where $r_t = p_t d_t$ is the revenue collected by the retailer at time t, following the pricing policy π.

In our Regret Definition (5), for outlier periods/customers, we still use the optimal revenue for typical customers as benchmarks for comparison. Using outlier customers as benchmarks, on the other hand, shall *not* change our upper regret bounds. In fact, using any benchmark in the regret definition for those outlier periods will lead to at most $O(\varepsilon T)$ regret in the regret upper bound. On the other hand, the term $O(\varepsilon T)$ is unavoidable in the regret (and also appears in our bound) because any policy cannot make a reasonable prediction for those outlier periods, when compared with a benchmark defined on nonoutlier customers. We also remark that in our regret metric $\mathfrak{R}_{T,E}(\pi; f_0, x_0)$, the underlying demand-rate function f_0 and the inventory rate x_0 stay fixed and do not change with time horizon T.

We note that another possible benchmark would be the optimal hindsight solution. In the setting without any inventory constraint, it corresponds to the benchmark $\sum_{t=1}^{T} p^* f_t(p^*)$, where p^* is a stationary benchmark price that maximizes $\sum_t pf_t(p)$ across T selling periods with different demand-rate functions f_t in hindsight. When using this hindsight benchmark, it is unclear whether $\widetilde{O}(\sqrt{T})$ regret is attainable. Because the expected revenue as a function of price is not concave, we cannot immediately apply bandit convex-optimization techniques; the discretization plus multiarmed-bandit approach is not likely to work either, as we have too many arms due to discretization, and the gap-distribution property might not be useful when the optimal fixed hindsight benchmark is used. Note also that there is a complicated relationship between $\max_p \sum_t pf_t(p)$ and Tr^*. In particular, in the case of all outlier customers being demand suppressed (i.e., $f_t \equiv 0$ if time period t belongs to an outlier), then $\max_p \sum_{t=1}^{T} pf_t(p) \le T \times r^*$; in the case of all outlier customers being demand active (i.e., $f_t \equiv 1$ if time period t belongs to an outlier), then $\max_p \sum_{t=1}^{T}$

$pf_t(p) \geq T \times r^*$. Moreover, in all cases, the absolute value difference $|\max_p \sum_{t=1}^{T} pf_t(p) - Tr^*|$ is upper-bounded by ϵT.

In some previous literature on dynamic pricing with demand learning—for example, Besbes and Zeevi (2009) and Wang et al. (2014)—a competitive ratio metric is used to evaluate the performance of a policy. For a policy π, its competitive ratio is defined as $1 - \frac{R^\pi}{R^*}$, where R^π is the expected reward of policy π and R^* is the expected reward of the optimal policy. When R^* scales linearly with T (e.g., $R^* = Tr^*$ in (5)), the competitive ratio and cumulative regret are equivalent, as an $O(1/\sqrt{T})$ competitive ratio would correspond to an $O(\sqrt{T})$ cumulative regret.

4. Dynamic Pricing Without Inventory Constraint

We first consider the unconstrained-inventory setting. As the dynamic pricing resembles multiarmed bandits, the popular upper confidence bound (UCB) policy might be a natural choice. However, with adversarial outliers, we will provide a toy example to illustrate the failure of the UCB policy. Similar constructions of the failed cases also appear in prior works (Lykouris et al. 2018, Gupta et al. 2019, Krishnamurthy et al. 2021). For the purpose of completeness, we provide a simple example with only two candidate prices, $p^l < p^h$, without any inventory constraint. Let $d^l = f_0(p^l), d^h = f_0(p^h)$ and $r^l = p^l d^l, r^h = p^h d^h$ be the expected demand rates and profits at the two price levels, respectively. Suppose $r^l > r^h$, meaning that for the majority of customers (i.e., typical customers), the lower price p^l results in higher revenue. For a UCB policy, because it is a deterministic policy, the adaptive adversary could realize adversary demand rates $\widetilde{d}^l, \widetilde{d}^h$ during the first $\sqrt{T}$ times p^l or p^h is offered, with $\widetilde{r}^l = p^l \widetilde{d}^l < p^h \widetilde{d}^h = \widetilde{r}^h$ and furthermore $\widetilde{r}^l < r^h$. In this case, because $\widetilde{r}^l < \widetilde{r}^h$, with overwhelming probability $1 - O(e^{-\Omega(T^{1/4})})$ the upper confidence bound of p^l is lower than the upper confidence bound of p^h when normal (typical) customers kick in. Furthermore, because $\widetilde{r}^l < r^h$, the upper confidence bound of p^l will remain lower than the upper confidence bound of p^h, even after normal (typical) customers start to appear at the p^h price. Therefore, with overwhelming probability, the UCB algorithm will commit to the wrong price p^h for the rest of the time horizon, leading to an $\Omega(T)$ linear regret, even when the outlier portion ε is as small as $O(1/\sqrt{T})$.

Instead of using the UCB policy, we propose to adopt the Follow-the-Regularized-Leader algorithm with a carefully designed regularizer that is robust to outlier customers. The FTRL algorithm is due to Audibert et al. (2014) and was also studied in the context of MAB with outliers (Zimmert and Seldin 2021). More discussions will be provided in the paragraph after Lemma 2 and Remark 1. As the price range is a continuous interval, we first discretize the interval $[p^o, \bar{p}]$ into N_1 candidate prices. According to our regret bound in Theorem 1, the discretization level N_1 will be set to $N_1 = \lceil T^{1/4} \rceil$. We use an online mirror descent strategy to balance the exploration and exploitation in the presence of an unknown number of outlier customers.

Algorithm 1 (A Follow-the-Regularized-Leader Algorithm for the $p^c < p^o$ Case with Unknown ε)

```
Parameters: $p^o \in [\underline{p}, \bar{p}]$, time horizon $T$, $N_1 \in \mathbb{N}$, regularizer $\psi$ and step sizes $\{\eta_t\}_{t=1}^{T}$ defined in (6);
Initialize: $p(1), \cdots, p(N_1)$ evenly spaced prices partitioning $[p^o, \bar{p}^o]$; $\widehat{L}_0 = 0 \in \mathbb{R}^{N_1 \times 1}$;
for $t = 1, 2, \cdots, T$ do
    Compute $w_t = \arg\max_{w \in \Delta^{N_1 - 1}} \langle w, \widehat{L}_{t-1} \rangle - \frac{1}{\eta_t} \psi(w)$, where $\Delta^{N_1-1}$ is the $N_1$-dimensional probability simplex;
    Sample $i_t \sim w_t$; offer price $p(i_t)$ and observe realized demand $d_t$;
    Update $\widehat{L}_t = \widehat{L}_{t-1} + \widehat{\ell}_t$ where
        $$\widehat{\ell}_{ti} = \begin{cases} \dfrac{p(i)d(t) - \bar{p}}{w_{ti}} + \bar{p} & \text{if} \quad i = i_t; \\ \bar{p} & \text{if} \quad i \neq i_t. \end{cases}$$
end for
```

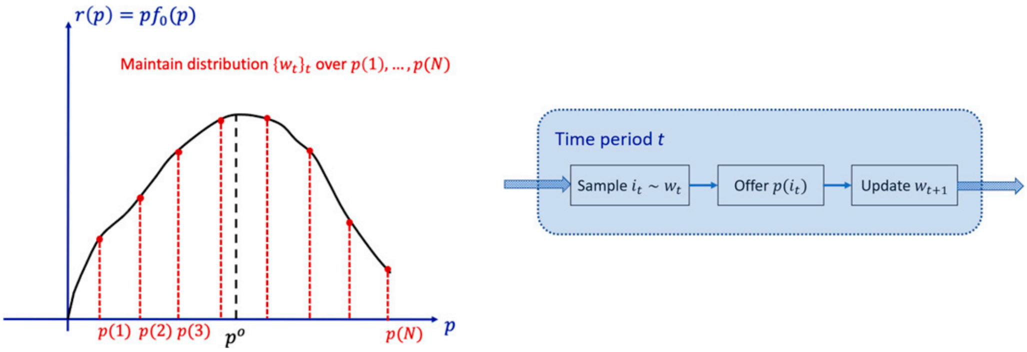
The pseudo-code of the proposed algorithm is given in Algorithm 1. At a high level, Algorithm 1 partitions the entire pricing interval $[0, 1]$ into N_1 discretized prices and maintains a probability distribution (denoted as w_t in Algorithm 1) over all candidate prices and sample one at each time period. The probability distributions are then updated by using the Follow-the-Regularized-Leader principle, taking into consideration both the realized demands from prior periods and a carefully chosen regularization term ψ. The FTRL principle has been used in adversarial bandit learning problems, by learning-probability distributions over actions sequentially based on both the historical data and a carefully designed regularizer (Zimmert and Seldin 2021). This approach does *not* require prior knowledge of the outlier proportion ε, making it suitable for the adaptive case. To better illustrate the key components in Algorithm 1, we provide the overall flow of the algorithm and key algorithmic ideas in Figure 1.

We will run Algorithm 1 with the following selection of the regularizer ψ and the step sizes $\{\eta_t\}$:

$$\psi(w) = \sum_{i=1}^{N_1} -\sqrt{w_i} - \sqrt{1 - w_i}, \quad \eta_t = \eta_0 / \sqrt{t}, \tag{6}$$

where $\eta_0 > 0$ is a small constant to be specified later. This regularizer is the α-Tsallis-Inf regularizer (Audibert and Bubeck 2009, Zimmert and Seldin 2021), which

Figure 1. (Color online) Plots of the Key Algorithmic Idea (Left) and Schematic (Right) of Algorithm 1



has also been used in reinforcement learning for best-of-both-world-type results (Jin and Luo 2020). The regularizer prevents the weight parameters $\{w_i\}$ from being too close to either zero or one, which is essential in establishing a self-bounding property in Lemma 1 presented later. We also note that Algorithm 1 and the corresponding regret bound in this section also apply to the setting where there is an inventory constraint, but the initial inventory level is "sufficient," as indicated by its fluid approximation (i.e., when $p^o > p^c$; see more discussions in Section 5). In such settings, we simply run Algorithm 1 for $t = 1, 2, \ldots, T$, but stop the algorithm when we run out of the inventory.

We first introduce the next technical lemma on the revenue gap between $\ell_{t,i^*} - \ell_{t,i_t}$ for any i^*, where $\ell_{ti} := p(i)f_t(p(i))$. This revenue gap plays an important role in regret analysis. At a high level, Lemma 1 is a self-bounding inequality: It upper bounds the expected regret of Algorithm 1 using the sampling weights $\{w_{ti}\}$ the algorithm itself produces over the T time periods.

Lemma 1. *Let $\{\ell_t\}_{t=1}^T \subseteq \mathbb{R}^{N_1}$ be defined as $\ell_{ti} = p(i)f_t(p(i))$, where f_t is the demand curve for customers in selling period t. Let η_0 be the step size chosen as $\eta_0 = 0.07/\overline{p}$. Then, for every $i^* \in [N_1]$, the revenue gap between the price $p(i^*)$ and the offered prices by our algorithm $p(i_t)$ for $t = 1, \ldots, T$ can be bounded as follows*:

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i^*} - \ell_{t,i_t}\right] \le 32\overline{p} \times \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i \ne i^*}\sqrt{\frac{w_{ti}}{t}}\right]. \tag{7}$$

The proof of Lemma 1 is based on the analysis of the works of Audibert et al. (2014), Zimmert et al. (2019), and Zimmert and Seldin (2021). The details of the proof are provided in the supplementary material. It is worth noting that traditionally a self-bounding inequality like the one in Equation (7) is used together with a Cauchy-Schwartz inequality to obtain $\widetilde{O}(\sqrt{NT})$ cumulative regret upper bound over *adaptively adversarial* bandit instances, by simply following $\sum_t\sum_i\sqrt{w_{ti}/t} \le \sqrt{NT} \times \sqrt{\sum_t\sum_i w_{ti}/t} \le \sqrt{NT} \times \sqrt{\sum_t 1/t} = O(\sqrt{NT\log T})$. However, in our case, an $\widetilde{O}(\sqrt{NT})$ upper bound is not sufficient because the number of discretized prices N scales polynomially with T. Instead, we combine Lemma 1 with the "gap" result in Lemma 3 to obtain much sharper regret upper bounds, as we detail in Lemma 2 later.

Based on Lemma 1, we obtain a more explicit upper bound of the term $\mathbb{E}[\sum_{t=1}^T \ell_{t,i^*} - \ell_{t,i_t}]$ in the next lemma by partitioning the prices to a subset $\mathcal{I}$ and its complement. Later, when we upper bound the expected regret of Algorithm 1, note that $\mathcal{I}$ corresponds to the prices that are close to $p(i^*)$ (corresponding to the $O(\sqrt{|\mathcal{I}|T})$ term in Lemma 2), and $\mathcal{I}^c$ to those prices far away from $p(i^*)$, whose expected revenues are much lower than the optimal reward (corresponding to the $\sum_{i\notin\mathcal{I}, i\ne i^*}\frac{O(\log T)}{\Delta\mu_i}$ term).

Lemma 2. *Let $\{\ell_t\}_{t=1}^T \subseteq \mathbb{R}^{N_1}$ be defined in Lemma 1, and suppose that the sampling probabilities $\{w_t\}_{t=1}^T \subseteq \Delta^{N_1-1}$ satisfy Equation (7). Let $i^* = \arg\max_{i\in[N_1]} p(i)f_0(p(i))$, and for every $i \ne i^*$, define $\Delta\mu_i = p(i^*)f_0(p(i^*)) - p(i)f_0(p(i)) \ge 0$. Then, for any subset $\mathcal{I} \subseteq [N_1]\backslash\{i^*\}$,*

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i^*} - \ell_{t,i_t}\right] &\le 64\overline{p}\sqrt{|\mathcal{I}|T} + \sum_{t=1}^{T}\sum_{i\notin\mathcal{I}, i\ne i^*}\frac{1}{2}\frac{(32\overline{p})^2}{t\Delta\mu_i} + \frac{1}{2}\overline{p}\varepsilon T\\ &\le O(\sqrt{|\mathcal{I}|T} + \varepsilon T) + \sum_{i\notin\mathcal{I}, i\ne i^*}\frac{O(\log T)}{\Delta\mu_i}.\end{aligned} \tag{8}$$

Lemma 2 is a key step in the analysis of the FTRL algorithm, and it is worth comparing Lemma 2 with similar results established in the existing literature under the ε-contamination model. For simplicity, we consider the case of $\mathcal{I} = \emptyset$, such that $|\mathcal{I}| = 0$. The closest result is from Zimmert and Seldin (2021), who also analyzed this FTRL procedure. A regret upper bound of $O\left(\sum_{i\ne i^*}\frac{\log T}{\Delta\mu_i} + \sqrt{\sum_{i\ne i^*}\frac{\log T}{\Delta\mu_i}\varepsilon T}\right)$ was shown in Zimmert and Seldin (2021). In Remark 1, we observe that this regret bound from Zimmert and Seldin (2021) yields worse regret bounds in the pricing with demand-learning setting. The work by Lykouris et al. (2018) and Gupta et al. (2019) studied different policies for

MAB with corruptions and derived regret upper bounds of $O\Big(\sum_{i\neq i^*}\frac{N\varepsilon T+\log T}{\Delta\mu_i}\log(NT)\Big)$ and $O\Big(\sum_{i\neq i^*}\frac{\log T}{\Delta\mu_i}+N\varepsilon T\Big)$, respectively. Again, in Remark 1, we show that these two regret bounds are not optimal in our setting.

Additionally, the reason for Lemma 2 separating prices (arms) into two sets is motivated by the observation that, for prices close to p^o (i.e., those belonging to $\mathcal{I}$), the suboptimality gap $\Delta\mu_i$ is small, and, hence, $1/(t\Delta\mu_i)$ would be too large. For these prices, it is better to use the gap-independent bound $\widetilde{O}(\sqrt{|\mathcal{I}|T})$. For prices (arms) that are far away from p^o (i.e., those not in $\mathcal{I}$), the gap-dependent result of $1/(t\Delta\mu_i)$ is important in establishing a tight regret bound that does not depend on the total number of arms (prices) in the discretization set.

In the proof of Lemma 2, the important technical step is to use the arithmetic mean-geometric mean (AM-GM) inequality to extract a $\frac{1}{2}\sum_t\sum_{i\notin\mathcal{I}}\Delta\mu_i w_{ti}$ term, in Equation (11). The $\frac{1}{2}$ coefficient is not important here, but it needs to be strictly smaller than one so that it could be absorbed by the left-hand side of the inequality. The other part arising from the AM-GM inequality would then have a nice dependency on $1/(t\Delta\mu_i)$.

Proof of Lemma 2. First, because there are at most εT outlier periods, we have that

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq i^*}\Delta\mu_i w_{ti}\right]-\overline{p}\varepsilon T\le\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,i^*}-\ell_{t,i_t}\right]$$
$$\le\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq i^*}\Delta\mu_i w_{ti}\right]+\overline{p}\varepsilon T. \qquad (9)$$

On the other hand, by Lemma 1, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,i^*}-\ell_{t,i_t}\right]\le\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq i^*}32\overline{p}\sqrt{\frac{w_{ti}}{t}}\right]$$
$$=\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i\in\mathcal{I}}32\overline{p}\sqrt{\frac{w_{ti}}{t}}+\sum_{i\notin\mathcal{I},\,i\neq i^*}32\overline{p}\sqrt{\frac{w_{ti}}{t}}\right]$$
$$\le\mathbb{E}\left[\sum_{t=1}^{T}32\overline{p}\sqrt{|\mathcal{I}|}\sqrt{\frac{\sum_{i\in\mathcal{I}}w_{ti}}{t}}\right]$$
$$+\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\notin\mathcal{I},\,i\neq i^*}32\overline{p}\sqrt{\frac{w_{ti}}{t}}\right], \qquad (10)$$
$$\le\mathbb{E}\left[\sum_{t=1}^{T}32\overline{p}\sqrt{|\mathcal{I}|}\sqrt{\frac{\sum_{i\in\mathcal{I}}w_{ti}}{t}}\right]$$
$$+\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\notin\mathcal{I},\,i\neq i^*}\frac{1}{2}\Delta\mu_i w_{ti}+\frac{1}{2}\frac{(32\overline{p})^2}{t\Delta\mu_i}\right], \qquad (11)$$
$$\le 64\overline{p}\sqrt{|\mathcal{I}|T}+\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\notin\mathcal{I},\,i\neq i^*}\frac{1}{2}\Delta\mu_i w_{ti}+\frac{1}{2}\frac{(32\overline{p})^2}{t\Delta\mu_i}\right]. \qquad (12)$$

Here, in Equation (10), we apply the Cauchy-Schwartz inequality that $\sum_{i\in\mathcal{I}}\sqrt{w_{ti}}\le\sqrt{|\mathcal{I}|}\sqrt{\sum_{i\in\mathcal{I}}w_{ti}}$ and the key term $\sum_{t=1}^{T}\sum_{i\notin\mathcal{I},i\neq i^*}\frac{1}{2}\Delta\mu_i w_{ti}$ can be viewed as a shadow regret; in Equation (11), we use the AM-GM inequality; in Equation (12), we use the fact that $\sum_{i\in\mathcal{I}}w_{ti}\le\sum_{i=1}^{N_1}w_{ti}\le 1$ and $\sum_{t=1}^{T}1/\sqrt{t}\le 2\sqrt{T}$.

Combining Equations (9) and (12), we have that

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,i^*}-\ell_{t,i_t}\right]\le 64\overline{p}\sqrt{|\mathcal{I}|T}+\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{2}\frac{(32\overline{p})^2}{t\Delta\mu_i}\right]$$
$$+\frac{1}{2}\left(\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,i^*}-\ell_{t,i_t}\right]+\overline{p}\varepsilon T\right),$$

because $\mathbb{E}[\sum_{t=1}^{T}\sum_{i\notin\mathcal{I},i\neq i^*}\Delta\mu_i w_{ti}]\le\mathbb{E}[\sum_{t=1}^{T}\sum_{i\neq i^*}\Delta\mu_i w_{ti}]\le\mathbb{E}[\sum_{t=1}^{T}\ell_{t,i^*}-\ell_{t,i_t}]+\overline{p}\varepsilon T$ thanks to Equation (9). Canceling out a $\frac{1}{2}\mathbb{E}[\sum_{t=1}^{T}\ell_{t,i^*}-\ell_{t,i_t}]$ term on both sides of the above inequality and rearranging the terms, we obtain

$$\frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,i^*}-\ell_{t,i_t}\right]\le 64\overline{p}\sqrt{|\mathcal{I}|T}+\sum_{t=1}^{T}\sum_{i\notin\mathcal{I},i\neq i^*}\frac{1}{2}\frac{(32\overline{p})^2}{t\Delta\mu_i}+\frac{1}{2}\overline{p}\varepsilon T$$
$$\le O(\sqrt{|\mathcal{I}|T})+\sum_{i\notin\mathcal{I},i\neq i^*}\frac{O(\log T)}{\Delta\mu_i}+O(\varepsilon T),$$

where the second inequality holds because $\sum_{t=1}^{T}1/t=O(\log T)$. Lemma 2 is thus proved. □

The next lemma shows that for a discretized price γ-distance away from the price p^o, the revenue gap is at least quadratic in γ when γ is not too small. Intuitively, this follows from the strong concavity assumptions imposed on the reward function r (as a function of the demand rate d) and the Lipschitz continuity of both f_0 and its inverse function f_0^{-1}. The parameters M, $\overline{L}_p$, and $\underline{L}_p$ are defined in Assumptions 1 and 2.

Lemma 3. *Suppose $\underline{p}^o\le p^o$, and let $i^\sharp=\operatorname{argmin}_{i\in[N_1]}|p(i)-p^o|$. Consider any $i\in[N_1]$, $i\neq i^\sharp$, such that $|i-i^\sharp|=\gamma$. Let $\zeta=(\overline{p}-\underline{p}^o)/N_1$ be the space between neighboring prices. If $\gamma\ge\frac{1}{2}+\frac{M\overline{L}_p}{\sqrt{2}\sigma\underline{L}_p}$, then*

$$r(i)\le r(i^\sharp)-\frac{\sigma^2\underline{L}_p^2(\gamma-1/2)^2\zeta^2}{4},$$

where $r(i)=p(i)f_0(p(i))$ and $r(i^\sharp)=p(i^\sharp)f_0(p(i^\sharp))$.

Lemma 3 is another important intermediate result that facilitates our analysis of Algorithm 1. It shows that as a price $p(i)$ moves away from the optimal price $p(i^\sharp)$, the expected revenue drops significantly, and the revenue drop is further associated with the distance between i and $i^\sharp$. This creates a nice optimality-gap structure among the discretized prices, which is essential in establishing a tight regret upper bound in Theorem 1. At a higher level, the results of Lemma 3 are similar to Kleinberg and Leighton (2003, lemma 3.11), which also utilized the concavity of revenue curves to prove optimality-gap structures. The key difference between Lemma 3 and Kleinberg and Leighton (2003, lemma 3.11) is that we assume the expected revenue r is concave as a function with respect to the *demand rate* d, whereas Kleinberg and Leighton (2003) assume r is concave with respect to the *price* p. It is well-known that the former assumption/condition includes more interesting demand distributions (e.g., the exponential and logistic demand distributions) and, thus, is more widely used in dynamic pricing problems (Chen and Shi 2019). The assumption that r is concave in the demand d instead of the price p requires a more robust suboptimality-gap analysis. Below, we give the complete proof of Lemma 3.

Proof of Lemma 3. Let $d(i) := f_0(p(i))$ be the expected demands of $p(1), \cdots, p(N)$. Let also $d^o = f_0(p^o)$. By the Lipschitz continuity of f_0 and f_0^{-1} (see Assumption 2), it holds that

$$|d(i^\sharp) - d^o| \le \overline{L}_p\zeta/2 \quad \text{and} \quad |d(i) - d^o| \ge \underline{L}_p(\gamma - 1/2)\zeta.$$

Using the strong concavity and smoothness of $r(d) = df_0^{-1}(d)$ (see Assumption 3), it holds that

$$r(d(i^\sharp)) \ge r(d^o) - \frac{M^2}{2}|d^o - d(i^\sharp)|^2 \ge \frac{M^2\overline{L}_p^2\zeta^2}{8};$$
$$r(d(i)) \le r(d^o) - \frac{\sigma^2}{2}|d^o - d(i)|^2 \le \frac{\sigma^2\underline{L}_p^2(\gamma - 1/2)^2\zeta^2}{2}.$$

With the condition that $\gamma \ge \frac{1}{2} + \frac{M\overline{L}_p}{\sqrt{2}\sigma\underline{L}_p}$, the term $\frac{M^2\overline{L}_p^2\zeta^2}{8}$ is upper bounded by one-half of $\frac{\sigma^2\underline{L}_p^2(\gamma-1/2)^2\zeta^2}{2}$. Subsequently,

$$r(d(i)) \le r(d(i^\sharp)) - \frac{\sigma^2\underline{L}_p^2(\gamma - 1/2)^2\zeta^2}{4},$$

which is to be demonstrated. □

We are now ready to state our main regret upper bound for Algorithm 1.

Theorem 1. *Suppose that Algorithm 1 runs with $N_1 = \lceil T^{1/4} \rceil$, and ψ, $\{\eta_t\}$ are chosen as in Equation (6), with $\eta_0 = 0.07/\overline{p}$. Suppose also that $p^c \le \underline{p}^o \le p^o$. Then, the regret of Algorithm 1 can be upper bounded by*

$$\begin{aligned}
&\mathfrak{R}_{T,\varepsilon T}(Alg.1\ ; f_0, x_0) \\
&= \frac{1}{8}M^2\overline{L}_d^2(\overline{p} - \underline{p})^2\sqrt{T} \\
&\quad + 128\sqrt{\frac{M\overline{L}_pT}{\sigma\underline{L}_p}} + \frac{1}{2}\overline{p}\varepsilon T + \frac{2145\overline{p}^2\sqrt{T}\ln(eT)}{\sigma^2\underline{L}_p^2(\overline{p} - \underline{p}^o)^2} \\
&= \widetilde{O}(\varepsilon T + \sqrt{T}),
\end{aligned}$$

where in the $\widetilde{O}(\cdot)$ notation, we drop polynomial dependency on $p^o, p^c, \underline{p}, \overline{p}, \underline{L}_p, \overline{L}_p, M^2, \sigma^2$ and $\log T$.

Theorem 1 is based on the result in Lemma 2, with the suboptimality gaps $\Delta\mu_i$ being replaced by their lower bounds proved in Lemma 3. The other parts of the proof of Theorem 1 are rather technical and deferred to the supplementary material.

Remark 1. It is worth comparing our results with existing results on robust multiarmed bandit (Lykouris et al. 2018, Gupta et al. 2019, Zimmert and Seldin 2021) and their consequences in the dynamic pricing with demand-learning problem. First, corollary 8 of section 5.1 in Zimmert and Seldin (2021) establishes a regret upper bound of

$$O\left(\sum_{i\neq i^*}\frac{\log T}{\Delta_i} + \sqrt{\sum_{i\neq i^*}\frac{\log T}{\Delta_i}C}\right),$$

where $\{\Delta_i\}$ are the suboptimality gaps, and C is the total number of adversarially corrupted periods, which is equal to εT in our notation. Because we have $N_1 \asymp T^{1/4}$ discretized prices and $r(p^*) - r(p') \gtrsim |p - p'|^2$, we have $\sum_{i\neq i^*}\frac{\log T}{\Delta_i} = O(\sqrt{T})$, and, subsequently, the upper bound in Zimmert and Seldin (2021) leads to an

$$\widetilde{O}(\sqrt{T} + \sqrt{\varepsilon T})$$

regret, which is *considerably worse* than our $\widetilde{O}(\sqrt{T} + \varepsilon T)$ regret because $\varepsilon \in (0, 1)$ is a small number characterizing the proportion of outlier/adversarial customers. The regret upper bound established in Lykouris et al. (2018) is (omitting logarithmic factors)

$$\widetilde{O}\left(\sum_{i\neq i^*}\frac{N_1C + \log T}{\Delta_i}\log(NT)\right),$$

where N_1 is the total number of arms ($N_1 \asymp T^{1/4}$ in our setting), and $C = \varepsilon T$. This translates to an $\widetilde{O}(\sqrt{T} + \varepsilon T^{7/4})$ regret, which is considerably worse than $\widetilde{O}(\sqrt{T} + \varepsilon T)$. Finally, the regret upper bound in Gupta et al. (2019) is

$$\widetilde{O}\left(\sum_{i\neq i^*}\frac{\log T}{\Delta_i} + N_1C\right),$$

which translates to $\widetilde{O}(\sqrt{T} + \varepsilon T^{5/4})$ with $C = \varepsilon T$, $N_1 \asymp T^{1/4}$, again worse than the desired $\widetilde{O}(\sqrt{T} + \varepsilon T)$ regret upper bound.

5. Dynamic Pricing with Binding Inventory Constraints

Now, we are ready to consider the impact of the inventory constraint. As we mentioned, the inventory constraint makes this problem challenging. In particular, we consider the case when the inventory constraint in the fluid-approximation problem is binding without corruptions, reflecting the impact of inventory constraints. We shall remark that, because it is *not* possible for a retailer to know a priori whether the inventory constraint in the fluid approximation is binding or not due to the unknown demand models and the presence of outlier customers, the algorithm presented in this section should be regarded as a “subroutine” for our meta-algorithm in Section 6. In practice, the robust bisection-search algorithm presented in this section should be used together with the meta-algorithm for general inventory settings designed in Section 6.

We first state a decomposition result, which tries to separate out the consideration of inventory constraints. It shows that if the (expected) demand rates at each time period are not significantly lower than the inventory rate x_0, then the regret of the policy can be characterized accurately by $\sum_t r^* - p_t f_0(p_t)$—that is, the gap without inventory consideration.

Proposition 1. *Fix an arbitrary pricing policy π. For selling period t, define $\delta_t := \max\{0, f_0(p_t) - x_0\}$. Then, it holds that*

$$\mathfrak{R}_{T,E}(\pi; f_0, x_0) \le \mathbb{E}^{\pi}\left[\sum_{t=1}^{T} r^* - p_t f_0(p_t)\right] + \frac{1}{x_0}\left(E + \sqrt{T \ln T} + \mathbb{E}^{\pi}[\overline{\delta}]\right),$$

where $\overline{\delta} := \sum_{t=1}^{T}\delta_t$. Furthermore, assume that $\overline{\delta} \le B$ with probability $1 - O(T^{-1})$ for some constant B. Then, with probability $1 - O(T^{-1})$, the inventory level will remain positive until the last $x_0^{-1}(B + \varepsilon T + O(\sqrt{T \ln T})) = O(E + \sqrt{T \ln T} + B)$ periods, where $E = \varepsilon T$ is the total number of outlier customers.

Proposition 1 is a technical result that will be used later in our regret analysis. The quantity δ_t plays an important role in this decomposition result. In particular, when the mean demand $f_0(p_t)$ at the price p_t exceeds the inventory rate x_0, δ_t will be positive and contribute to the regret because it is necessary to consider the inventory constraint. Otherwise, we simply truncate δ_t at zero and ignore the effect of inventory in the regret.

At a higher level, the proof of Proposition 1 is as follows: Using standard concentration inequalities, we can show that the cumulative realized demand over the first t selling periods is upper bounded by $x_0 t + \mathbb{E}[\overline{\delta}] + E + O(\sqrt{t \log T})$. Let T^+ be the index of the time period when inventory is completely depleted. Thus, the total demand over the first T^+ periods exceeds $x_0 T$. This implies that $\mathbb{E}[T - T^+]$ is upper bounded by $x_0^{-1}(\mathbb{E}[\overline{\delta}] + \varepsilon T + O(\sqrt{T \log T}))$. This implication would then lead to Proposition 1 because the regret incurred by the last $T - T^+$ periods is at most $\overline{p}(T - T^+)$. A complete proof of Proposition 1 is given in the supplementary material.

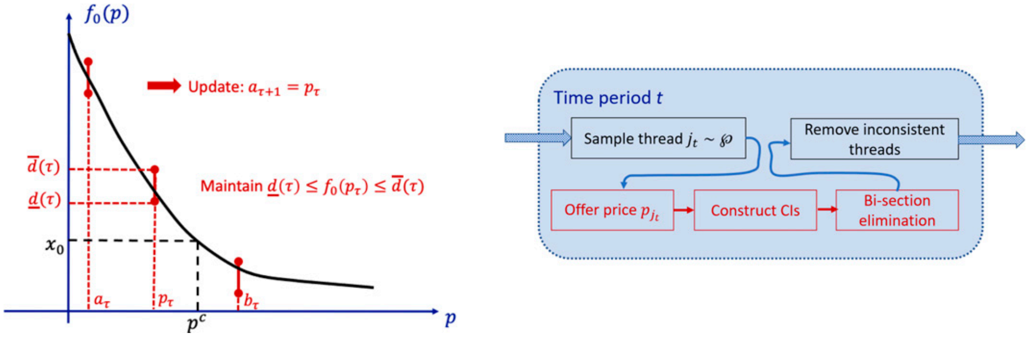
When the inventory is sufficient, the inventory constraint is ineffective. To measure the sufficiency of the inventory, we leverage the fluid approximation from Gallego and Van Ryzin (1994). As discussed in Section 3.2, the optimal solution to the fluid approximation is the maximum of the revenue-maximization price and the clearance price—that is, $p^* = \max(p^o, p^c)$. The condition $p^c \equiv f_0^{-1}(x_0) > p^o$ indicates that the total inventory $x_T = x_0 T$ is small, and, thus, the retailer needs to charge a higher price than p^o. This section mainly focuses on the insufficient-inventory case with $p^c > p^o$. On the other hand, when the inventory is sufficient with $p^c < p^o$, one can directly adopt the FTRL algorithm (see Algorithm 1), which achieves the same regret guarantee in Theorem 1. For ease of presentation, we will refer to the sufficient-inventory or “unconstrained-inventory” case as the $p^c < p^o$ case (note that unconstrained inventory corresponds to $p^c = \underline{p}$).

In the case of $p^c > p^o$, we first propose a (robust) bisection approach to quickly identify prices that lead to the total expected demand close to the inventory level x_0, with the knowledge of the outlier proportion ε. We then extend the robust bisection approach to the case when ε is *not* known, using a multithread coordination strategy. We provide Figure 2 to illustrate the key algorithmic idea of the robust bisection search and the schematic for the multithread robust search algorithm (Algorithm 3).

5.1. A Robust Bisection-Search Policy

Before we present our robust bisection-search policy, we first illustrate why the standard bisection-search algorithm will not work in the setting with adversarial outliers. Indeed, without consideration of outlier customers, a popular method for identifying p^c is to use *bisection search*, with each iteration building upper and lower confidence bounds of $f_0(p_{mid}) \in [\underline{d}_{mid}, \overline{d}_{mid}]$ at the price-interval median p_{mid}. Then, it shortens the price interval to the right of p_{mid} if $\underline{d}_{mid} > x_0$ or to the left of p_{mid} if $\overline{d}_{mid} < x_0$, based on the monotonicity of the demand-rate function f_0. In the presence of outlier customers, such a bisection protocol would easily fail when the first few customers are outliers, leading to significantly inaccurate estimates of $f_0(p_{mid})$ and a subsequently incorrect decision of whether $p^c > p_{mid}$ or $p^c < p_{mid}$. Once the first interval shrinkage is incorrect, there is no opportunity for the bisection-search algorithm to correct itself in later time periods, as the price p^c is eliminated once and for all. Therefore, the bisection-search strategy without consideration of outlier customers must suffer an $\Omega(T)$ linear

Figure 2. (Color online) Plots of the Key Algorithmic Idea (Left) of Algorithm 2 and Schematic (Right) of Algorithm 3



regret, even if only the first few customers exhibit outlier purchase behaviors.

Algorithm 2 (A Robust Bisection Algorithm for the $p^c > p^o$ Case with Known ε)

```
1: Parameters: time horizon $T$, inventory level $x_0 \in (0,1]$, an upper bound of $\varepsilon$.
2: Initialize: $I(1) = [a(1), b(1)] = [0,1]$, $C_\varepsilon(0) = 1$;
3: for $\tau = 1, 2, \cdots,$ until $T$ selling periods are reached do
4:    Set $T(\tau) = 2^\tau$, $p(\tau) = (a(\tau) + b(\tau))/2$ and $\widehat{d}(\tau) = 0$;
5:    for the next $T(\tau)$ selling periods, or until $T$ selling periods are reached do
6:       Offer price $p(\tau)$ and observe realized demand $d_t$;
7:       Update $\widehat{d}(\tau) \leftarrow \widehat{d}(\tau) + d_t$;
8:    end for
9:    Compute $[\underline{d}(\tau), \bar{d}(\tau)] = \frac{\widehat{d}(\tau)}{T(\tau)} \pm C_\varepsilon(\tau)$, where $C_\varepsilon(\tau) = \min\{1, \varepsilon T/T(\tau)\} + \sqrt{\log(2T^2)/T(\tau)} + 2\log(2T^2)/T(\tau)$;
10:   Update $I(\tau+1) = [a(\tau+1), b(\tau+1)]$ as follows:
       - If $x_0 < \underline{d}(\tau)$, then set $a(\tau+1) = p(\tau)$ and $b(\tau+1) = b(\tau)$;
       - If $x_0 > \bar{d}(\tau)$, then set $a(\tau+1) = a(\tau)$ and $b(\tau+1) = p(\tau)$;
       - If $x_0 \in [\underline{d}(\tau), \bar{d}(\tau)]$, then set $a(\tau+1) = \max\{a(\tau), p(\tau) - 2\bar{L}_d C_\varepsilon(\tau)\}$ and $b(\tau+1) = \min\{b(\tau), p(\tau) + 2\bar{L}_d C_\varepsilon(\tau)\}$;
11: end for
```

Pseudo-code of the proposed robust bisection-search algorithm is given in Algorithm 2. At a high level, Algorithm 2 uses the monotonicity of the demand curve f_0 to perform bisection and accurately identifies the price p^c at which $f_0(p^c) = x_0$. More specifically, Algorithm 2 maintains intervals $I(\tau)$ containing the clearance price p^c with high probability and attempts to halve the lengths of $I(\tau)$ at the end of every epoch τ. If $f_0(d_\tau)$ is deemed to be higher than x_0, then the algorithm moves the left endpoint of $I(\tau)$ to its midpoint, and vice versa. Because $f_0(d_\tau)$ is unknown, Algorithm 2 uses (corrupted) samples to construct upper and lower edges $[\underline{d}(\tau), \bar{d}(\tau)]$, which contain $f_0(d_\tau)$ with high probability. Finally, the third case in step 10 in Algorithm 2 reflects the possibility of the algorithm *not* being able to determine (with high probability) whether $f_0(d_\tau)$ is higher or lower than x_0, which happens when $f_0(p_\tau)$ is very close to x_0. In such a case, however, Algorithm 2 is still able to shorten the interval $I(\tau)$ considerably, by utilizing the (inverse) Lipschitz continuity of f_0, meaning $f_0(p_\tau)$ being close to x_0 must imply the demand rate at the left endpoint of $I(\tau)$ being much higher than x_0.

The following theorem upper bounds the regret of Algorithm 2 under the $p^o < p^c$ setting, with the value of the outlier proportion ε (or its upper bound) being known and fed into the algorithm as an input parameter.

Theorem 2. *Suppose $p^c > p^o$. The regret of Algorithm 2 can be upper bounded by*

$$
\begin{aligned}
&\Re_{T,\varepsilon T}(Alg.2\ ; f_0, x_0)\\
&\le (x_0^{-1} + 4\bar{p}\bar{L}_p\bar{L}_d)\varepsilon T \log_2 T\\
&\quad + (x_0^{-1} + 14\bar{p}\bar{L}_p)\sqrt{T\ln(2T^2)} + 6\bar{p}\bar{L}_p\bar{L}_d \ln^2(2T^2) + O(1)\\
&= \widetilde{O}(\varepsilon T + \sqrt{T}),
\end{aligned}
$$

where in the $\widetilde{O}(\cdot)$ notation, we drop polynomial dependency on $\underline{p}, \bar{p}, p^c, p^o, \underline{L}_p, \bar{L}_p, \underline{L}_d, \bar{L}_d$ and $\log T$.

To establish Theorem 2, we will introduce several technical lemmas. The first lemma shows that for an epoch τ, $[\underline{d}(\tau), \bar{d}(\tau)]$ covers the expected demand at the price $p(\tau)$ with high probability (see Lemma EC.2 in the supplementary material). The second lemma provides an upper bound on the length of the searching interval $|I(\tau)| := (b(\tau) - a(\tau))$ (see Lemma EC.5 in the supplementary material). We will then establish the result in Theorem 2 based on these two key lemmas. Because of space constraints, all the details of the technical lemmas and the proofs are relegated to the supplementary material.

5.2. Multithread Coordination of Bisection Searches

In the unknown-ε case with $p^c > p^o$, we will run multiple threads of Algorithm 2 with different ε values and carefully coordinate them, so that an adaptive regret upper bound can be achieved.

More specifically, let $\{\varepsilon_j = 2^{-j}\}_{j=1}^J$ be a geometric grid of ε values, with $\varepsilon_J = 2^{-J} = 1/\sqrt{T}$. Recall that T, as used in Algorithm 2, is the total number of selling periods on which the algorithm runs. Each thread $j \in [J]$ is associated with one independent instantiation of Algorithm 2, with its own intervals $I_j(\tau)$ and confidence bands $C_{\varepsilon_j}(\tau)$. An algorithm that coordinates these threads in parallel is presented in Algorithm 3.

Algorithm 3 (A Multithread Bisection-Search Algorithm for the $p^c > p^o$ Case with Unknown ε)

```
1: Parameters: time horizon $T$, inventory level $x_0 \in (0,1]$.
2: Let $\{\varepsilon_j = 2^{-j}\}_{j=1}^J$ be a geometric grid with $J = \lceil \log_2 \sqrt{T} \rceil$;
3: For each $j \in [J]$ initialize $I_j(1) = [a_j(1), b_j(1)] := [\underline{p}, \overline{p}]$, $\Delta_{\varepsilon_j}(0) = 1$, $\wp_j := 2^{-(J-j)}/2(1-2^{-J})$;
4: for $\tau = 1, 2, \cdots,$ until $T$ selling periods are reached do
5:   Set $T(\tau) = 2^\tau$ and $T_j(\tau) = \wp_j T(\tau)$, $p_j(\tau) = (a_j(\tau) + b_j(\tau))/2$, $\widehat{d}_j(\tau) = 0$ for all $j \in [J]$;
6:   for each of the next $T(\tau)$ selling period $t$ do
7:     Sample $j_t \in [J]$ randomly such that $\Pr[j_t = j] = \wp_j$;
8:     Offer price $p_{j_t}(\tau)$ and observe realized demand $d_t$;
9:     Update $\widehat{d}_{j_t}(\tau) \leftarrow \widehat{d}_{j_t}(\tau) + d_t$;
10:  end for
11:  for $j = 1, 2, \cdots, J$ do
12:    $[\underline{d}_j(\tau), \overline{d}_j(\tau)] = \frac{\widehat{d}_j(\tau)}{T_j(\tau)} \pm C_{\varepsilon_j}(\tau)$, where $C_{\varepsilon_j}(\tau) = \min\left\{1, \frac{\varepsilon_j T}{T_j(\tau)}\right\} + \sqrt{\frac{\log(2T^2)}{T_j(\tau)}} + \frac{2\log(2T^2)}{T_j(\tau)}$;
13:    Update $I_j(\tau+1)$ based on $[\underline{d}_j(\tau), \overline{d}_j(\tau)]$ using step 10 in Algorithm 2;
14:    If $j > 1$ then further update $I_j(\tau+1) \leftarrow I_j(\tau+1) \cap I_{j-1}(\tau+1)$;
15:  end for
16:  If $I_J(\tau+1) = \emptyset$, then set $J \leftarrow J-1$ and recalculate $\{\wp_j\}_{j=1}^J$ as in step 3;
17: end for
```

The high-level idea behind Algorithm 3 is as follows: Threads with $\varepsilon_j \geq \varepsilon$ are more conservative with the elimination of suboptimal prices, at the cost of potentially larger regret incurred per period (due to insufficient elimination). Hence, we sample threads of larger ε_j with smaller probability to control the total regret incurred by these threads. On the other hand, threads with $\varepsilon_j < \varepsilon$ are more aggressive with the elimination of prices and, therefore, incur much less regret, *provided that the targeted clearance price p^c is not eliminated*. To check whether p^c is potentially eliminated, we compare active prices of a thread with those in threads with larger ε_j values and eliminate a thread (i.e., decreasing the value of J) whenever inconsistency is spotted, meaning that p^c is likely to be eliminated by mistake in thread J.

In the following, we state important technical lemmas in the analysis of the regret of Algorithm 3. Because of the technical nature of the proofs, we relegate all proofs in this section to the supplementary material and only describe high-level intuitions and explanations of lemmas we present.

First, we state a lemma showing that, for those threads with ε_j levels higher than the true outlier proportion ε, the upper and lower bounds $[\underline{d}_j(\tau), \overline{d}_j(\tau)]$ and the bisection intervals $I_j(\tau)$ behave well (with high probability) in these threads.

Lemma 4. *With probability $1 - O(T^{-2} J \log T)$, the following holds for all τ and j such that $\varepsilon_j \geq \varepsilon$:*

1. $\underline{d}_j(\tau) \leq f_0(p_j(\tau)) \leq \overline{d}_j(\tau)$;
2. $p^c \in I_j(\tau)$;
3. $|I_j(\tau)| = (b_j(\tau) - a_j(\tau)) \leq 2\overline{L}_d C_{\varepsilon_j}(\tau - 1)$.

The first property in Lemma 4 states that, with high probability, the lower and upper confidence edges $[\underline{d}_j(\tau), \overline{d}_j(\tau)]$ contain the true demand rate $f_0(p_j(\tau))$ evaluated at the offered price $p_j(\tau)$. The second and third properties of Lemma 4 show that, with high probability, the bisection intervals $I_j(\tau)$ will never exclude the target clearance price p^c, and the lengths of the bisection intervals decrease over epochs. Note that Lemma 4 only applies to those threads with $\varepsilon_j \geq \varepsilon$ and may not hold true for the other threads with smaller ε_j values.

The next corollary shows that the J value, which could potentially be decreased in step 16 of Algorithm 3, will not fall below the level of the true corruption level ε with high probability.

Corollary 1. *With probability $1 - O(T^{-2} J \log T)$, the parameter J in Algorithm 3 satisfies $\varepsilon_J \leq \varepsilon$ throughout.*

We are now ready to state our main regret theorem. The proof of our main theorem based on Lemma 4 and Corollary 3 is relegated to the supplementary material.

Theorem 3. *Suppose $p^c > p^o$. Then, the regret of Algorithm 3 can be upper bounded by*

$$\begin{aligned}
&\mathfrak{R}_{T_3,\varepsilon T}(Alg.3\ ; f_0, x_0) \\
&\leq (44\overline{p}\overline{L}_d^2 + x_0^{-1})\varepsilon T \ln(2T^2) \\
&\quad + (116\overline{p}\overline{L}_d^2 + x_0^{-1})\sqrt{T}\ln(2T^2) + 32\overline{p}\overline{L}_d^2 \ln(2T^2) + O(1) \\
&\leq \widetilde{O}(\varepsilon T + \sqrt{T}),
\end{aligned}$$

where in the $\widetilde{O}(\cdot)$ notation, we drop polynomial dependency on $\underline{p}, \overline{p}, p^c, p^o, \underline{L}_p, \overline{L}_p, \underline{L}_d, \overline{L}_d$ and $\log T$.

Before presenting the proof of Theorem 3, we remark that the main algorithmic idea behind this multithread searching of unknown ε is due to Lykouris et al. (2018)

for MAB. In Lykouris et al. (2018), the number of actions (i.e., number of arms) is finite and small, whereas in the dynamic pricing, there are an infinite number of arms because the price range (action space) is infinite. Later, in Krishnamurthy et al. (2021), the multithread approach was extended to a more complicated multidimensional setting. The analysis of Algorithm 3, on the other hand, is slightly different from the works of Lykouris et al. (2018) and Krishnamurthy et al. (2021), as it utilizes the Lipschitz and "inverse" Lipschitz properties of the demand-rate function $f_0(p)$ (see Assumption 1).

We also explain why we adopt the multithread coordination idea in Algorithm 3, specifically for the $p^o < p^c$ setting, while using a Follow-the-Regularized-Leader strategy for the unconstrained inventory (or $p^c < p^o$) setting in the previous section. As we discussed in Remark 1, a multithread coordination algorithm will inevitably lead to suboptimal regret (e.g., $\widetilde{O}(\sqrt{T} + \varepsilon T^{7/4})$ instead of $\widetilde{O}(\sqrt{T} + \varepsilon T)$) in the unconstrained- inventory case. Indeed, the multithread coordination analysis in Lykouris et al. (2018) is not optimal in terms of dependency on the number of arms, and the number of discretized prices is fairly large in the dynamic pricing (i.e., $N_1 \asymp T^{1/4}$). On the other hand, there are technical difficulties applying bisection search directly to the unconstrained-inventory case. In the $p^o < p^c$ case, we only need to check whether the expected demand rate at the bisection-search midpoint p_{mid} is above or below x_0. In contrast, in the unconstrained-inventory setting, we need to estimate the *derivative* of the revenue function at p_{mid} in order to decide whether $p_{mid} < p^o$ or $p_{mid} > p^o$. This makes bisection search harder to implement for the unconstrained-inventory case, as estimating the derivative of the unknown revenue curve is challenging, especially in the presence of outlier customers.

Finally, we discuss why the Follow-the-Regularized-Leader strategy for the unconstrained-inventory setting is unlikely to be applicable to the bisection-search threads in the $p^o < p^c$ setting. First, using FTRL or similar methods to coordinate several independent *bandit-learning* threads usually leads to suboptimal regret guarantees, as illustrated in Cheung et al. (2018) and Agarwal et al. (2017). It is also difficult to predict or estimate the final incurred regret of a bisection-search thread under certain ε values, rendering Bayesian optimization and GP-UCB-type methods (see, e.g., Toscano-Palmerin and Frazier 2018) inapplicable to our setting.

Proof of Theorem 3. We will prove this theorem by upper bounding the regret incurred by all threads separately. Recall that, by Proposition 1, it suffices to upper bound $\mathbb{E}[\sum_{t=1}^{T}|f_0(p_t) - x_0|]$, where $x_0 = f_0(p^c)$. We will also condition the rest of the proof on the success events of Lemma 4 and Corollary 1, which occur with probability $1 - O(T^{-2}J\log T) = 1 - O(T^{-2}\log^2 T)$.

First, consider threads $j \in [J]$ with $\varepsilon_j \geq \varepsilon$. By Lemma 4, we have that $p^c \in I_j(\tau)$ for all epochs, and, furthermore, $|I_j(\tau)| \leq 2\overline{L}_d C_{\varepsilon_j}(\tau - 1)$. This means that, for every selling period during which thread j is selected (with probability $\wp_j = 2^{-(J-j)}/2(1 - 2^{-J}) \leq 2^{1-J+j}$), the regret incurred by $|f_0(p_j(\tau)) - x_0|$ is upper bounded by $|I_j(\tau)|$. Because thread j is selected in epoch τ for $T_j(\tau) = \wp_j T(\tau) \leq 2^{\tau+1-J+j}$ selling periods in expectation, the total regret incurred on thread j is upper bounded by

$$
\begin{aligned}
\sum_{\tau=1}^{\tau_0} \wp_j T(\tau)|I_j(\tau)| &\leq \sum_{\tau=1}^{\tau_0} 2^{\tau+1-J+j} \\
&\quad \times 2\overline{L}_d\left[\min\left\{1, \frac{2^{-j}T}{2^{\tau-1}}\right\} + \sqrt{\frac{\ln(2T^2)}{2^{\tau-J+j}}} + \frac{2\ln(2T^2)}{2^{\tau-J+j}}\right] \\
&\leq 8\overline{L}_d \sum_{\tau=1}^{\tau_0}\left(2^{-J}T + \sqrt{2^{\tau-J+j}\ln(2T^2)} + 2\ln(2T^2)\right) \\
&\leq 8\overline{L}_d\left(2^{-J}T\tau_0 + 3.5\sqrt{2^{\tau_0-J+j}\ln(2T^2)} + 2\ln(2T^2)\right) \\
&\leq 8\overline{L}_d\left(\sqrt{T}\log_2 T + 3.5\sqrt{T\ln(2T^2)} + 2\ln(2T^2)\right),
\end{aligned} \tag{13}
$$

where the last inequality holds because $2^{-J} \leq 2^{-j}$, and τ_0 being the last epoch must satisfy $|T(\tau_0)| = 2^{\tau_0} \leq T$.

We next consider those threads with $\varepsilon_j < \varepsilon$. Let τ_j be the last epoch, such that $I_j(\tau_j) \neq \emptyset$. For any epoch $\tau \leq \tau_j$, recall that $p_j(\tau)$ is the price advertised by thread j. Now, let $j^* < J$ be the thread with the smallest $\varepsilon_{j^*} \geq \varepsilon$, implying that $\varepsilon_{j^*} \geq \varepsilon > \varepsilon_{j^*+1} \geq \varepsilon_j$. Because $I_j(\tau) \subseteq I_{j^*}(\tau)$, we conclude that the mismatched demand $|f_0(p_j(\tau)) - x_0|$ is upper bounded by $|I_{j^*}(\tau)| \leq 2\overline{L}_d C_{\varepsilon_{j^*}}(\tau - 1)$. Because thread j is selected in epoch τ for, at most, $T(\tau)$ selling periods in total, the total regret incurred on thread j is upper bounded by

$$
\begin{aligned}
\sum_{\tau=1}^{\tau_j} T(\tau)|I_{j^*}(\tau)| &\leq \sum_{\tau=1}^{\tau_j} 2^{\tau} \times 2\overline{L}_d C_{\varepsilon_{j^*}}(\tau - 1) \\
&\leq \sum_{\tau=1}^{\tau_j} 2^{\tau} \times 2\overline{L}_d\left[\min\left\{1, \frac{2^{-j^*}T}{2^{\tau-1}}\right\} + \sqrt{\frac{\ln(2T^2)}{2^{\tau-J+j^*}}}\right. \\
&\qquad \left. + \frac{2\ln(2T^2)}{2^{\tau-J+j^*}}\right] \\
&\leq 4\overline{L}_d \sum_{\tau=1}^{\tau_j}\left(2^{-j^*}T + \sqrt{2^{\tau+J-j^*}\ln(2T^2)}\right. \\
&\qquad \left. + 2^{J-j^*}\ln(2T^2)\right) \\
&\leq 4\overline{L}_d\left(2\tau_j\varepsilon T + 3.5\sqrt{2^{\tau_j+J-j^*}\ln(2T^2)}\right. \\
&\qquad \left. + 2^{J-j^*}\tau_j\ln(2T^2)\right),
\end{aligned} \tag{14}
$$

$$
\begin{aligned}
&\leq 4\overline{L}_d(2\varepsilon T\log_2 T + 3.5T^{3/4}\sqrt{2\varepsilon\ln(2T^2)} \\
&\qquad + 2\varepsilon\sqrt{T}\ln(2T^2)),
\end{aligned} \tag{15}
$$

$$\le 4\overline{L}_d(2\varepsilon T\log_2 T + 2.5(\varepsilon T + \sqrt{T})\sqrt{\ln(2T^2)} + 2\varepsilon\sqrt{T}\ln(2T^2)), \tag{16}$$

$$\le 22\overline{L}_d\varepsilon T\ln(2T^2) + 18\overline{L}_d\sqrt{T}\ln(2T^2). \tag{17}$$

Here, Equation (14) holds because $2^{-j^*} = \varepsilon_{j^*} \in [\varepsilon, 2\varepsilon]$ and $2^{-J} = \varepsilon_J \ge 1/\sqrt{T}$. Equation (15) holds because $\tau_j \le \tau_0 \le \log_2 T$ and $2^J = 1/\varepsilon_J \le \sqrt{T}$. Equation (16) holds because $T^{3/4}\varepsilon^{1/2} = \sqrt{\varepsilon T} \times T^{1/4} \le (\sqrt{\varepsilon T} + T^{1/4})^2/4 \le 0.5(\varepsilon T + \sqrt{T})$, thanks to the AM-GM inequality.

Finally, combine Equations (13) and (17) and note that the regret of Algorithm 3 is upper bounded by $(\max\{p^c, \overline{L}_d\} + 1)\mathbb{E}[\sum_t |f_0(p_t) - x_0|] + x_0^{-1}(\varepsilon T + \sqrt{T\ln T})$ (see, e.g., Eq. (EC.25) in the supplementary material). We have that

$$\begin{aligned}
&\mathfrak{R}_{T_3,\varepsilon T}(Alg.3\,; f_0, x_0)\\
&\le O(1) + x_0^{-1}(\varepsilon T + \sqrt{T\ln T})\\
&\quad + (\max\{p^c, \overline{L}_d\} + 1) \times\\
&\Big[8\overline{L}_d\Big(\sqrt{T}\log_2 T + 3.5\sqrt{T\ln(2T^2)} + 2\ln(2T^2)\Big)\\
&\quad + 22\overline{L}_d\varepsilon T\ln(2T^2) + 18\overline{L}_d\sqrt{T}\ln(2T^2)\Big]\\
&\le (44\overline{p}\overline{L}_d^2 + x_0^{-1})\varepsilon T\ln(2T^2) + (116\overline{p}\overline{L}_d^2 + x_0^{-1})\sqrt{T}\ln(2T^2)\\
&\quad + 32\overline{p}\overline{L}_d^2\ln(2T^2) + O(1)\\
&\le \widetilde{O}(\varepsilon T + \sqrt{T}).
\end{aligned}$$

Theorem 3 is thus proved. □

6. Meta-Algorithm for the General Inventory Setting

The previous section considers the insufficient-inventory setting with $p^c \equiv f_0^{-1}(x_0) > p^o$. However, in practice, the true demand rate f_0 is unknown. Thus, the retailer has no prior knowledge about the relationship between p^c and p^o. In this section, we will develop a meta-algorithm that combines our previously presented methods without knowing the relative relationship between p^c and p^o. Again, we note that the unconstrained-inventory case can be viewed as a special case of $p^c < p^o$, and our FTRL algorithm in Section 4 also fits in the case of $p^c < p^o$. Because of space constraints, all proofs to technical lemmas in this section are relegated to the supplementary material.

6.1. Key Ideas

The main idea of the aggregation meta-algorithm is to use a partial exploration subroutine to obtain estimates of p^c, p^o and a carefully designed mechanism to invoke the partial exploration subroutine to mitigate the existence of adversarial customers. In the partial exploration subroutine (see Section 6.2 and Algorithm 4), we use a simple discretization idea to discretize the entire pricing interval $[\underline{p}, \overline{p}]$ into N_3 evenly spaced prices and estimate $\widehat{p}^o, \widehat{p}^c$ on the discretized prices by testing a total of T_3 selling periods (N_3, T_3 are tunable algorithm parameters). Clearly, the larger N_3 and T_3 are, the more accurate the estimation of $\widehat{p}^o, \widehat{p}^c$, but the larger regret the partial exploration subroutine incurs.

Then, in the main aggregation algorithm (see Section 6.2 and Algorithm 5), we invoke the partial exploration subroutine iteratively. For iteration (epoch) ζ, the length of the epoch (i.e. the number of time periods in the epoch) is $2^\zeta T_0$, which is geometrically increasing with ζ. On the other hand, the probability of a time period being used in the pure exploration Algorithm 4 is $1/\sqrt{T(\zeta)}$, which is decreasing with ζ. This means that for later epochs ζ, the total number of time periods devoted to estimate $\widehat{p}^o(\zeta), \widehat{p}^c(\zeta)$ increases (because $T(\zeta) \times 1/\sqrt{T(\zeta)} \to \infty$ as $\zeta \to \infty$), ensuring that after a certain epoch ζ, the partial exploration subroutine will return sufficiently accurate $\widehat{p}^c(\zeta), \widehat{p}^o(\zeta)$ estimates so that the relationship between p^o and p^c can be reliably determined for all remaining time periods. On the other hand, the exploration probability $1/\sqrt{T(\zeta)}$ decreases as ζ increases, which makes sure that the cumulative regret incurred during time periods assigned to Algorithm 4 is small, even though the epoch lengths $T(\zeta)$ increase geometrically. We also use randomly assigned time periods for partial exploration to avoid the concentration of adversarial customers in the assigned exploration periods.

Note also that, when there is no corruption ($\varepsilon = 0$), neither doubling trick nor randomized exploration is needed to aggregate the two subalgorithms. However, in the presence of an adaptive adversary, we need to use randomized exploration to evenly sample corrupted periods and also a doubling trick to keep track of the current progress of p^o and p^c estimates because corruption could be concentrated in *the block of initial time periods.*

6.2. Partial Exploration over Selected Selling Periods

Algorithm 4 (Partial Exploration for Crude Estimates of p^c and p^o)

```
**Parameters**: selected subset of selling periods $\mathcal{T}$ with $|\mathcal{T}| = T_3$, intervals $N_3$, $T_0 = \lceil\sqrt{T}\rceil$.
Let $p(1), \cdots, p(N_3) \in [\underline{p}, \overline{p}]$ be evenly spaced points on $[\underline{p}, \overline{p}]$, with $p(1) = \underline{p}$, $p(N_3) = \overline{p}$;
Initialize $\widehat{d}(1), \cdots, \widehat{d}(N_3) = 0$;
**for** every $t \in \mathcal{T}$ **do**
    Select $i \in [N_3]$ uniformly at random;
    Offer price $p_t = p(i)$ and observe realized demand $d_t$;
    Update $\widehat{d}(i) \leftarrow \widehat{d}(i) + d_t$;
**end for**
For every $i \in [N_3]$ compute $\widetilde{d}(i) = N_3\widehat{d}(i)/T_0$;
**Output**: $\widehat{p}^c = p(\widehat{i}^c)$, $\widehat{p}^o = p(\widehat{i}^o)$, where $\widehat{i}^c = \mathrm{argmin}_{i\in[N_3]} |\widetilde{d}(i) - x_0|$ and $\widehat{i}^o = \mathrm{argmax}_{i\in[N_3]} p(i)\widetilde{d}(i)$.
```

To distinguish between the $p^c < p^o$ and $p^c > p^o$ cases, we use randomized partial exploration to get crude estimates $\widehat{p}^c$ and $\widehat{p}^o$. Our robust partial exploration method is different from the common "initial exploration" strategy in the existing dynamic pricing literature (see, e.g., Besbes and Zeevi 2009, Broder and Rusmevichientong 2012, and Wang et al. 2014). More specifically, in the prior literature Wang et al. (2014), Besbes and Zeevi (2009), and Broder and Rusmevichientong (2012), the explorations are done in a centralized manner, utilizing the first $T_0 \ll T$ selling periods. In the presence of outlier customers, however, such an approach is likely to fail, as the outlier customers might concentrate on the beginning of the time periods. To overcome this challenge, we design an algorithm based on doubling epochs and perform *randomized* exploration (i.e., designating whether a selling period is used for exploration at random) to hedge against the possibility of clustered outlier periods. Our partial exploration strategy is presented in Algorithm 4, which will be used as the core building block for our meta-algorithm introduced in the next subsection. Our algorithm is closely related to randomized exploration strategy in robust multiarmed-bandit algorithms (Gupta et al. 2019). Nevertheless, one key difference is that in Gupta et al. (2019), the randomized exploration is applied to a discrete set of arms, whereas in Algorithm 4, the exploration is applied to coordinate two separate bandit algorithms.

We provide the theoretical properties for Algorithm 4. We first upper bound the deviation of $\widehat{p}^c, \widehat{p}^o$ from p^c, p^o, using the "relative density" of the outlier periods (indicated by $\iota_t = 1$) among the exploration phases, denoted as Z in Lemma 5. The larger Z is, the more outlier periods there are in the exploration phases, and, thus, the larger errors of $|\widehat{p}^c - p^c|$ and $|\widehat{p}^o - p^o|$ are expected.

Lemma 5. *Let $\widehat{p}^c, \widehat{p}^o$ be the output of Algorithm 4. Suppose also that $\sum_{t\in\mathcal{T}} \iota_t \le ZT_3$ with probability $1 - O(T^{-2})$ for some $Z > 0$. Then, with probability $1 - O(T^{-2})$, it holds that*

$$|\widehat{p}^c - p^c| \le \overline{L}_d\left[\frac{\overline{L}_p(\overline{p}-\underline{p})}{N_3} + 2H_Z(N_3,T_3)\right];$$

$$|\widehat{p}^o - p^o| \le \frac{2\overline{L}_d}{\sigma^2}\left[\frac{M\overline{L}_p(\overline{p}-\underline{p})}{\sqrt{2}N_3} + \sqrt{2\overline{p}H_Z(N_3,T_3)}\right],$$

where

$$H_Z(N_3,T_3) = \min\{1, Z\} + \sqrt{\frac{N_3 \log(2N_3T^2)}{T_3}} + \frac{2N_3 \log(2N_3T^2)}{T_3}.$$

Algorithm 5 (A Meta-Algorithm Combining $p^c < p^o$ and $p^c > p^o$ Policies)

```
1: Parameters: time horizon $T$, initial inventory level $x_T = x_0T$, policies $\pi^o$ (Algorithm 1) and $\pi^c$ (Algorithm 3).
2: Initialize: $\widehat{p}^c(0) = \underline{p}$ and $\widehat{p}^o(0) = \overline{p}$; $T_0 = \lceil\sqrt{T}\rceil$;
3: for each epoch $\zeta = 1, 2, \cdots$ until inventory runs out, or $T$ selling periods are reached do
4:     Let $T(\zeta) = 2^\zeta T_0$ and let $\mathcal{T}(\zeta)$ be the next $T(\zeta)$ selling periods;
5:     For each selling period in $\mathcal{T}(\zeta)$, place it in an exploration phase set $\mathcal{G}(\zeta)$ with probability $1/\sqrt{T(\zeta)}$;
6:     if $\widehat{p}^c(\zeta-1) < \widehat{p}^o(\zeta-1)$ then
7:         Run policy $\pi^o$ with $\underline{p}^o = (\widehat{p}^c(\zeta-1) + \widehat{p}^o(\zeta-1))/2$ and initial inventory $x_\zeta = x_0T(\zeta)$ on $\mathcal{T}(\zeta)\backslash\mathcal{G}(\zeta)$; *
8:         Run Algorithm 4 with $T_3 = |\mathcal{G}(\zeta)|$ and $N_3 = \lceil\sqrt{T_3}\rceil$ on $\mathcal{G}(\zeta)$;
9:     else
10:        Run policy $\pi^c$ with initial inventory $x_\zeta = x_0T(\zeta)$ on $\mathcal{T}(\zeta)\backslash\mathcal{G}(\zeta)$;*
11:        Run Algorithm 4 with $T_3 = |\mathcal{G}(\zeta)|$ and $N_3 = \lceil\sqrt{T_3}\rceil$ on $\mathcal{G}(\zeta)$;
12:    end if
13:    Update estimates $\widehat{p}^o(\zeta), \widehat{p}^c(\zeta)$ from Algorithm 4 run on $\mathcal{G}(\zeta)$;
14: end for
```

*If the designated inventory level x_ζ runs out, then offer price $\overline{p}$ in the rest of epoch τ, during which $\widehat{\pi}^o$ or $\widehat{\pi}^c$ are run.

6.3. A Meta-Policy Combining $p^c < p^o$ and $p^c > p^o$ Policies and the Lower Bound

Now, we are ready to introduce the meta-policy that combines the cases of $p^c < p^o$ and $p^c > p^o$. Suppose we have access to two policies: The first policy in Algorithm 1, denoted as π^o, achieves $\widetilde{O}(\varepsilon T + \sqrt{T})$ regret over T selling periods under the condition that $p^c < p^o$; and the second policy in Algorithm 3, denoted as π^c, achieves $\widetilde{O}(\varepsilon T + \sqrt{T})$ regret under the condition that $p^c > p^o$. The first policy π^o also requires a parameter $\underline{p}^o$ that is between p^c and p^o. In step 7 of Algorithm 5, we simply set $\underline{p}^o$ to be $(\widehat{p}^c + \widehat{p}^o)/2$, where $\widehat{p}^c$ and $\widehat{p}^o$ are estimated prices from Algorithm 4. Our meta-algorithm is presented in Algorithm 5, and its regret bound is presented in the next theorem. A schematic figure of the meta-policy is also given in Figure 3.

Theorem 4. *Suppose $p^c \ne p^o$, and the policies π^o, π^c satisfy $\mathfrak{R}_{T'',\varepsilon T}(\pi^o; f_0, x_0) = \widetilde{O}(\varepsilon T + \sqrt{T''})$ and $\mathfrak{R}_{T'',\varepsilon T}(\pi^c; f_0, x_0) = \widetilde{O}(\varepsilon T + \sqrt{T''})$ under the conditions of $p^c < p^o$ and $p^c > p^o$, respectively. Then, the regret of the Meta-Algorithm 5 can be upper bounded by*

$$\mathfrak{R}_{T,\varepsilon T}(Alg.6\,; f_0, x_0) = \widetilde{O}(\varepsilon T + \sqrt{T}),$$

for sufficiently large T, where in the $\widetilde{O}(\cdot)$ notation, we drop polynomial dependency on $p^o, p^c, \underline{p}, \overline{p}, \underline{L}_p, \overline{L}_p, M^2, \sigma^2$ and $\log T$.

Theorem 4 is optimal in the sense that no policy is capable of achieving regret lower than $\Omega(\varepsilon T + \sqrt{T})$ when there are εT outlier customers, as shown by the following theorem. Its proof is given in the supplementary material.

Theorem 5. *There exists a universal constant $C > 0$, such that, for any policy π, time horizon T, and outlier portion $\varepsilon \in (0,1)$, there exists f_0 and $x_0 \in (0,1]$ such that*

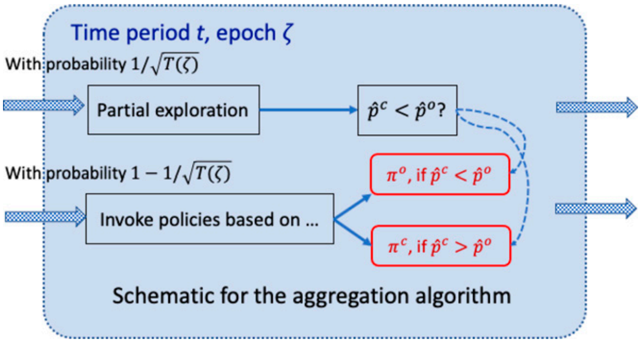
$$\Re_{T,\varepsilon T}(\pi; f_0, x_0) \geq C \times (\varepsilon T + \sqrt{T}).$$

7. Numerical Results

We use synthetic data to verify the effectiveness of our proposed algorithms on dynamic pricing with demand learning and outlier customers' purchase activities. We study a linear demand-rate model of $f_0(p) = 1 - p$ and set $x_0 = 0.8$ for the $p^c < p^o$ case and $x_0 = 0.2$ for the $p^c > p^o$ case. Simple calculations show that in the $x_0 = 0.8$ case, the optimal price with respect to the fluid approximation is $p^* = 0.5$ with expected per-period revenue $r(p^*, x_0) = 0.25$. In the $x_0 = 0.2$ case, the optimal price is $p^* = 0.8$, and the expected per-period revenue is $r(p^*, x_0) = 0.16$. For the outlier customers, we corrupt the first $\lceil \varepsilon T \rceil$ time periods with $f_t(p) \equiv 0.05$ for all advertised prices p. Our algorithm has a robust performance under different ways of corruptions.

We first test our proposed algorithms for the separate cases of $p^c < p^o$ and $p^c > p^o$, using Algorithms 1 and 3, respectively. Note that in both algorithms, the corruption level ε is *not* known a priori. For a baseline algorithm, we use a recent trisection algorithm developed in Lei et al. (2014), which has an $\widetilde{O}(\sqrt{T})$ regret without the existence of outlier customers. It should be noted that the algorithm developed in Lei et al. (2014) features a pure exploration phase, followed by two completely separate bisection procedures for the $p^c < p^o$ and $p^c > p^o$ cases. In the first comparison, we abolish the pure exploration phase of the baseline algorithm in Lei et al. (2014) because the relationship between p^c and p^o is known.

Figure 3. (Color online) A Schematic Figure for Algorithm 5, the Aggregation Meta-Policy

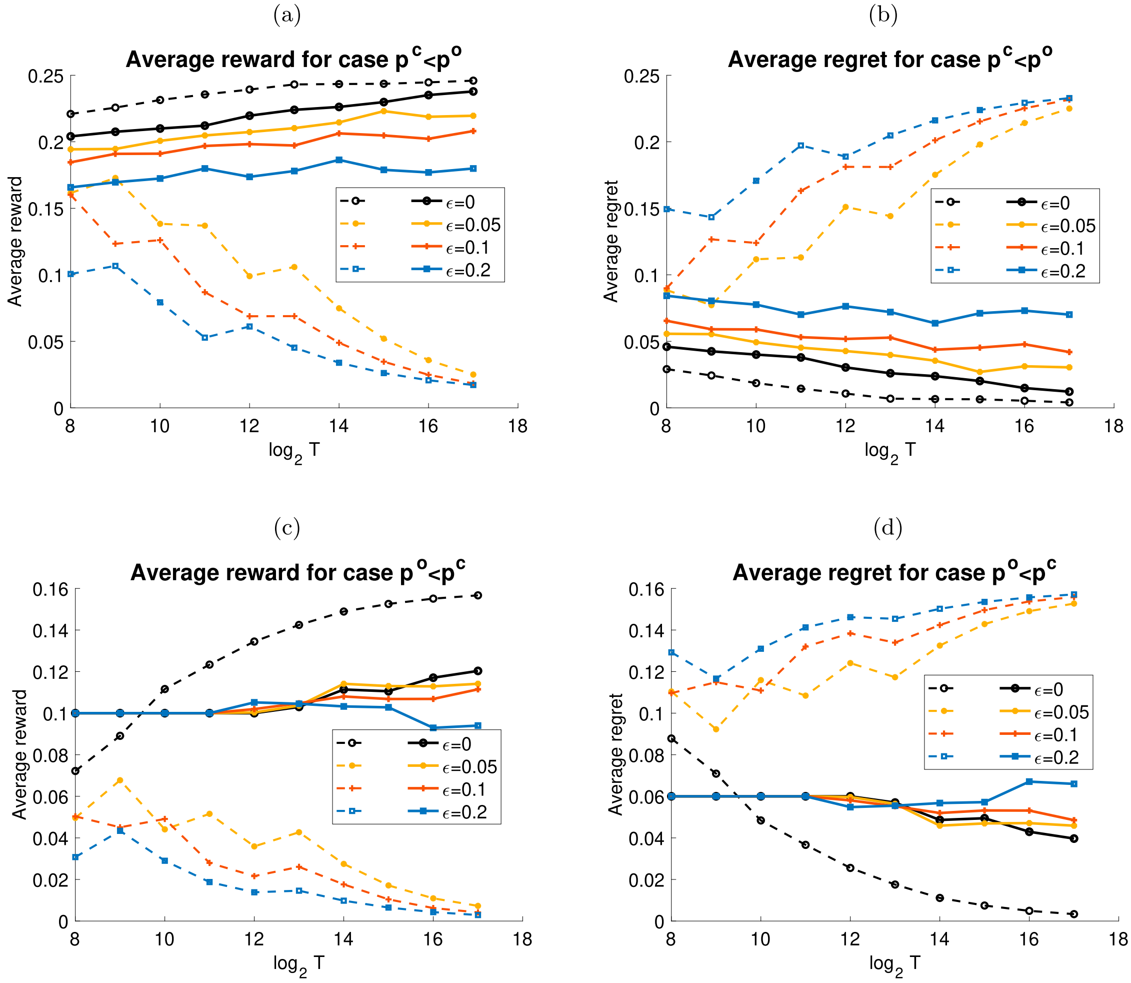

Time period t, epoch ζ
With probability $1/\sqrt{T(\zeta)}$
Partial exploration
$\hat{p}^c < \hat{p}^o$?
With probability $1 - 1/\sqrt{T(\zeta)}$
Invoke policies based on ...
π^o, if $\hat{p}^c < \hat{p}^o$
π^c, if $\hat{p}^c > \hat{p}^o$
Schematic for the aggregation algorithm

In Figure 4, we compare the average revenues ((a) and (c); the higher the better) and the average regret ((b) and (d); the lower the better) of our proposed algorithms with the baseline method under various time horizons (T) and outlier levels (ε). Each setting is repeated for 100 independent trials, and the mean average revenue/regret is reported. The standard deviation is relatively small and, thus, omitted for better visualization. As we can see in Figure 4, our proposed algorithms consistently outperform the baseline method, which does not take into consideration the presence of outlier customers. Indeed, with the presence of outliers, the bisection algorithms in Lei et al. (2014) are heavily biased toward higher prices, resulting in very large and even increasing average regret. On the other hand, our algorithms not only correctly avoided the influences of the outlier customers, but also deliver stable revenue and regret performances under varying ε settings. We also remark that without outlier customers (i.e., $\varepsilon = 0$), the regret of our proposed algorithms is worse than the baseline algorithm, due to the additional overhead incurred by adapting to unknown outlier portions over T time periods. We also note that when ε is large (e.g., $\varepsilon = 0.2$), the average revenue of our method (and average regret) does not increase (decrease) with respect to $\log_2 T$. This is because in the accumulated regret bound, the term εT becomes the dominating term, and, thus, the average regret does not decrease over T.

Next, we compare the performances of our proposed meta-algorithm (Algorithm 5) with the baseline method when the relationship between p^o and p^c is *not* known a priori and has to be learned in the process of the dynamic pricing procedure, in order to determine the right subalgorithm to use. The demand models and initial inventory-level settings are identical to the ones used in Figure 4.

Figure 5 shows the performances of our algorithm and the baseline algorithm under various T and ε settings, in which the relationship between p^o and p^c is unknown. As we can see from Figure 5, the regret of the baseline algorithm (dotted lines) is small when there is no outlier, but increases significantly with even only 5% of customers being outliers. We also see that, in the presence of outliers, the regret of the baseline algorithm does not decrease and sometimes even increases significantly with more time periods/customers available. This is partly due to the fact that the first εT customers are outliers, which could sway the baseline algorithm's judgment of the relationship between p^o, p^c, and the bisection procedure in the

Figure 4. (Color online) The Average (Avg.) Revenues ((a) and (c)) and Regret ((b) and (d)) of Our Proposed Algorithm and the Baseline Algorithm in the Separate Cases of $p^c < p^o$ (Algorithm 1 in (a) and (c)) and $p^c > p^o$ (Algorithm 3 in (b) and (d)) Under Various Outlier Levels (ε)


(a)
(b)
(c)
(d)

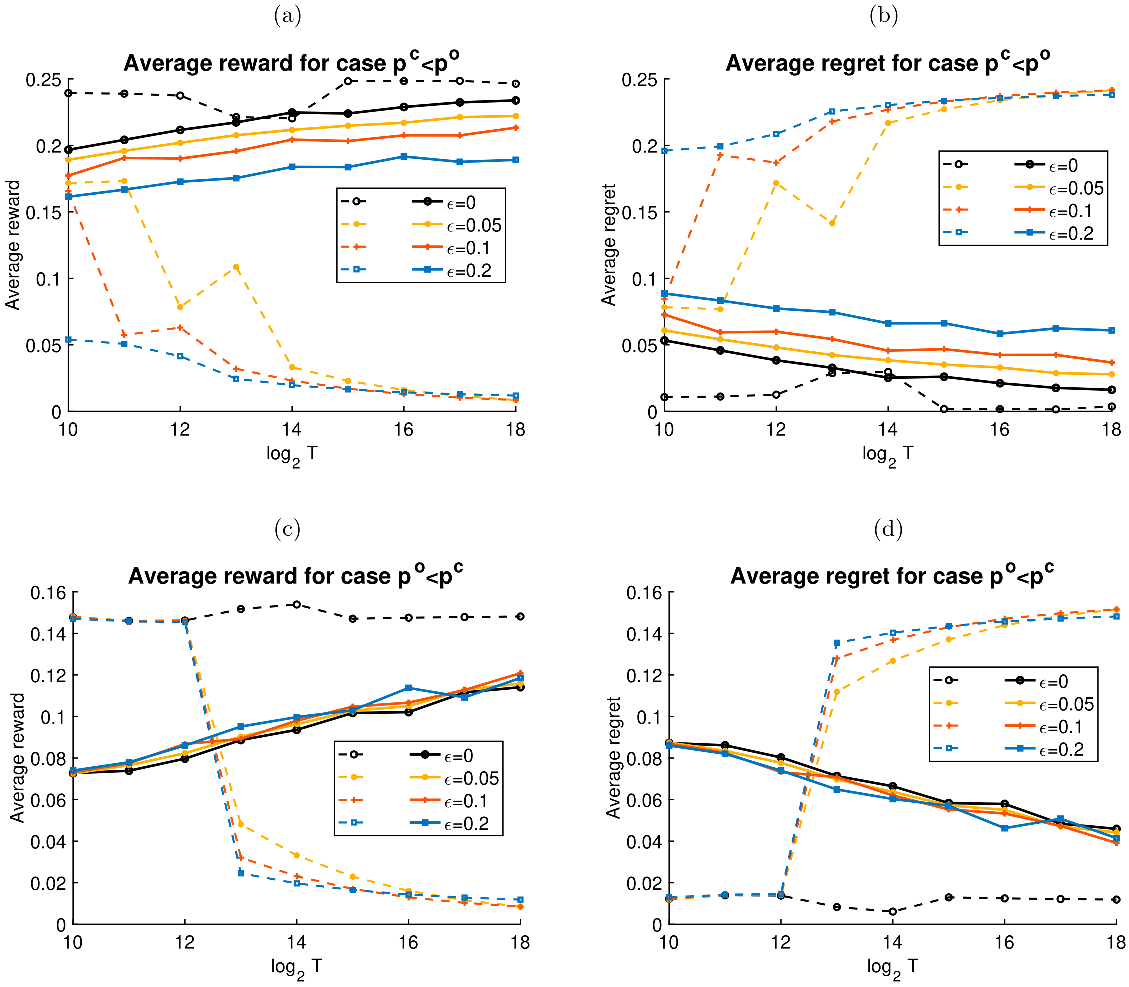
Notes. The dashed lines correspond to performances of the baseline (comparative) algorithm, and the solid lines correspond to performances of our proposed methods. More details are in the main text. (a) Avg. reward vs. $\log_2 T$. (b) Avg. regret vs. $\log_2 T$. (c) Avg. reward vs. $\log_2 T$. (d) Avg. regret vs. $\log_2 T$.

baseline algorithm might also eliminate good price candidates due to the influence of outliers. On the other hand, the performance of our proposed algorithm is much more stable in the presence of outlier customers.

We report additional sets of numerical results concerning alternative adversarial customer patterns and their associated outlier demand distributions. In this experiment (and following ones), we fix $\varepsilon = 0.1$. We first explore settings in which outlier customers are not simply concentrated among the first εT time periods. In particular, for some $\eta \in [0, 1)$, we assume $\eta\varepsilon T$ outlier customers still arrive during the first few time periods, while the remaining $(1 - \eta)\varepsilon T$ outliers are distributed uniformly at random among the remaining time periods. Results of our algorithms and their baseline competitors for different η settings are displayed in Figure 6. As we can see from Figure 6, the difference in how outlier customers are distributed has very little impact on the overall performance of our proposed methods, validating their robustness against diverse patterns of outliers.

We also study the performance of our algorithm under settings where the outlier demand distributions $g_t(\cdot)$ differ from the demand of typical customers $f_0(\cdot)$ at different levels. More specifically, recall that the demand distribution of typical customers is $f_0(p) = 1 - p$. To construct outlier demand distributions $g_t(\cdot)$ that are different from f_0, we consider $g_t(p) = 1 - e^{a_t} p$, where $a_t \sim U[0, u]$ is i.i.d. distributed from a uniform distribution on the interval of $[0, u]$, for a certain range parameter $u > 0$. Clearly, the range parameter u

Figure 5. (Color online) The Average (Avg.) Revenues ((a) and (c)) and Regret ((b) and (d)) of our Proposed Algorithm and the Baseline Algorithm Without Knowing the Relationship Between p^o and p^c (Algorithm 5)



(a)

(b)

(c)

(d)

Notes. The dashed lines correspond to performances of the baseline (comparative) algorithm, and the solid lines correspond to performances of our proposed methods. More details are in the main text. (a) Avg. reward vs. $\log_2 T$. (b) Avg. regret vs. $\log_2 T$. (c) Avg. reward vs. $\log_2 T$. (d) Avg. regret vs. $\log_2 T$.

controls how far away g_t deviates from the typical demand distribution f_0, with larger u values indicating that g_t deviates farther away from f_0 and vice versa.

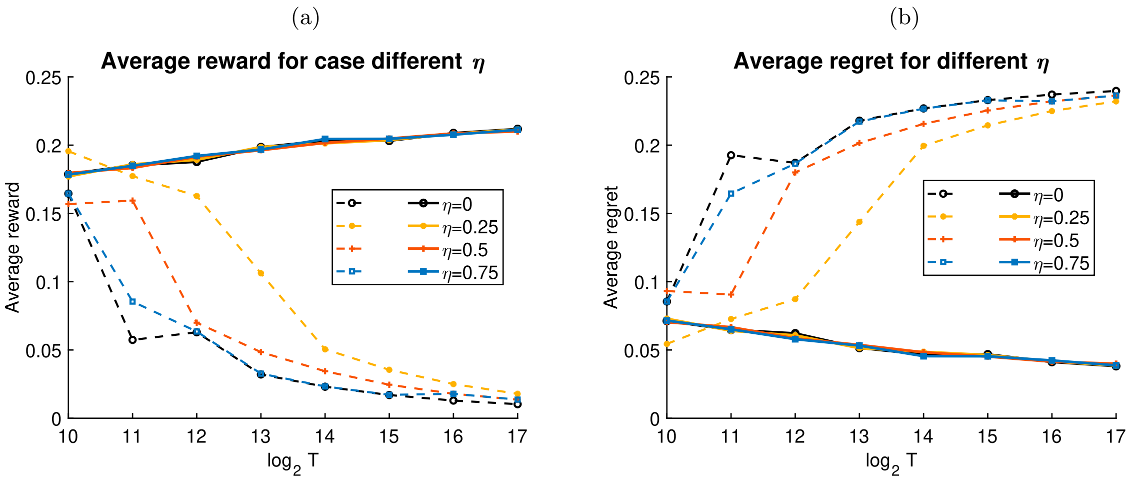
In Figure 7, we report the results of our proposed policies and their baseline competitors for problem settings with different u values. As we can see, with smaller u values (i.e., less deviation of g_t from f_0), the performance of our proposed robust policies improves, which is intuitive, as less adversarial deviation makes robust estimation of typical customers' demands easier. We also remark that the performance of the baseline methods fluctuates quite significantly compared with other experimental settings, primarily because of the highly randomized nature of adversaries (outlier customers) in this setting, which is different from our previous experimental settings, where g_t is fixed at the constant level of 0.05.

7.1. Adding Outlier Detection

In this subsection, we compare our proposed algorithm with the baseline algorithm (Lei et al. 2014) equipped with an outlier-detection component. More specifically, when an observation (p_t, d_t) arrives, the baseline algorithm first tries to detect whether the observation is an outlier. If an outlier is detected, the observation (p_t, d_t) is discarded, and it will not affect the execution flow of the underlying pricing algorithm. Otherwise, the observation is regarded as "trustworthy" by the algorithm and is handled just like it is not corrupted by any adversary.

To perform outlier detection, we consider a parametric demand model $f_0(p) = a - bp$. After every $T_0 = \lceil\sqrt{T}\rceil$

Figure 6. (Color online) The Average (Avg.) Revenue ((a)) and Regret ((b)) of Our Proposed Algorithm and the Baseline Algorithm Under Different Outlier-Occurring Patterns Quantified by η (with $\varepsilon = 0.1$)


(a)
(b)

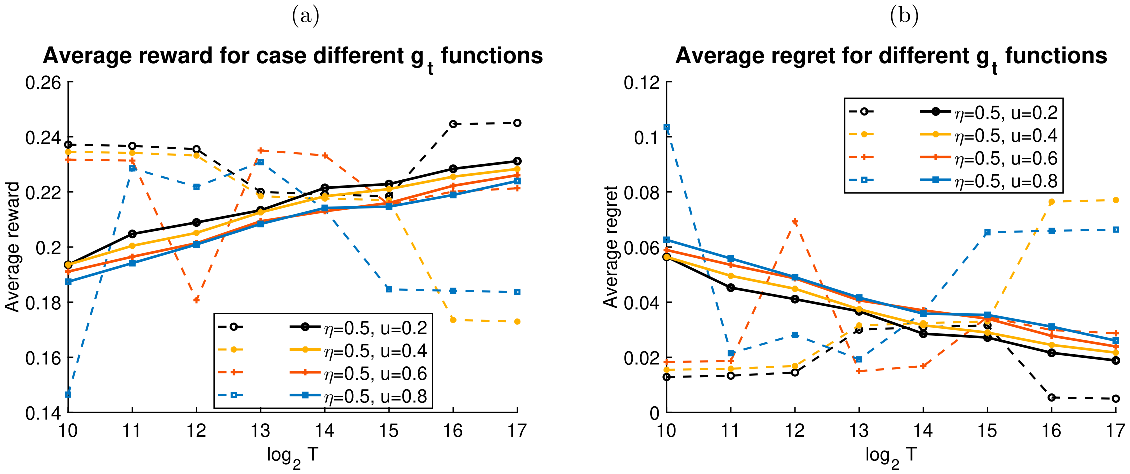
Notes. The dashed lines correspond to performances of the baseline (comparative) algorithm, and the solid lines correspond to performances of our proposed methods. More details are in the main text. (a) Avg. reward vs. $\log_2 T$. (b) Avg. regret vs. $\log_2 T$.

selling periods, the algorithm uses the collected data to fit a ordinary least-squares (OLS) estimate of the model $\widehat{f}_0(p) = \widehat{a} - \widehat{b}p$. With the estimated model, a new observation (p, d) is deemed as an outlier if $|(\widehat{d} - d)/\sqrt{MSE(1 - h_p)}| > \omega$, with $\omega = 2.0$ corresponding to the 2σ tail of a standard centered Gaussian distribution, and $\widehat{d} = \widehat{a} - \widehat{b}p$ is the estimated demand rate, *MSE* is the mean-square error of the fitted OLS model, and h_p is the leverage score corresponding to the observation (p, d). This outlier-detection procedure follows naturally the classical residual analysis in linear-regression models. To warm-start the outlier detector, for the first T_0 selling periods, the algorithm sets prices uniformly at random.

In Figure 8, we report average rewards and regrets of our proposed algorithms and compare them with the performance of the baseline algorithm equipped with the above outlier-detection component. Figure 9 further shows the percentages of observations labeled as outliers under different ε and time horizon (T) settings. It also displays the accuracy/precision and recall of the outlier-detection procedure.[1]

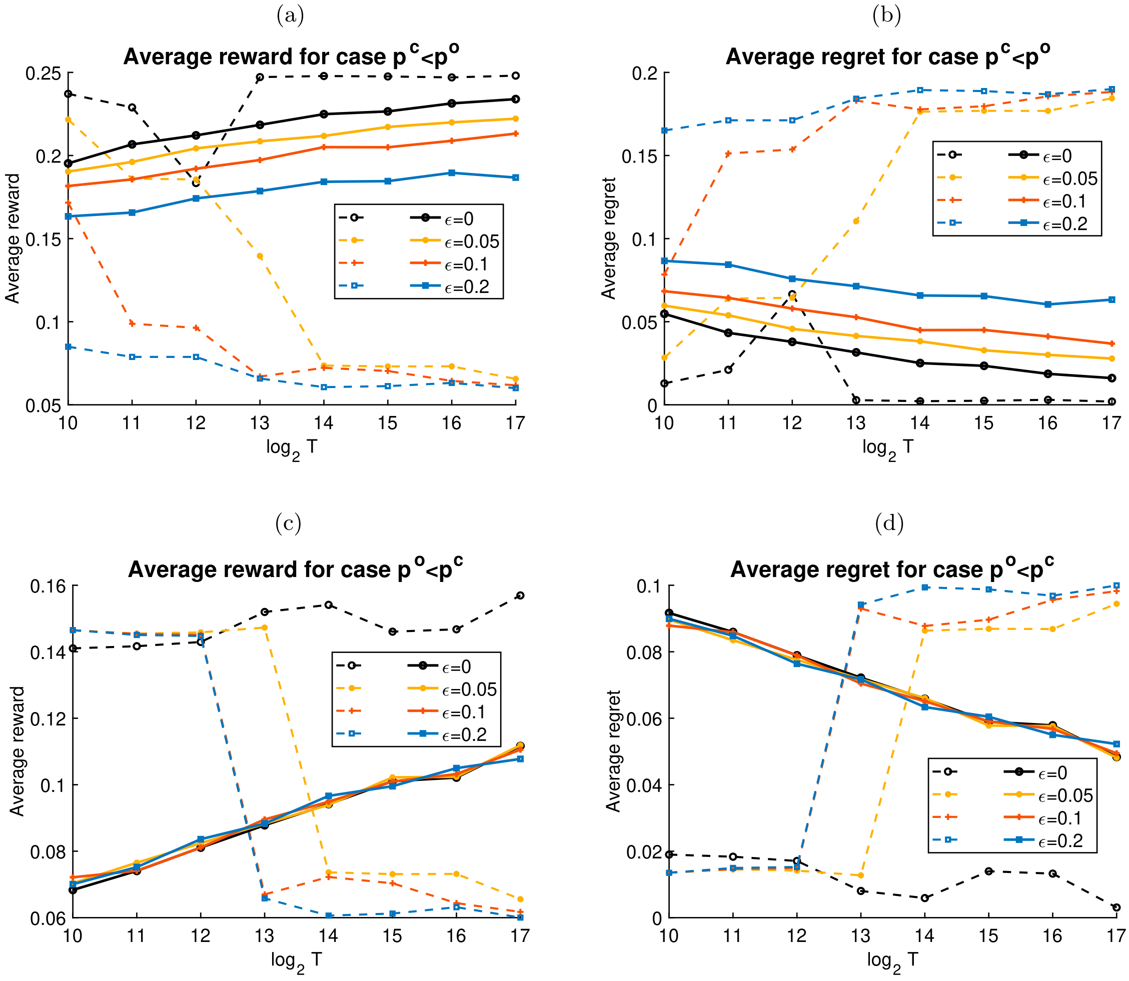
Finally, in Figure 10, we report the behavior of the outlier-detection procedure with different cutoff thresholds ω. Note that a smaller ω value means that more

Figure 7. (Color online) The Average (Avg.) Revenue ((a)) and Regret ((b)) of Our Proposed Algorithm and the Baseline Algorithm Under Different Levels of Deviations of Outlier Demand Functions Quantified by u (with $\varepsilon = 0.1, \eta = 0.5$)


(a)
(b)

Notes. The dashed lines correspond to performances of the baseline (comparative) algorithm, and the solid lines correspond to performances of our proposed methods. More details are in the main text. (a) Avg. reward vs. $\log_2 T$. (b) Avg. regret vs. $\log_2 T$.

Figure 8. (Color online) The Average (Avg.) Revenues ((a) and (c)) and Regret ((b) and (d)) of Our Proposed Algorithm and the Baseline Algorithm with an Outlier-Detection Component, Without Knowing the Relationship Between p^o and p^c (Algorithm 5)



Notes. The dashed lines correspond to performances of the baseline (comparative) algorithm, and the solid lines correspond to performances of our proposed methods. More details are in the main text. (a) Avg. reward vs. $\log_2 T$. (b) Avg. regret vs. $\log_2 T$. (c) Avg. reward vs. $\log_2 T$. (d) Avg. regret vs. $\log_2 T$.

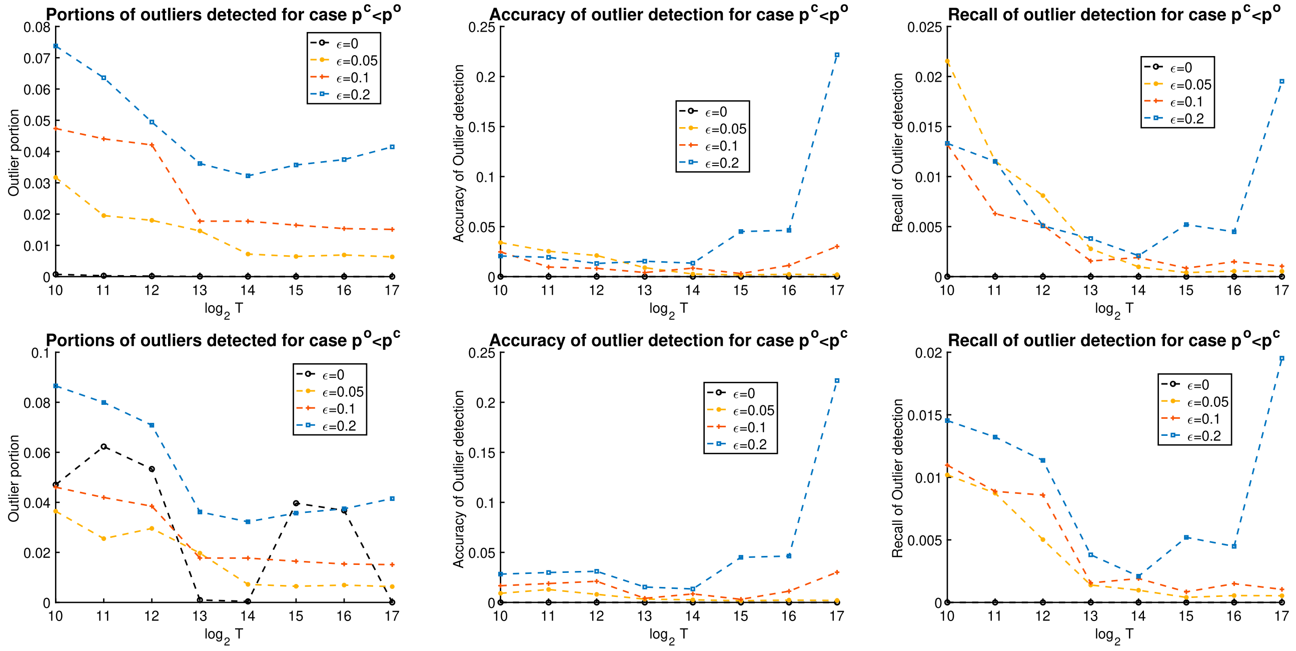
observations will be detected as outliers, which would potentially increase the recall (more actual outliers are being detected), but decrease the accuracy (more false positives in the detected outliers). As we can see from the first graph on the second row in Figure 10, with $\omega = 1.0$, as many as 35% observations are detected as outliers, and with $\omega \geq 2.0$, fewer than 10% of observations are marked as outliers, which sandwich the true outlier proportion of $\varepsilon = 0.1 = 10\%$. However, for this wide range of cutoff thresholds ω, the accuracy and recall of the outlier-detection procedure are still low. From the first row of Figure 10, the overall revenue/regret performances of the outlier-detection approach still trail behind our proposed algorithm for larger time horizons T.

8. Conclusion

In this paper, we study the robust dynamic pricing problem under an online extension of the fundamental "ε-contamination model" from statistics and machine learning. For both known and unknown outlier-proportion ε cases, we propose efficient pricing policies that are robust to adversarial corruptions and establish near-optimal regret bounds.

As for future work, it is interesting to further extend the paper to the fully adversarial setting. It is also interesting to extend the methods developed in this paper to network revenue management, where multiple products are present for sale, and their demand rates/resource consumptions are correlated. However, both are technically challenging problems.

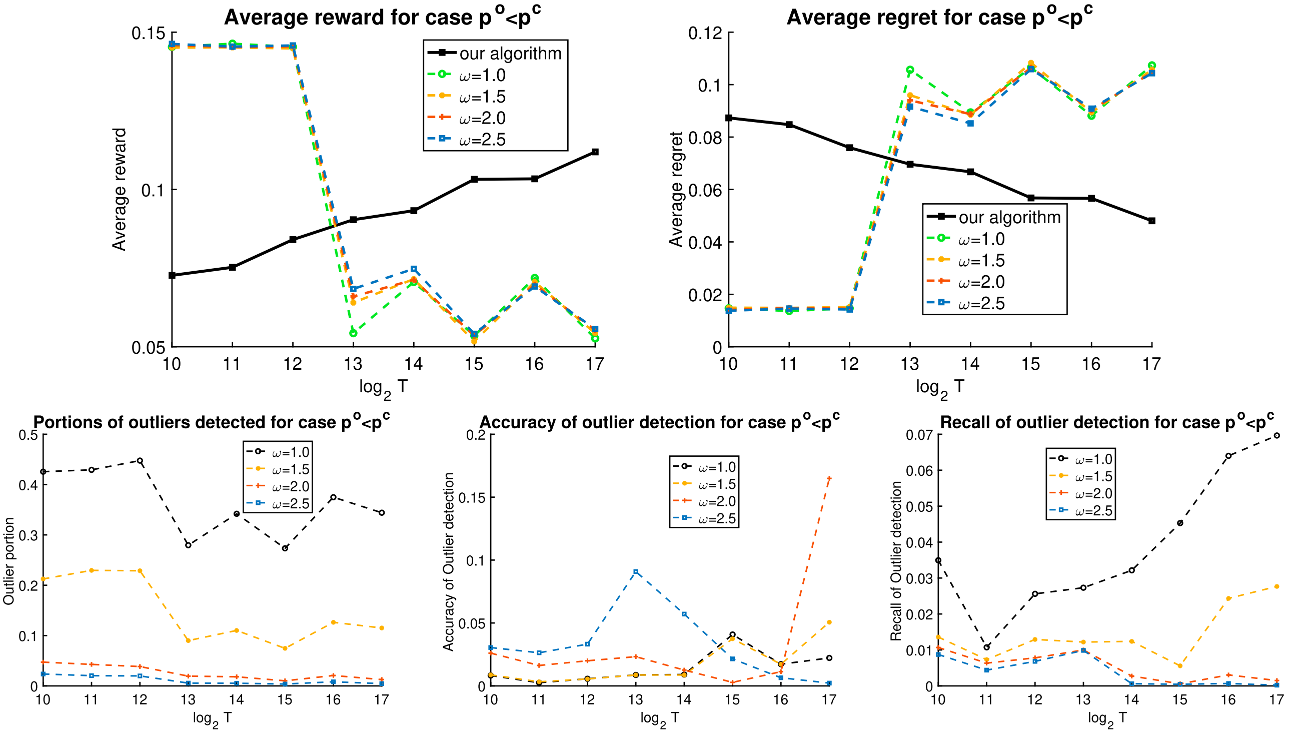
Figure 9. (Color online) The Proportion of Observations Detected as Outliers and the Accuracy/Precision and Recall of Such Detection Under Different ε Settings



Note. More details are in the main text.

In addition, our work could motivate the investigation of many other operations problems with inventory constraints under the online ε-contamination model. For example, it would be interesting to extend the robust online assortment optimization from Chen et al. (2019) to the inventory-constrained setting.

Figure 10. (Color online) Regret, Revenue, Proportions of Outliers Detected, and Accuracy/Recall of Outlier Detection Under Different ω Thresholds, with the Actual Outlier Proportion $\varepsilon = 0.1$



Note. More details are in the main text.

Acknowledgments

The authors thank the department editor, the associate editor, and the anonymous referees for many useful suggestions and feedback, which greatly improved the paper. Author names are listed in alphabetical order.

Endnote

[1] Accuracy/precision is defined as the number of true detections among all those detected as outliers, and recall is defined as the number of detections among those that are actual outliers. As we can see, the performance of the baseline algorithm is still inferior to our algorithm, despite adding the outlier-detection component. We also note that both the accuracy and recall of the standard outlier-detection procedure is quite low, which is, in general, below 10% for accuracy and hovering around 1%–2% for recall. It suggests that, with *adversarial* outlier arrival patterns, detecting outlier consumer behaviors could be very difficult in a dynamic pricing setting.

References

Agarwal A, Agarwal S, Patil P (2021) Stochastic dueling bandits with adversarial corruption. Feldman V, Ligett K, Sabato S, eds. *Proc. 32nd Internat. Conf. Algorithmic Learn. Theory* (PMLR), 217–248.

Agarwal A, Luo H, Neyshabur B, Schapire RE (2017) Corralling a band of bandit algorithms. *Conf. Learn. Theory.*

Agrawal S, Devanur NR (2014) Bandits with concave rewards and convex knapsacks. *EC'14 Proc. 15th ACM Conf. Econom. Comput.* (Association for Computing Machinery, New York), 989–1006.

Agrawal S, Devanur NR (2015) Linear contextual bandits with knapsacks. Lee DD, von Luxburg U, Garnett R, Sugiyama M, Guyon I, eds. *NIPS'16 Proc. 30th Conf. Neural Inform. Processing Systems (NeurIPS)* (Curran Associates, Red Hook, NY), 3458–3467.

Araman VF, Caldentey R (2009) Dynamic pricing for nonperishable products with demand learning. *Oper. Res.* 57(5):1169–1188.

Audibert J-Y, Bubeck S (2009) Minimax policies for adversarial and stochastic bandits. *Proc. 22nd Annu. Conf. Learn. Theory COLT.*

Audibert J-Y, Bubeck S, Lugosi G (2014) Regret in online combinatorial optimization. *Math. Oper. Res.* 39(1):31–45.

Badanidiyuru A, Kleinberg R, Slivkins A (2018) Bandits with knapsacks. *J. ACM* 65(3):1–55.

Ban G-Y, Keskin NB (2017) Personalized dynamic pricing with machine learning. Preprint, submitted May 25, https://dx.doi.org/10.2139/ssrn.2972985.

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

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

Besbes O, Gur Y, Zeevi A (2015) Non-stationary stochastic optimization. *Oper. Res.* 63(5):1227–1244.

Bitran G, Caldentey R (2003) An overview of pricing models for revenue management. *Manufacturing Service Oper. Management* 5(3):203–229.

Bogunovic I, Krause A, Jonathan S (2020) Corruption-tolerant Gaussian process bandit optimization. *Internat. Conf. Artificial Intelligence Statist. AISTATS.*

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

Bubeck S, Lee YT, Eldan R (2017) Kernel-based methods for bandit convex optimization. *STOC 2017 Proc. 49th Annu. ACM SIGACT Sympos. Theory Comput.* (Association for Computing Machinery, New York), 72–85.

Chen N, Gallego G (2021) Nonparametric learning and optimization with covariates. *Oper. Res.* 69(3):974–984.

Chen Y, Shi C (2019) Network revenue management with online inverse batch gradient descent method. Preprint, submitted February 10, https://dx.doi.org/10.2139/ssrn.3331939.

Chen M, Gao C, Ren Z (2016) A general decision theory for Huber's ϵ-contamination model. *Electronic J. Statist.* 10(2):3752–3774.

Chen Q, Jasin S, Duenyas I (2015) Real-time dynamic pricing with minimal and flexible price adjustment. *Management Sci.* 62(8): 2437–2455.

Chen X, Krishnamurthy A, Wang Y (2019) Robust dynamic assortment optimization in the presence of outlier customers. Preprint, submitted October 9, https://arxiv.org/abs/1910.04183.

Chen X, Miao S, Wang Y (2021a) Differential privacy in personalized pricing with nonparametric demand models. Preprint, submitted September 10, https://arxiv.org/abs/2109.04615.

Chen X, Simchi-Levi D, Wang Y (2022) Privacy-preserving dynamic personalized pricing with demand learning. *Management Sci.* Forthcoming.

Chen X, Owen Z, Pixton C, Simchi-Levi D (2021b) A statistical learning approach to personalization in revenue management. *Management Sci.* 68(3):1923–1937.

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

Cheung WC, Simchi-Levi D, Zhu R (2018) Hedging the drift: Learning to optimize under non-stationarity. Preprint, submitted October 5, https://dx.doi.org/10.2139/ssrn.3261050.

Cooper WL, Homem-de Mello T, Kleywegt AJ (2006) Models of the spiral-down effect in revenue management. *Oper. Res.* 54(5): 968–987.

den Boer AV (2015) Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys Oper. Res. Management Sci.* 20(1):1–18.

den Boer AV, Zwart B (2013) Simultaneously learning and optimizing using controlled variance pricing. *Management Sci.* 60(3): 770–783.

Diakonikolas I, Kamath G, Kane DM, Li J, Moitra A, Stewart A (2017) Being robust (in high dimensions) can be practical. Precup D, Teh YW, eds. *ICML'17 Proc. 34th Internat. Conf. Machine Learn.* (JMLR.org), 999–1008.

Diakonikolas I, Kamath G, Kane D, Li J, Moitra A, Stewart A (2018) Robustly learning a Gaussian: Getting optimal error, efficiently. *Proc. ACM-SIAM Sympos. Discrete Algorithms* (Society for Industrial and Applied Mathematics, Philadelphia), 2683–2702.

Elmaghraby W, Keskinocak P (2003) Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. *Management Sci.* 49(10):1287–1309.

Esfandiari H, Korula N, Mirrokni V (2018) Allocation with traffic spikes: Mixing adversarial and stochastic models. *ACM Trans. Econom. Comput.* 6(3–4):1–23.

Farias VF, Van Roy B (2010) Dynamic pricing with a prior on market response. *Oper. Res.* 58(1):16–29.

Ferreira KJ, Simchi-Levi D, Wang H (2018) Online network revenue management using Thompson sampling. *Oper. Res.* 66(6):1586–1602.

Flaxman AD, Kalai AT, McMahan HB (2004) Online convex optimization in the bandit setting: Gradient descent without a gradient. *SODA'05 Proc. Annu. ACM-SIAM Sympos. Discrete Algorithms* (Society for Industrial and Applied Mathematics, Philadelphia), 385–394.

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

Gallego G, Van Ryzin G (1997) A multiproduct dynamic pricing problem and its applications to network yield management. *Oper. Res.* 45(1):24–41.

Golrezaei N, Jaillet P, Liang JCN (2019) Incentive-aware contextual pricing with non-parametric market noise. Preprint, submitted November 8, https://arxiv.org/abs/1911.03508.

Golrezaei N, Manshadi V, Schneider J, Sekar S (2020) Learning product rankings robust to fake users. Preprint, submitted September 2, https://dx.doi.org/10.2139/ssrn.3685465.

Gupta A, Koren T, Talwar K (2019) Better algorithms for stochastic bandits with adversarial corruptions. *Proc. Conf. Learn. Theory.*

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

Hazan E, Levy K (2014) Bandit convex optimization: Toward tight bounds. Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems 27 NIPS 2014* (Curran Associates, Red Hook, NY).

Huber PJ (1964) Robust estimation of a location parameter. *Ann. Math. Statist.* 35(1):73–101.

Javanmard A, Nazerzadeh H (2019) Dynamic pricing in high-dimensions. *J. Machine Learn. Res.* 20(9):1–49.

Jin T, Luo H (2020) Simultaneously learning stochastic and adversarial episodic MDPs with known transition. *Proc. 34th Conf. Neural Inform. Processing Systems NeurIPS.*

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

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

Kleinberg R, Leighton T (2003) The value of knowing a demand curve: Bounds on regret for online posted-price auctions. *Proc. 44th Annu. IEEE Sympos. Foundations Comput. Sci. FOCS.* (IEEE, Piscataway, NJ), 594–605.

Krishnamurthy A, Lykouris T, Podimata C, Schapire RE (2021) Contextual search in the presence of irrational agents. *STOC 2021 Proc. 53rd Annu. ACM SIGACT Sympos. Theory Comput.* (Association for Computing Machinery, New York), 910–918.

Lei YM, Jasin S, Sinha A (2014) Near-optimal bisection search for nonparametric dynamic pricing with inventory constraint. Preprint, submitted October 1, https://dx.doi.org/10.2139/ssrn.2509425.

Lobel I, Leme RP, Vladu A (2018) Multidimensional binary search for contextual decision-making. *Oper. Res.* 66(5):1346–1361.

Lykouris T, Mirrokni V, Leme RP (2018) Stochastic bandits robust to adversarial corruptions. *STOC 2018 Proc. 50th Annu. ACM SIGACT Sympos. Theory Comput.* (Association for Computing Machinery, New York), 114–122.

Lykouris T, Simchowitz M, Slivkins A, Sun W (2019) Corruption robust exploration in episodic reinforcement learning. Preprint, submitted November 20, https://arxiv.org/abs/1911.08689.

Miao S, Chen X, Chao X, Liu J, Zhang Y (2019) Context–based dynamic pricing with online clustering. Preprint, submitted February 17, https://arxiv.org/abs/1902.06199.

Nambiar M, Simchi-Levi D, Wang H (2019) Dynamic learning and price optimization with endogeneity effect. *Management Sci.* 65(11):4980–5000.

Toscano-Palmerin S, Frazier P (2018) Effort allocation and statistical inference for 1-dimensional multistart stochastic gradient descent. *2018 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 1850–1861.

Wang Z, Deng S, Ye Y (2014) Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Oper. Res.* 62(2):318–331.

Wang Y, Chen X, Chang X, Ge D (2021) Uncertainty quantification for demand prediction in contextual dynamic pricing. *Production Oper. Management* 30(6):1703–1717.

Zimmert J, Seldin Y (2021) Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *J. Machine Learn. Res.* 22(28):1–49.

Zimmert J, Luo H, Wei C-Y (2019) Beating stochastic and adversarial semi-bandits optimally and simultaneously. *ICML Proc. Internat. Conf. Machine Learn.*

Xi Chen is an associate professor at the Department of Technology, Operations, and Statistics at Stern School of Business, New York University. His research interests include statistical machine learning, stochastic optimization, and data-driven operations management.

Yining Wang is an associate professor in the Department of Operations Management, Naveen Jindal School of Management, University of Texas at Dallas. His major research interests are active learning, online learning, and bandit optimization methods, as well as their applications to revenue-management problems, such as assortment optimization and dynamic pricing.

Proofs of Statements

EC.1. Proof of Lemma 1

We first review some notations and concepts necessary to our proof. Let $\mathcal{D} \subseteq \mathbb{R}^{N_1}$ be a convex open set, and $\overline{\mathcal{D}}$ be the closure of $\mathcal{D}$. We say a function $\psi : \overline{\mathcal{D}} \to \mathbb{R}$ is *Legendre* if it satisfies the following conditions:

1. ψ is strictly convex and continuously differentiable on $\mathcal{D}$;
2. $\lim_{w \to \overline{\mathcal{D}} \setminus \mathcal{D}} \|\nabla \psi(w)\| = \infty$.

It is easy to verify that our choice of $\psi(w) = \sum_{i=1}^{N_1} -\sqrt{w_i} - \sqrt{1 - w_i}$ is Legendre with $\mathcal{D} = (0,1)^{N_1}$.

Let $\mathcal{D}^* = \nabla \psi(\mathcal{D})$ be the dual space of $\mathcal{D}$. With our choice of ψ, $\mathcal{D}^* = \mathbb{R}^{N_1}$. For a Legendre function $\psi : \mathcal{D} \to \mathbb{R}$, its *Legendre-Fenchel transform* (also known as the *convex conjugate*) $\psi^* : \mathcal{D}^* \to \mathbb{R}$ is defined as

$$\psi^*(u) = \sup_{w \in \overline{\mathcal{D}}} \langle w, u \rangle - \psi(w). \tag{EC.1}$$

The following properties are standard results of convex conjugates. See for example, the reference of (Cesa-Bianchi & Lugosi 2006, Rockafellar 1970), or (Audibert et al. 2014).

Fact 1 *Suppose ψ is Legendre. Then $\psi^{**} = \psi$ and $\nabla \psi^* = (\nabla \psi)^{-1}$. Furthermore, if ψ is also twice continuously differentiable on $\mathcal{D}$, then $\nabla^2 \psi^*(u) = [\nabla^2 \psi(w)]^{-1}$ for every pair of $w = \nabla \psi^*(u)$ or $u = \nabla \psi(w)$.*

Given a Legendre function $\psi : \overline{\mathcal{D}} \to \mathbb{R}$, its *Bregman divergence* $D_\psi : \overline{\mathcal{D}} \times \mathcal{D} \to \mathbb{R}$ is defined as

$$D_\psi(x,y) = \psi(x) - \psi(y) - \langle x - y, \nabla \psi(y) \rangle. \tag{EC.2}$$

If ψ is twice continuously differentiable on $\mathcal{D}$, then by Taylor expansion with the Lagrangian remainder, we have for every $x, y \in \mathcal{D}$ that

$$D_\psi(x,y) \le \frac{1}{2}(y - x)^\top \nabla^2 \psi(z)(y - x), \tag{EC.3}$$

where $z = x + \alpha(y - x)$ for some $\alpha \in (0,1)$.

At time period t, define $\psi_t = \frac{1}{\eta_t}\psi$, where η_t is the step size (or learning rate) at time t. Define function $\phi_t : \mathbb{R}^{N_1} \to \mathbb{R}$ as

$$\phi_t(u) = \sup_{w \in \Delta^{N_1 - 1}} \langle w, u \rangle - \psi_t(w). \tag{EC.4}$$

Comparing ϕ_t with the convex conjugate ψ_t^* defined in Eq. (EC.1), the only difference lies in the additional constraint of $w \in \Delta^{N_1-1}$ in the definition of ϕ_t. The following properties are simple and elementary to verify.

Fact 2 *For any Legendre ψ_t let ψ_t^* be its convex conjugate and ϕ_t be defined in Eq. (EC.4). The following properties hold:*

1. *For any $u \in \mathbb{R}^{N_1}$, $\phi_t(u) \le \psi_t^*(u)$;*
2. *For any $u \in \mathbb{R}^{N_1}$ and $c \in \mathbb{R}$, $\phi_t(u + c\mathbf{1}) = \phi_t(u) + c$, where $\mathbf{1} = (1, \cdots, 1) \in \mathbb{R}^{N_1}$;*
3. *For any $w \in \Delta^{N_1-1}$, $\phi_t(\nabla\psi_t(w)) = \psi_t^*(\nabla\psi_t(w))$;*
4. *For any $u \in \mathbb{R}^{N_1}$, let $w^* = \arg\max_{w\in\Delta^{N_1-1}} \langle w, u\rangle - \psi_t(w)$. Then there exists $\lambda \in \mathbb{R}$ depending on u, such that $\nabla\psi_t(w^*) = u + \lambda\mathbf{1}$.*

Proof of Fact 2. The first property is obvious because ϕ_t has the same objective with ψ_t^*, but with a smaller feasible region. The second property holds because $\phi_t(u + c\mathbf{1}) = \sup_{w\in\Delta^{k-1}} \langle w, u + c\mathbf{1}\rangle - \psi_t(w) = c + \sup_{w\in\Delta_{k-1}} \langle w, u\rangle - \psi_t(w) = c + \phi_t(u)$.

To see the third property, note that $\nabla\psi_t^*(\nabla\psi_t(w)) = w$, thanks to Fact 1. This means that w is the maximizer of $\langle \cdot, \nabla\psi_t(w)\rangle - \psi_t(\cdot)$ on $\mathbb{R}^d$. Since $w \in \Delta^{N_1-1}$, it is also the maximizer of the same objective on $\mathbb{R}^d$. Hence, $\phi_t(\nabla\psi_t(w)) = \psi_t^*(\nabla\psi_t(w))$.

For the fourth property, consider the maximization question of $\max_{w\in\mathbb{R}^{N_1-1}} \langle w, u\rangle - \psi_t(w)$ and let w^* be the maximizer. Using the Lagrangian multiplier, we know that $u - \nabla\psi_t(w^*) + \lambda\mathbf{1} = 0$. Hence, $\nabla\psi_t(w^*) = u + \lambda\mathbf{1}$, which is to be demonstrated. □

Let $e_{i^*} = (0, \cdots, 0, 1, 0, \cdots, 0) \in \mathbb{R}^{N_1}$ be the indicator vector corresponding to the price $p(i^*)$. Because $i_t \sim w_t$, we know that $\mathbb{E}[\ell_{t,i_t}] = \mathbb{E}[\langle \ell_t, w_t\rangle]$. Subsequently, the regret $\mathbb{E}[\sum_{t=1}^T \ell_{t,i^*} - \ell_{t,i_t}]$ can be decomposed as follows:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T \ell_{t,i^*} - \ell_{t,i_t}\right] &= \mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, e_{i^*} - w_t\rangle\right] \\
&= \underbrace{\mathbb{E}\left[\sum_{t=1}^T \langle \ell_t, -w_t\rangle - \phi_t(\widehat{L}_{t-1}) + \phi_t(\widehat{L}_t)\right]}_{\text{the stability term}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T \phi_t(\widehat{L}_{t-1}) - \phi_t(\widehat{L}_t) + \langle \ell_t, e_{i^*}\rangle\right]}_{\text{the penalty term}}. \qquad \text{(EC.5)}
\end{aligned}$$

In the rest of this proof, we will upper bound the stability term and the penalty term separately.

EC.1.1. Upper bounding the penalty term

By definition of ϕ_t, $\phi_t(\widehat{L}_{t-1}) = \sup_{w\in\Delta^{N_1-1}}\langle w, \widehat{L}_{t-1}\rangle - \psi_t(w)$ where $\psi_t(w) = \frac{1}{\eta_t}\psi(w)$. Because $w_t \in \Delta^{N_1-1}$ is the maximizer of $\langle w, \widehat{L}_{t-1}\rangle - \psi_t(w)$, we have that

$$\phi_t(\widehat{L}_{t-1}) = \langle w_t, \widehat{L}_{t-1}\rangle - \frac{1}{\eta_t}\psi(w_t). \tag{EC.6}$$

Similarly, for any $w' \in \Delta^{N_1-1}$, it holds that

$$\phi_t(\widehat{L}_t) \geq \langle w', \widehat{L}_t\rangle - \frac{1}{\eta_t}\psi(w'). \tag{EC.7}$$

Combine Eqs. (EC.6,EC.7) and set w' in Eq. (EC.7) as $w' = w_{t+1}$ for $t < T$, and $w' = e_{i^*}$ for $t = T$. We then have

$$\begin{aligned}
&\sum_{t=1}^{T}\phi_t(\widehat{L}_{t-1}) - \phi_t(\widehat{L}_t)\\
&= \left[\sum_{t=1}^{T}\langle w_t, \widehat{L}_{t-1}\rangle - \frac{1}{\eta_t}\psi(w_t)\right] - \left[\sum_{t=1}^{T-1}\langle w_{t+1}, \widehat{L}_t\rangle - \frac{1}{\eta_t}\psi(w_{t+1})\right] - \langle \widehat{L}_T, e_{i^*}\rangle + \frac{1}{\eta_T}\psi(e_{i^*})\\
&= \left[\sum_{t=2}^{T}\left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}\right)\psi(w_t)\right] - \frac{1}{\eta_1}\psi(w_1) - \langle \widehat{L}_T, e_{i^*}\rangle + \frac{1}{\eta_T}\psi(e_{i^*}),
\end{aligned}$$

where the last equality holds because $\widehat{L}_0 = 0$ by definition. Notice that $\widehat{L}_T = \sum_{t=1}^{T}\widehat{\ell}_t$ satisfies $\mathbb{E}[\widehat{L}_T] = \sum_{t=1}^{T}\mathbb{E}[\widehat{\ell}_t] = \sum_{t=1}^{T}\ell_t$. Hence,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\phi_t(\widehat{L}_{t-1}) - \phi_t(\widehat{L}_t) + \langle \ell_t, e_{i^*}\rangle\right] &= \mathbb{E}\left[\sum_{t=2}^{T}\left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}\right)\psi(w_t) - \frac{1}{\eta_1}\psi(w_1) + \frac{1}{\eta_T}\psi(e_{i^*})\right]\\
&= \mathbb{E}\left[\sum_{t=2}^{T}\left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}\right)(\psi(w_t) - \psi(e_{i^*})) - \frac{1}{\eta_1}(\psi(w_1) - \psi(e_{i^*}))\right],
\end{aligned}\tag{EC.8}$$

where the last equality holds because the terms involving $\psi(e_{i^*})$ sum to $\frac{1}{\eta_T}\psi(e_{i^*})$.

Next, we analyze the differences between $\psi(w_t)$ and $\psi(e_{i^*})$. Recall that, for $w \in \Delta^{N_1-1}$, $\psi(w)$ is defined as $\psi(w) = \sum_{i=1}^{N_1} -\sqrt{w_i} - \sqrt{1-w_i}$. Also, because each component of e_{i^*} is either 1 or 0, we have $\psi(e_{i^*}) = -N_1$. Subsequently,

$$\psi(e_{i^*}) - \psi(w_t) = \sum_{i=1}^{N_1}\sqrt{w_{ti}} + \sqrt{1-w_{ti}} - 1 \leq \min\{\sqrt{w_{ti}}, \sqrt{1-w_{ti}}\}. \tag{EC.9}$$

Plugging Eq. (EC.9) into Eq. (EC.8), and noting that $\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} = (\sqrt{t} - \sqrt{t-1})/\eta_0 \leq 1/(\eta_0\sqrt{t})$, we have that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \phi_t(\widehat{L}_{t-1}) - \phi_t(\widehat{L}_t) + \langle \ell_t, e_{i^*}\rangle\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N_1} \frac{\min\{\sqrt{w_{ti}}, \sqrt{1-w_{ti}}\}}{\eta_0\sqrt{t}}\right] \\
&\leq \frac{1}{\eta_0}\mathbb{E}\left[\sum_{t=1}^{T}\sqrt{\frac{1-w_{ti^*}}{t}} + \sum_{i\neq i^*}\sqrt{\frac{w_{ti}}{t}}\right] = \frac{1}{\eta_0}\mathbb{E}\left[\sum_{t=1}^{T}\sqrt{\frac{\sum_{i\neq i^*} w_{ti}}{t}} + \sum_{i\neq i^*}\sqrt{\frac{w_{ti}}{t}}\right] \\
&\leq \frac{2}{\eta_0}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq i^*}\sqrt{\frac{w_{ti}}{t}}\right]. \qquad \text{(EC.10)}
\end{aligned}$$

EC.1.2. Upper bounding the stability term

Recall the definition that $w_t = \arg\max_{w\in\Delta^{N_1-1}} \langle w, \widehat{L}_{t-1}\rangle - \psi_t(w)$. By the fourth property of Fact 2, there exists $\lambda_t \in \mathbb{R}$ such that $\nabla\psi_t(w_t) = \widehat{L}_{t-1} + \lambda_t\mathbf{1}$. Subsequently,

$$\begin{aligned}
-\phi_t(\widehat{L}_{t-1}) + \phi_t(\widehat{L}_t) &= -\phi_t(\nabla\psi_t(w_t) - \lambda_t\mathbf{1}) + \phi_t(\nabla\psi_t(w_t) - \lambda_t\mathbf{1} + \widehat{\ell}_t) \\
&= -\phi_t(\nabla\psi_t(w_t)) + \phi_t(\nabla\psi_t(w_t) + \widehat{\ell}_t), \qquad \text{(EC.11)}
\end{aligned}$$

where the second equality holds thanks to the second property of Fact 2 (so that the $\lambda_t\mathbf{1}$ terms are canceled), and that $\widehat{L}_t = \widehat{L}_{t-1} + \widehat{\ell}_t$. By the first and the third properties of Fact 2, we know that $\phi_t(\nabla\psi_t(w_t)) = \psi_t^*(\nabla\psi_t(w_t))$ and $\phi_t(\nabla\psi_t(w_t) + \widehat{\ell}_t) \leq \psi_t^*(\nabla\psi_t(w_t) + \widehat{\ell}_t)$. Subsequently,

$$\begin{aligned}
-\langle \ell_t, w_t\rangle - \phi_t(\widehat{L}_{t-1}) + \phi_t(\widehat{L}_t) &\leq -\langle \ell_t, w_t\rangle - \psi_t^*(\nabla\psi_t(w_t)) + \psi_t^*(\nabla\psi_t(w_t) + \widehat{\ell}_t) \\
&= D_{\psi_t^*}(\nabla\psi_t(w_t) + \widehat{\ell}_t, \nabla\psi_t(w_t)), \qquad \text{(EC.12)}
\end{aligned}$$

where the last equality holds because $\nabla\psi_t^*(\nabla\psi_t(w_t)) = w_t$, thanks to the $\nabla\psi_t^* = (\nabla\psi_t)^{-1}$ property in Fact 1. Using Eq. (EC.3) and the relationship between $\nabla^2_{\psi_t^*}, \nabla^2_{\psi_t}$ in Fact 1, we have that

$$\begin{aligned}
D_{\psi_t^*}(\nabla\psi_t(w_t) + \widehat{\ell}_t, \nabla\psi_t(w_t)) &\leq \frac{1}{2}\widehat{\ell}_t^\top \nabla^2\psi_t^*(\nabla\psi_t(w_t) + \alpha_t\widehat{\ell}_t)\widehat{\ell}_t \\
&= \frac{1}{2}\widehat{\ell}_t^\top \left[\nabla^2\psi_t(\nabla\psi_t^*(\nabla\psi_t(w_t) + \alpha_t\widehat{\ell}_t))\right]^{-1}\widehat{\ell}_t, \qquad \text{(EC.13)}
\end{aligned}$$

where $\alpha_t \in (0,1)$ is a certain interpolation parameter.

The following lemma upper bounds the discrepancy between $\nabla\psi_t^*(\nabla\psi_t(w_t) + \alpha_t\widehat{\ell}_t)$ and $\nabla\psi_t^*(\nabla\psi_t(w_t)) = w_t$.

LEMMA EC.1. *Let $\widetilde{w} = \nabla\psi_t^*(\nabla\psi_t(w) + \alpha\delta)$ for some $w \in \Delta^{K_1-1}$, $\alpha \in (0,1)$ and $\delta \in \mathbb{R}^{N_1-1}$ satisfying $-\frac{\overline{p}}{w_i} \leq \delta_i \leq \overline{p}$. The potential function is chosen as $\psi_t(w) = \frac{1}{\eta_t}\psi(w)$ where $\psi(w) = \sum_{i=1}^{N_1} -\sqrt{w_i} - \sqrt{1-w_i}$. Suppose $\eta_t \leq \frac{1}{4\overline{p}}(1 - \frac{1}{\sqrt{2}})$. Then it holds for all $i \in [N_1]$ that $2w_i - 1 \leq \widetilde{w}_i \leq 2w_i$.*

We will prove Lemma EC.1 in the next section. For the rest of the proof, note that the condition $\eta_0 = 0.07/\overline{p} \le \frac{1}{4\overline{p}}(1-\frac{1}{\sqrt{2}})$ in Lemma 1 implies the condition on η_t in Lemma EC.1 holds, because $\eta_t = \eta_0/\sqrt{t} \le \eta_0$. Note also that, for any $w \in [0,1]^{N_1}$ and $j \in [N_1]$,

$$\partial^2_{jj}\psi(w) = \frac{1}{4}\left(\frac{1}{w_j^{3/2}} + \frac{1}{(1-w_j)^{3/2}}\right).$$

With Lemma EC.1 and the notation that $\widetilde{w}_t = w_t + \alpha_t \widehat{\ell}_t$, and conditioned on the event that $i_t = i$, it holds that

$$\begin{aligned}
\widehat{\ell}_t^\top \left[\nabla^2\psi_t(\widetilde{w}_t)\right]^{-1}\widehat{\ell}_t &= \sum_{j\neq i}\overline{p}^2[\partial^2_{jj}\psi_t(\widetilde{w}_{tj})]^{-1} + \left(\frac{p(i)d_t - \overline{p}}{w_{ti}} + \overline{p}\right)^2[\partial^2_{ii}\psi_t(\widetilde{w}_{ti})]^{-1} \\
&\le \sum_{j=1}^{N_1}\overline{p}^2[\partial^2_{jj}\psi_t(\widetilde{w}_{tj})]^{-1} + \frac{\overline{p}^2}{w_{ti}^2}[\partial^2_{ii}\psi_t(\widetilde{w}_{ti})]^{-1} && \text{(EC.14)} \\
&= 4\overline{p}^2\eta_t\sum_{j=1}^{N_1}\left(\frac{1}{\widetilde{w}_{tj}^{3/2}} + \frac{1}{(1-\widetilde{w}_{tj})^{3/2}}\right)^{-1} + \frac{4\overline{p}^2\eta_t}{w_{ti}^2}\left(\frac{1}{\widetilde{w}_{ti}^{3/2}} + \frac{1}{(1-\widetilde{w}_{ti})^{3/2}}\right)^{-1} \\
&\le 8\sqrt{2}\overline{p}^2\eta_t\sum_{j=1}^{N_1}\min\left\{w_{tj}^{3/2}, (1-w_{tj})^{3/2}\right\} + \frac{8\sqrt{2}\overline{p}^2\eta_t}{w_{ti}^2}\min\left\{w_{ti}^{3/2}, (1-w_{ti})^{3/2}\right\}. && \text{(EC.15)}
\end{aligned}$$

Here, Eq. (EC.14) holds because $0 \le p(i)d_t \le \overline{p}$ and hence $|\frac{p(i)d_t - \overline{p}}{w_{ti}} + \overline{p}| \le \frac{\overline{p}}{w_{ti}}$. Eq. (EC.15) holds because $(\widetilde{w}_{tj}^{-3/2} + (1-\widetilde{w}_{ti})^{-3/2})^{-1} \le \min\{\widetilde{w}_{tj}^{3/2}, (1-\widetilde{w}_{tj})^{3/2}\} \le \min\{(2w_{tj})^{3/2}, (2-2w_{tj})^{3/2}\}$, thanks to Lemma EC.1. Because $i_t = i$ with probability w_{ti}, the right-hand side of Eq. (EC.15) can be further upper bounded by

$$\begin{aligned}
&8\sqrt{2}\overline{p}^2\eta_t\sum_{j=1}^{N_1}\min\left\{w_{tj}^{3/2}, (1-w_{tj})^{3/2}\right\} + 8\sqrt{2}\overline{p}^2\eta_t\sum_{i=1}^{N_1}\frac{\min\{w_{ti}^{3/2}, (1-w_{ti})^{3/2}\}}{w_{ti}} \\
&\le 16\sqrt{2}\overline{p}^2\eta_t\sum_{i=1}^{N_1}\frac{\min\{w_{ti}^{3/2}, (1-w_{ti})^{3/2}\}}{w_{ti}} \le 16\sqrt{2}\overline{p}^2\eta_t\sum_{i=1}^{N_1}\min\{\sqrt{w_{ti}}, \sqrt{1-w_{ti}}\} \\
&\le 16\sqrt{2}\overline{p}^2\eta_t\left[\sqrt{1-w_{ti^*}} + \sum_{i\neq i^*}\sqrt{w_{ti}}\right] = 16\sqrt{2}\overline{p}^2\eta_t\left[\sqrt{\sum_{i\neq i^*}w_{ti}} + \sum_{i\neq i^*}\sqrt{w_{ti}}\right] \\
&\le 32\sqrt{2}\overline{p}^2\eta_t\left[\sum_{i\neq i^*}\sqrt{w_{ti}}\right] = 32\sqrt{2}\overline{p}^2\eta_0\left[\sum_{i\neq i^*}\sqrt{\frac{w_{ti}}{t}}\right]. && \text{(EC.16)}
\end{aligned}$$

EC.1.3. Proof of Lemma EC.1

Because $\nabla\psi_t^* = (\nabla\psi_t)^{-1}$ thanks to Fact 1, we have that

$$\nabla\psi_t(\widetilde{w}) = \nabla\psi_t(w) + \alpha\delta.$$

Fix an arbitrary $i \in [N_1]$ and let $w_i, \widetilde{w}_i$ be the ith components of w and $\widetilde{w}$, respectively. The above inequality then implies that

$$\partial_i \psi_t(w_i) - \alpha\delta_i \le \partial_i \psi_t(\widetilde{w}_i) \le \partial_i \psi_t(w_i) + \alpha\delta_i. \tag{EC.17}$$

We first prove $\widetilde{w}_i \le 2w_i$. If $w_i \ge 1/2$ then the inequality automatically holds because $\widetilde{w} \in \mathrm{Range}(\nabla\psi_t^*) \subseteq [0,1]^{N_1}$. Hence, we shall assume that $w_i < 1/2$. Assume by way of contradiction that $\widetilde{w}_i > 2w_i$. Because $\partial_i\psi_t(w_i) = \frac{1}{\eta_t}[-\frac{1}{2\sqrt{w_i}} + \frac{1}{2\sqrt{1-w_i}}]$ is strictly monotonically increasing with w_i, we have that

$$\alpha\delta_i = \partial_i\psi_t(\widetilde{w}_i) - \partial_i\psi_t(w_i) > \partial_i\psi_t(2w_i) - \partial_i\psi_t(w_i).$$

Subsequently,

$$\alpha\delta_i > \frac{1}{\eta_t}\left[-\frac{1}{2\sqrt{2w_i}} + \frac{1}{2\sqrt{1-2w_i}} + \frac{1}{2\sqrt{w_i}} - \frac{1}{2\sqrt{1-w_i}}\right] \ge \frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right)\frac{1}{\eta_t w_i} \ge \overline{p},$$

where the last inequality holds because $w_i \le 1$ and $\eta_t \le \frac{1}{2\overline{p}}(1 - \frac{1}{\sqrt{2}})$. This contradicts the condition that $\alpha \in (0,1)$ and $\delta_i \le \overline{p}$.

We next prove $\widetilde{w}_i \ge 2w_i - 1$. If $w_i \le 1/2$ then the inequality automatically holds because $\widetilde{w} \in \mathrm{Range}(\nabla\psi_t^*) \subseteq [0,1]^{N_1}$. Hence we shall assume that $w_i > 1/2$. Assume by way of contradiction that $\widetilde{w}_i < 2w_i - 1$. Again by the strict monotonicity of $\partial_i\psi_t$, we have that

$$\alpha\delta_i = \partial_i\psi_t(\widetilde{w}_i) - \partial_i\psi_t(w_i) < \partial_i\psi_t(2w_i - 1) - \partial_i\psi_t(w_i).$$

Subsequently,

$$\begin{aligned}\alpha\delta_i &< \frac{1}{\eta_t}\left[-\frac{1}{2\sqrt{2w_i-1}} + \frac{1}{2\sqrt{2(1-w_i)}} + \frac{1}{2\sqrt{w_i}} - \frac{1}{2\sqrt{1-w_i}}\right] \le \frac{1}{\eta_t}\left[\frac{1}{2\sqrt{2(1-w_i)}} - \frac{1}{2\sqrt{1-w_i}}\right]\\ &\le -\frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right)\frac{1}{\eta_t\sqrt{1-w_i}} \le -\frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right)\frac{1}{2\eta_t w_i},\end{aligned}$$

where the last inequality holds because $\sqrt{1-w_i} \le 2w_i$ for all $w_i \in (1/2, 1]$. Because $\eta_t \le \frac{1}{4\overline{p}}(1 - \frac{1}{\sqrt{2}})$, we have that

$$\alpha\delta_i < -\overline{p}/w_i,$$

which contradicts the condition that $\alpha \in (0,1)$ and $\delta_i \ge -\overline{p}/w_i$.

EC.1.4. Putting everything together

Combine Eqs. (EC.5,EC.10,EC.16). We have that

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,i^*}-\ell_{t,i_t}\right]\leq\left(32\sqrt{2}\overline{p}^2\eta_0+\frac{2}{\eta_0}\right)\times\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\neq i^*}\sqrt{\frac{w_{ti}}{t}}\right].$$

Plugging in the scaling that $\eta_0=0.07/\overline{p}$ we complete the proof of Lemma 1.

EC.2. Proof of Theorem 1

Recall that $\ell_{t,i}=p(i)f_t(p(i))$ and $i^*=\arg\max_{i\in[N_1]}p(i)f_0(p(i))$ is the revenue maximizer among prices $\{p(i)\}_{i=1}^{N_1}$ for typical customers.

We then have

$$\mathfrak{R}_{T,\varepsilon T}(\text{Alg.3};f_0,x_0)\leq T\big|p^of_0(p^o)-p(i^*)f_0(p(i^*))\big|+\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,i^*}-\ell_{t,i_t}\right].\tag{EC.18}$$

To upper bound the second term on the right-hand side of Eq. (EC.18), note that because $\{p(i)\}_{i=1}^{N_1}$ are N_1 prices evenly partitioning $[\underline{p}^o,\overline{p}]$, there exists $i^\sharp\in[N_1]$ such that $|p(i^\sharp)-p^o|\leq(\overline{p}-\underline{p})/(2N_1)$. Other the other hand, because $r(d)=df_0^{-1}(d)$ is strongly smooth thanks to Assumption (A2), we have that $r(d^o)-r(d(i^\sharp))\leq\frac{M^2}{2}|d^o-d(i^\sharp)|^2$, where $d^o=f_0(p^o)$ and $d(i^\sharp)=f_0(p(i^\sharp))$. Subsequently,

$$\begin{aligned}p^of_0(p^o)-p(i^*)f_0(p(i^*))\leq p^of_0(p^o)-p(i^\sharp)f_0(p(i^\sharp))&=d^of_0^{-1}(d^o)-d(i^\sharp)f_0^{-1}(d(i^\sharp))\\&=r(d^o)-r(d(i^\sharp))\\&\leq\frac{M^2}{2}\big|d^o-d(i^\sharp)\big|^2\leq\frac{M^2}{2}\overline{L}_d^2\big|p^o-p(i^\sharp)\big|^2\qquad\text{(EC.19)}\\&\leq\frac{M^2\overline{L}_d^2}{2}\frac{(\overline{p}-\underline{p})^2}{4N_1^2}=\frac{M^2\overline{L}_d^2(\overline{p}-\underline{p})^2}{8\sqrt{T}}.\qquad\text{(EC.20)}\end{aligned}$$

Here the second inequality in Eq. (EC.19) holds because $|d^o-d(i^\sharp)|=|f_0^{-1}(p^o)-f_0^{-1}(p(i^\sharp))|\leq\overline{L}_d|p^o-p(i^\sharp)|$. Subsequently,

$$T\big|p^of_0(p^o)-p(i^*)f_0(p(i^*))\big|\leq\frac{1}{8}M^2\overline{L}_d^2(\overline{p}-\underline{p})^2\sqrt{T}.\tag{EC.21}$$

We next turn to the third term on the right-hand side of Eq. (EC.18). By Lemma 2, for every $\mathcal{I}\subseteq[N_1]\backslash\{i^*\}$, it holds that

$$\mathbb{E}\left[\sum_{t=1}^T \ell_{t,i^*}-\ell_{t,i_t}\right]\leq 64\overline{p}\sqrt{|\mathcal{I}|T}+\sum_{t=1}^T\sum_{i\notin\mathcal{I},i\neq i^*}\frac{1}{2}\frac{(32\overline{p})^2}{t\Delta\mu_i}+\frac{1}{2}\overline{p}\varepsilon T, \tag{EC.22}$$

where $\Delta\mu_i=p(i^*)f_0(p(i^*))-p(i)f_0(p(i))$. For any $i\in[N_1]$, define $\gamma=|i-i^\sharp|$ where $i^\sharp=\arg\min_{i\in[N_1]}|p(i)-p^o|$. By Lemma 3, if $\gamma\geq\frac{1}{2}+\frac{M\overline{L}_p}{\sqrt{2}\sigma\underline{L}_p}$ then it holds that

$$\Delta\mu_i\geq p(i^\sharp)f_0(p(i^\sharp))-p(i)f_0(p(i))\geq\frac{\sigma^2\underline{L}_p^2(\gamma-1/2)^2\zeta^2}{4},$$

where $\zeta=(\overline{p}-\underline{p}^o)/N_1$. Now let $\mathcal{I}=\{i\in[N_1]:i\neq i^*,|i-i^\sharp|\leq\frac{1}{2}+\frac{M\overline{L}_p}{\sqrt{2}\sigma\underline{L}_p}\}$. Clearly $|\mathcal{I}|\leq 2+\sqrt{2}M\overline{L}_p/(\sigma\underline{L}_p)$. Subsequently, Eq. (EC.22) can be reduced to

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T\ell_{t,i^*}-\ell_{t,i_t}\right]\\
&\leq 64\overline{p}\sqrt{(2+\sqrt{2}M\overline{L}_p/(\sigma\underline{L}_p))T}+\frac{1}{2}\overline{p}\varepsilon T+\sum_{\gamma=1}^{N_1}512\overline{p}^2(\ln T+1)\times\left[\frac{\sigma^2\underline{L}_p^2(\gamma-1/2)^2\zeta^2}{4}\right]^{-1}\\
&\leq 128\sqrt{\frac{M\overline{L}_pT}{\sigma\underline{L}_p}}+\frac{1}{2}\overline{p}\varepsilon T+\frac{2048\overline{p}^2\ln(eT)}{\sigma^2\underline{L}_p^2\zeta^2}\times\sum_{\gamma=1}^{N_1}\frac{1}{(\gamma-1/2)^2}\\
&\leq 128\sqrt{\frac{M\overline{L}_pT}{\sigma\underline{L}_p}}+\frac{1}{2}\overline{p}\varepsilon T+\frac{2048\overline{p}^2\ln(eT)}{\sigma^2\underline{L}_p^2\zeta^2}\times\frac{\pi}{3}.
\end{aligned}\tag{EC.23}$$

Note that $\zeta=(\overline{p}-\underline{p}^o)/N_1$ and $N_1\approx T^{1/4}$. Combining Eqs. (EC.18,EC.21,EC.23) we obtain

$$\begin{aligned}
\mathfrak{R}_{T,\varepsilon T}(\text{Alg.3};f_0,x_0)&\leq\frac{1}{8}M^2\overline{L}_d^2(\overline{p}-\underline{p})^2\sqrt{T}+128\sqrt{\frac{M\overline{L}_pT}{\sigma\underline{L}_p}}+\frac{1}{2}\overline{p}\varepsilon T+\frac{2145\overline{p}^2\sqrt{T}\ln(eT)}{\sigma^2\underline{L}_p^2(\overline{p}-\underline{p}^o)^2}\\
&=\widetilde{O}(\sqrt{T}+\varepsilon T),
\end{aligned}$$

which is to be demonstrated.

EC.3. Proof of Proposition 1.

For notation simplicity we shall omit the π superscript in this proof. For every $t<T$, it holds that $\mathbb{E}[\sum_{t'\leq t}d_{t'}]\leq\min\{x_T,x_0t+\mathbb{E}[\sum_{t'\leq t}\delta_{t'}]+E\}\leq x_0t+\mathbb{E}[\overline{\delta}]+E$. Using Hoeffding's inequality and the union bound, we have with probability $1-T^{-1}$ that

$$\frac{1}{t}\sum_{t'\leq t}d_{t'}\leq x_0+\frac{\mathbb{E}[\overline{\delta}]+E}{t}+\sqrt{\frac{\ln T}{t}}$$

for every $t < T$. Or equivalently,

$$\sum_{t' \le t} d_{t'} \le x_0 t + \mathbb{E}[\overline{\delta}] + E + \sqrt{t \ln T}. \tag{EC.24}$$

Now let $T^+ \le T$ be the random variable of the last episode such that $x(T^+) > 0$. This implies that $\sum_{t' \le T^+} d_{t'} > x_T = x_0 T$. Comparing this with Eq. (EC.24), we have that

$$\mathbb{E}[x_0 T^+ + \overline{\delta} + \varepsilon T + \sqrt{T^+ \log T}] \ge x_0 T.$$

Re-arranging the inequality and noting that $T^+ \le T$, $x_0 \in (0,1]$, we have that

$$\mathbb{E}[T - T^+] \le \frac{1}{x_0}\left(\mathbb{E}[\overline{\delta}] + \varepsilon T + \sqrt{T \log T}\right).$$

This completes the proof, because $\overline{p}(T - T^+)$ is an upper bound on the regret incurred by lost sales over the T time periods.

To prove the high-probability claim, note that the above argument remains valid if one condition on the event that $\overline{\delta} \le \overline{B}$ and taking expectations over the randomness of $\{f_t(p_t) - d_t\}$ only. □

EC.4. Proofs of Technical Results in Sec. 5.1

To prove the Theorem 2, we first introduce some technical lemmas. The following lemma shows that for an epoch τ, $[\underline{d}(\tau), \overline{d}(\tau)]$ covers the expected demand at the price $p(\tau)$ with high probability.

LEMMA EC.2. *Suppose for some epoch τ, the inventory levels kept positive throughout the epoch. Then with probability $1 - O(T^{-2})$, $\underline{d}(\tau) \le f_0(p(\tau)) \le \overline{d}(\tau)$.*

Proof of Lemma EC.2. To prove Lemma EC.2 we first present and prove another lemma, where we try to bound the gap between "the averaged realized demand over T' periods" during which the price p was offered and the actual demand at price p.

LEMMA EC.3. *Let $\mathcal{T}$ be a set of T' selling periods during which the inventory level remained positive. Let $p \in [\underline{p}, \overline{p}]$ be a fixed price, which is offered at each selling period $t \in \mathcal{T}$ with probability $q \in (0,1]$. Let $\widehat{d}$ be the total realized demands over $t \in \mathcal{T}$ during which price p is offered. Suppose also that the total number of corrupted periods during the T' selling periods considered is upper bounded by εT almost surely. Then for any $\delta \in (0,1)$, with probability $1 - \delta$ it holds that*

$$\left|\frac{\widehat{d}}{qT'} - f_0(p)\right| \le \min\{1, \varepsilon T / T'\} + \sqrt{\frac{\log(2/\delta)}{qT'}} + \frac{2\log(2/\delta)}{qT'}.$$

Proof of Lemma EC.3. For each $t \in \mathcal{T}$, let $\nu_t = \mathbf{1}\{p_t = p\}$ be the indicator random variable denoting whether price p is offered at time t. Recall that $\iota_t \in \{0,1\}$ indicates whether t is an outlier period. Define also $z_t := f_t(p)$. We then have

$$\widehat{d} = \sum_{t\in\mathcal{T}} \nu_t\iota_t z_t + \nu_t(1-\iota_t)f_0(p) = \sum_{t\in\mathcal{T}} \nu_t[\iota_t(z_t - f_0(p)) + f_0(p)].$$

Because ν_t is statistically independent from both ι_t and z_t, and that $\Pr[\nu_t = 1] = q$, we have that

$$\mathbb{E}[\widehat{d}] = qZ + qT' f_0(p),$$

where $Z = \sum_{t\in\mathcal{T}} \iota_t z_t$.

To upper bound the deviation of $\widehat{d}$ from $\mathbb{E}[\widehat{d}]$, we need the following result, which is Theorem 1.2A cited from (Victor 1999).

Lemma EC.4. *Let $\{w_i, \mathcal{F}_i\}$ be a martingale difference sequence with $\mathbb{E}[w_j|\mathcal{F}_{j-1}] = 0$, $\mathbb{E}[w_j^2|\mathcal{F}_{j-1}] = \sigma_j^2$, $V_n^2 = \sum_{j=1}^n \sigma_j^2$. Furthermore, assume that $\Pr[|w_j| \le c|\mathcal{F}_{j-1}] = 1$ for some $0 < c < \infty$. Then, for all $\epsilon, y > 0$, it holds that*

$$\Pr\left[\sum_{i=1}^n w_i \ge \epsilon, V_n^2 \le y \text{ for some } n\right] \le \exp\left\{-\frac{\epsilon^2}{2(y + c\epsilon)}\right\}.$$

We now go back to the definition of $\widehat{d}$ and write it as $\widehat{d} - \mathbb{E}[\widehat{d}] = \sum_{t\in\mathcal{T}} w_t - \mathbb{E}[w_t|\mathcal{F}_{t-1}]$, where $w_t = \nu_t[\iota_t(z_t - f_0(p)) + f_0(p)]$ and $\mathcal{F}_{t-1}$ denotes the filtration prior to time t. Clearly $\{w_t - \mathbb{E}[w_t|\mathcal{F}_{t-1}]\}_{t\in\mathcal{T}}$ forms a martingale difference sequence with zero mean. Furthermore, $|w_t| \le 1$ almost surely, and $\mathbb{E}[w_t^2|\mathcal{F}_{t-1}] \le \mathbb{E}[\nu_t|\mathcal{F}_{t-1}] = q$. This means that V_n^2 in Lemma EC.4 satisfies $V_n^2 \le T'q$ almost surely. With $n = T'$, $c = 1$, $y = T'q$ and ϵ appropriately set in Lemma EC.4, we have with probability $1-\delta$ that

$$\left|\sum_{t\in\mathcal{T}} w_t - \mathbb{E}[w_t|\mathcal{F}_{t-1}]\right| \le \log(2/\delta) + \sqrt{\log^2(2/\delta) + y\log(2/\delta)} \le 2\log(2/\delta) + \sqrt{T'q\log(2/\delta)}.$$

Recall that $\sum_{t\in\mathcal{T}} w_t = \widehat{d}$ and $\sum_{t\in\mathcal{T}} \mathbb{E}[w_t|\mathcal{F}_{t-1}] = \mathbb{E}[\widehat{d}] = qZ + qT'f_0(p)$. We then have with probability $1-\delta$ that

$$\left|\widehat{d} - qT'f_0(p)\right| \le qZ + 2\log(2/\delta) + \sqrt{T'q\log(2/\delta)} \le q\max\{T', \varepsilon T\} + 2\log(2/\delta) + \sqrt{T'q\log(2/\delta)},$$

where the last inequality holds since at most εT periods in $\mathcal{T}$ can be corrupted. Dividing both sides of the above inequality by qT' we proved Lemma EC.3. □

Invoking Lemma EC.3 with $T' = T(\tau)$, $q = 1$ and $\delta = 1/T^2$. We then have with probability $1 - O(T^{-2})$ that

$$\left|\frac{d}{T(\tau)} - f_0(p(\tau))\right| \le \min\left\{1, \frac{\varepsilon T}{T(\tau)}\right\} + \sqrt{\frac{\log(2T^2)}{T(\tau)}} + \frac{2\log(2T^2)}{T(\tau)} = C_\varepsilon(\tau),$$

which proves Lemma EC.2. □

Given Lemma EC.2, we are ready to state the next key Lemma for proving Theorem 2. The next lemma provides an upper bound on the length of the searching interval $|I(\tau)| := (b(\tau) - a(\tau))$.

LEMMA EC.5. *With probability $1 - O(T^{-2}\log T)$ the following holds: at the beginning of every epoch τ, $p^c \in I(\tau)$ and the length of $I(\tau)$ is upper bounded by:*

$$|I(\tau)| = (b(\tau) - a(\tau)) \le 2\overline{L}_d C_\varepsilon(\tau - 1).$$

Proof of Lemma EC.5. Throughout this proof we will assume that $f_0(p(\tau)) \in [\underline{d}(\tau), \overline{d}(\tau)]$ for every τ, at the end of each epoch. By Lemma EC.2, this occurs with probability $1 - O(T^{-2}\log T)$.

We first prove $p^c \in I(\tau)$ for any epoch τ. Recall that p^c is the unique price for which $f_0(p^c) = x_0$. Also note that f_0 is strictly monotonically decreasing in p. Hence, in the cases of $x_0 \notin [\underline{d}(\tau), \overline{d}(\tau)]$, it is clear that p^c remains in the shrunk interval $I(\tau+1)$. For the other case of $x_0 \in [\underline{d}(\tau), \overline{d}(\tau)]$, denote $d(\tau) := f_0(p(\tau))$. Since f_0 is strictly decreasing, f_0^{-1} is also a strictly decreasing function. By Assumption (A1), we have that

$$p^c = f_0^{-1}(x_0) \ge f_0^{-1}(\overline{d}(\tau)) \ge f_0^{-1}(d(\tau)) - \overline{L}_d(\overline{d}(\tau) - d(\tau)) \ge f_0^{-1}(d(\tau)) - \overline{L}_d(\overline{d}(\tau) - \underline{d}(\tau));$$
$$p^c = f_0^{-1}(x_0) \le f_0^{-1}(\underline{d}(\tau)) \le f_0^{-1}(d(\tau)) + \overline{L}_d(d(\tau) - \underline{d}(\tau)) \le f_0^{-1}(d(\tau)) + \overline{L}_d(\overline{d}(\tau) - \underline{d}(\tau)).$$

Note that $f_0^{-1}(d(\tau)) = p(\tau)$ and $\overline{d}(\tau) - \underline{d}(\tau) = 2C_\varepsilon(\tau)$. This justifies that $p^c \in I(\tau+1)$ in the case of $x_0 \in [\underline{d}(\tau), \overline{d}(\tau)]$.

We next use induction to prove that $|I(\tau)| \le 2\overline{L}_d C_\varepsilon(\tau-1)$. The base case of $\tau = 1$ clearly holds because $|I(\tau)| \le 1$ and $C_\varepsilon(0) = 1$. Now consider epoch $\tau+1$, assuming the claim holds for epoch τ, or more specifically $|I(\tau)| \le 2\overline{L}_d C_\varepsilon(\tau-1)$. If $x_0 \in [\underline{d}(\tau), \overline{d}(\tau)]$ then clearly $|I(\tau+1)| \le 2\overline{L}_d C_\varepsilon(\tau)$ by definition. For the other case of $x_0 \notin [\underline{d}(\tau), \overline{d}(\tau)]$ we have that $|I(\tau+1)| = |I(\tau)|/2 \le \overline{L}_d C_\varepsilon(\tau-1)$. Note that $C_\varepsilon(\tau) \ge C_\varepsilon(\tau-1)/2$. The lemma is thus proved. □

EC.4.1. Proof of Theorem 2

We first consider the regret incurred by potential running out of inventory (at the end of the T total selling periods). Recall the definition that $\delta_t = \max\{0, f_0(p_t) - x_0\}$. By the Lipschitz continuity of f_0, we have that

$$\sum_{t\le T}\delta_t \le \sum_{t\le T}\left|f_0(p_t)-x_0\right| \le \sum_{t\le T}\overline{L}_p\left|p_t-p^c\right|.$$

Invoking Proposition 1, the regret of Algorithm 2 can be upper bounded by

$$\begin{aligned}\mathfrak{R}_{T,\varepsilon T}(\text{Alg.2}; f_0, x_0) &\le \mathbb{E}\left[\sum_{t=1}^{T} p^c x_0 - p_t f_0(p_t) + \overline{L}_p\big|p_t - p^c\big|\right] + \frac{1}{x_0}(\varepsilon T + \sqrt{T\ln T})\\ &\le (\max\{p^c\overline{L}_p, 1\} + \overline{L}_p)\times\mathbb{E}\left[\sum_{t=1}^{T}\big|p_t - p^c\big|\right] + \frac{1}{x_0}(\varepsilon T + \sqrt{T\ln T}). \qquad \text{(EC.25)}\end{aligned}$$

Here the second inequality holds because, if $p_t > p^c$ then $p^c x_0 - p_t f_0(p_t) \le p^c(f_0(p^c) - f_0(p_t)) \le p^c \times \overline{L}_p|p_t - p^c|$, and if $p_t < p^c$ then $p^c x_0 - p_t f_0(p_t) \le (p^c - p_t)x_0 \le |p^c - p_t|$.

We next upper bound the $\mathbb{E}[\sum_{t=1}^{T}|p_t - p^c|]$ term in Eq. (EC.25). We condition on the success event of $p^c \in I(\tau)$ for all epochs τ, which occurs with probability $1 - O(T^{-2}\log T)$ thanks to Lemma EC.5. Let τ_0 be the last complete epoch with positive inventory levels, satisfying that $2^{\tau_0} \le T$. We then have that

$$\begin{aligned}\sum_{t=1}^{T}\big|p_t - p^c\big| &\le \sum_{\tau=1}^{\tau_0} T(\tau)\times\big|b(\tau) - a(\tau)\big| \le \sum_{\tau=1}^{\tau_0} 2^{\tau}\times 2\overline{L}_d C_{\varepsilon}(\tau-1) \qquad \text{(EC.26)}\\ &\le \sum_{\tau=1}^{\tau_0} 2\overline{L}_d\left(\varepsilon T + 2^{\tau/2}\sqrt{\ln(2T^2)} + 2\ln(2T^2)\right)\\ &\le 2\overline{L}_d\varepsilon T\log_2 T + \frac{2\sqrt{2T\ln(2T^2)}}{\sqrt{2}-1} + 2\overline{L}_d\ln(2T^2)\log_2 T. \qquad \text{(EC.27)}\end{aligned}$$

Here Eq. (EC.26) holds thanks to Lemma EC.5, and the last inequality holds because $2^{\tau_0} \le T$. Combining Eqs. (EC.25) and (EC.27) we have that

$$\begin{aligned}&sR_{T,\varepsilon T}(\text{Alg.2}; f_0, x_0)\\ &\le O(1) + \frac{1}{x_0}(\varepsilon T + \sqrt{T\ln T}) + (\max\{p^c\overline{L}_p, 1\} + \overline{L}_p)\left[2\overline{L}_d\varepsilon T\log_2 T + \frac{2\sqrt{2T\ln(2T^2)}}{\sqrt{2}-1} + 2\overline{L}_d\ln(2T^2)\log_2 T\right]\\ &\le (x_0^{-1} + 4\overline{p}\overline{L}_p\overline{L}_d)\varepsilon T\log_2 T + (x_0^{-1} + 14\overline{p}\overline{L}_p)\sqrt{T\ln(2T^2)} + 6\overline{p}\overline{L}_p\overline{L}_d\ln^2(2T^2) + O(1)\\ &= \widetilde{O}(\varepsilon T + \sqrt{T}),\end{aligned}$$

which completes the proof of Theorem 2.

EC.5. Proofs of Technical Lemmas in Sec. 5.2

EC.5.1. Proof of Lemma 4

Fix a specific epoch τ and thread j such that $\varepsilon_j \geq \varepsilon$. Invoke Lemma EC.3 with $\delta = 1/T^2$, $T' = T(\tau)$ and $q = \wp_j$, we have that with probability $1 - O(T^{-2})$,

$$\left|\frac{\widehat{d}_j(\tau)}{\wp_j T(\tau)} - f_0(p_j(\tau))\right| \leq \min\left\{1, \frac{\varepsilon T}{T(\tau)}\right\} + \sqrt{\frac{\log(2T^2)}{\wp_j T(\tau)}} + \frac{2\log(2T^2)}{\wp_j T(\tau)}$$
$$\leq \min\left\{1, \frac{\varepsilon_j T}{T(\tau)}\right\} + \sqrt{\frac{\log(2T^2)}{\wp_j T(\tau)}} + \frac{2\log(2T^2)}{\wp_j T(\tau)},$$

where the second inequality holds because $\varepsilon \leq \varepsilon_j$. The first property of Lemma 4 is then proved, by the definition of $C_{\varepsilon_j}(\tau)$ in Algorithm 3 and that $\wp_j T(\tau) = T_j(\tau)$.

The second and third properties can be proved in the same vein as the proof of Lemma EC.5, via an induction argument with the union bound over all epochs and threads.

EC.5.2. Proof of Corollary 1

According to Algorithm 3, J only decreases when $I_J(\tau) = \emptyset$. If $\varepsilon_J \leq \varepsilon$, then by Lemma 4 it holds that $p^c \in I_j(\tau)$ for all $j \leq J$ and τ. This means that J will never be further decreased since $I_J(\tau) \neq \emptyset$ throughout.

EC.6. Proofs of technical results in Sec. 6

EC.6.1. Proof of Lemma 5.

Invoking Lemma EC.3 with $T' = T_3$, $q = 1/N_3$, $\varepsilon T = ZT_3$ and $\delta = 1/(N_3 T^2)$, it holds with probability $1 - T^2$ uniformly over all $i \in [N_3]$ that

$$\left|\frac{N_3 \widehat{d}(i)}{T_3} - f_0(p(i))\right| \leq \min\{1, Z\} + \sqrt{\frac{N_3 \log(2N_3T^2)}{T_3}} + \frac{2N_3 \log(2N_3T^2)}{T_3} = H_Z(N_3, T_3). \qquad \text{(EC.28)}$$

We now prove the upper bound on $|\widehat{p}^c - p^c|$. Recall that p^c is the unique solution to $f_0(p^c) = x_0$, or equivalently $p^c = f_0^{-1}(x_0)$. Because of the Lipschitz continuity of f_0 (Assumption 2) and the fact that $p(i)$ and $p(i+1)$ are $(\overline{p} - \underline{p})/N_3$ distance apart, there exists $i^\sharp \in [N_3]$ such that $|f_0(p(i^\sharp)) - x_0| \leq \overline{L}_p(\overline{p} - \underline{p})/N_3$. This implies that, with probability $1 - O(T^{-2})$,

$$\left|f_0(p(\widehat{i^c})) - x_0\right| \leq \left|\widetilde{d}(\widehat{i^c}) - f_0(p(\widehat{i^c}))\right| + \left|\widetilde{d}(\widehat{i^c}) - x_0\right|$$
$$\leq \left|\widetilde{d}(\widehat{i^c}) - f_0(p(\widehat{i^c}))\right| + \left|\widetilde{d}(i^\sharp) - x_0\right| \qquad \text{(EC.29)}$$

$$
\begin{aligned}
&\le \big|\widetilde{d}(\widehat{i}^c) - f_0(p(\widehat{i}^c))\big| + \big|\widetilde{d}(i^\sharp) - f_0(p(i^\sharp))\big| + \big|f_0(p(i^\sharp)) - x_0\big| \\
&\le 2H_Z(N_3,T_3) + \overline{L}_p(\overline{p}-\underline{p})/N_3. \qquad \text{(EC.30)}
\end{aligned}
$$

Here, Eq. (EC.29) holds because $\widehat{i}^c$ minimizes $|\widetilde{d}(i)-x_0|$ by definition, and the last inequality holds with probability $1-O(T^{-2})$ by applying Eq. (EC.28). Consequently,

$$
\begin{aligned}
\big|p(\widehat{i}^c) - p^c\big| &\le \big|f_0^{-1}(x_0 \pm [\overline{L}_p(\overline{p}-\underline{p})/N_3 + 2H_Z(N_3,T_3)]) - f_0^{-1}(x_0)\big| \\
&\le \overline{L}_d \times \left[\frac{\overline{L}_p(\overline{p}-\underline{p})}{N_3} + 2H_Z(N_3,T_3)\right],
\end{aligned}
$$

where the last inequality holds thanks again to Assumption (A1). This proves the upper bound on $|\widehat{p}^c - p^c|$.

Next we prove the upper bound on $|\widehat{p}^o - p^o|$. Let $d^o = f_0(p^o)$ and $i^* \in [N_3]$ be the index such that $|p(i^*) - p^o| \le (\overline{p}-\underline{p})/N_3$. By the Lipschitz continuity of f_0 (see Assumption A2) this means that $|d(p(i^*)) - d^o| \le \overline{L}_p(\overline{p}-\underline{p})/N_3$. Recall also the definition that $r(p) := pf_0(p)$ and $r(d) = df_0^{-1}(d)$. By the strong smoothness of $r(d)$ (see Assumption A3), we have that

$$
\big|r(d(p(i^*))) - r(d^o)\big| \le \frac{M^2}{2}\big|d(p(i^*)) - d^o\big|^2 \le \frac{M^2\overline{L}_p^2(\overline{p}-\underline{p})^2}{2N_3^2}. \qquad \text{(EC.31)}
$$

On the other hand, $\widehat{i}^o$ is selected such that $p(\widehat{i}^o)\widetilde{d}(\widehat{i}^o) \ge p(i^*)\widetilde{d}(i^*)$. Therefore,

$$
\begin{aligned}
r(p(i^*)) - r(p(\widehat{i}^o)) &= p(i^*)f_0(p(i^*)) - p(\widehat{i}^o)f_0(p(\widehat{i}^o)) \\
&\le \overline{p}\big|\widetilde{d}(\widehat{i}^o) - f_0(p(\widehat{i}^o))\big| + \overline{p}\big|\widetilde{d}(i^*) - f_0(p(i^*))\big| + p(i^*)\widetilde{d}(i^*) - p(\widehat{i}^o)\widetilde{d}(\widehat{i}^o) \\
&\le \overline{p}\big|\widetilde{d}(\widehat{i}^o) - f_0(p(\widehat{i}^o))\big| + \overline{p}\big|\widetilde{d}(i^*) - f_0(p(i^*))\big| \le 2\overline{p}H_Z(N_3,T_3). \qquad \text{(EC.32)}
\end{aligned}
$$

Combining Eqs. (EC.31) and (EC.32), we have that

$$
r(d^o) - r(f_0(p(\widehat{i}^o))) \ge \frac{M^2\overline{L}_p^2(\overline{p}-\underline{p})^2}{2N_3^2} + 2\overline{p}H_Z(N_3,T_3), \qquad \text{(EC.33)}
$$

By strong concavity of $r(d)$ (see Assumption A3), Eq. (EC.33) implies that

$$
\big|d^o - f_0(p(\widehat{i}^o))\big| \le \frac{2}{\sigma^2}\sqrt{\frac{M^2\overline{L}_p^2(\overline{p}-\underline{p})^2}{2N_3^2} + 2\overline{p}H_Z(N_3,T_3)} \le \frac{2}{\sigma^2}\left[\frac{M\overline{L}_p(\overline{p}-\underline{p})}{\sqrt{2}N_3} + \sqrt{2\overline{p}H_Z(N_3,T_3)}\right].
$$

Subsequently, using the Lipschitz continuity of f_0^{-1} (see Assumption A2) we have that

$$
\big|\widehat{p}^o - p^o\big| = \big|f_0^{-1}(f_0(p(\widehat{i}^o))) - f_0^{-1}(d^o)\big| \le \overline{L}_d\big|f_0(p(\widehat{i}^o)) - d^o\big|
$$

$$\leq \frac{2\overline{L}_d}{\sigma^2}\left[\frac{M\overline{L}_p(\overline{p}-\underline{p})}{\sqrt{2}N_3}+\sqrt{2\overline{p}H_Z(N_3,T_3)}\right],$$

which is to be demonstrated.

The next lemma is an immediate corollary of Lemma 5, showing that (with high probability) the estimation errors of p^c and p^o are smaller relative to the gap between p^c and p^o, when the parameters in Lemma 5 are tuned appropriately.

COROLLARY EC.1. *Suppose $p^c \neq p^o$, and for some positive constants $\beta' > \gamma' > 0$, that $T_3 = T^{\beta'}$ and $N_3 = T^{\gamma'}$. Suppose also that the Z parameter defined in Lemma 5 satisfies $Z \leq \frac{1}{3}\max\{\frac{|p^o-p^c|}{40\overline{L}_d}, \frac{1}{2\overline{p}}(\frac{\sigma^2|p^o-p^c|}{40\overline{L}_d})^2\}$. Then there exists a* polynomial *function $\varphi(\log T, \overline{L}_d, \overline{L}_p, \overline{p}, 1/|p^o-p^c|)$, whose degrees depend on β', γ', such that if $T \geq \varphi(\log T, \overline{L}_d, \overline{L}_p, \overline{p}, 1/|p^o-p^c|)$, then with probability $1-O(T^{-2})$ it holds that*

$$\left|\widehat{p}^c - p^c\right| + \left|\widehat{p}^o - p^o\right| \leq 0.2\left|p^o - p^c\right|.$$

EC.6.2. Proof of Corollary EC.1.

Following Lemma 5, it suffices to prove, separately, that all of $\frac{\overline{L}_d\overline{L}_p(\overline{p}-\underline{p})}{N_3}$, $2\overline{L}_dH_Z(N_3,T_3)$, $\frac{2\overline{L}_dM\overline{L}_p(\overline{p}-\underline{p})}{\sqrt{2}\sigma^2N_3}$ and $2\sigma^{-2}\overline{L}_d\sqrt{2\overline{p}H_Z(N_3,T_3)}$ terms are upper bounded by $0.05|p^o-p^c|$. to simplify notations, we shall also denote $\Delta p := |p^o - p^c|$ throughout the rest of this proof.

First we consider the $\frac{\overline{L}_d\overline{L}_p(\overline{p}-\underline{p})}{N_3} \leq 0.05\Delta p$ constraint. Re-arranging the terms we have that $N_3 \geq \frac{\overline{L}_d\overline{L}_p(\overline{p}-\underline{p})}{0.05\Delta p}$. Since $N_3 = T^{\gamma'}$, the condition can be reduced to

$$T \geq \left[\frac{\overline{L}_d\overline{L}_p(\overline{p}-\underline{p})}{0.05\Delta p}\right]^{1/\gamma'}. \tag{EC.34}$$

Second we consider the $\frac{2\overline{L}_dM\overline{L}_p(\overline{p}-\underline{p})}{\sqrt{2}\sigma^2N_3} \leq 0.05\Delta p$ constraint. Again, with $N_3 = T^{\gamma'}$, the constraint is satisfied if

$$T \geq \left[\frac{2\overline{L}_dM\overline{L}_p(\overline{p}-\underline{p})}{0.1\sigma^2\Delta p}\right]^{1/\gamma'}. \tag{EC.35}$$

We next consider the constraints $2\overline{L}_dH_Z(N_3,T_3) \leq 0.05\Delta p$ and $2\sigma^{-2}\overline{L}_d\sqrt{2\overline{p}H_Z(N_3,T_3)} \leq 0.05\Delta p$. We first simplify both constraints as conditions involving $H_Z(N_3,T_3)$ only, as

$$H_Z(N_3,T_3) \leq \max\left\{\frac{\Delta_p}{40\overline{L}_d}, \frac{1}{2\overline{p}}\left(\frac{\sigma^2\Delta p}{40\overline{L}_d}\right)^2\right\} =: \overline{H}.$$

By definition of $H_Z(N_3,T_3)$, it suffices to prove that $Z\leq\overline{H}/3$, $\sqrt{\frac{N_3\log(2N_3T^2)}{T_3}}\leq\overline{H}/3$ and $\frac{2N_3\log(2N_3T^2)}{T_3}\leq\overline{H}_3/3$. Note that $Z\leq\overline{H}/3$ holds directly from the condition imposed on Z in Corollary EC.1. On the other hand, noting that $T_3=T^{\beta'}$, $T_3/N_3=T^{\beta'-\gamma'}$ and $\log(2N_3T^2)\leq 3\log(2T)$, the conditions can be reduced to

$$T\geq(3\sqrt{3\log(2T)}/\overline{H})^{2/(\beta'-\gamma')}; \tag{EC.36}$$

$$T\geq(18\log(2T)/\overline{H})^{1/(\beta'-\gamma')}. \tag{EC.37}$$

Corollary EC.1 is subsequently proved, by noting that the right-hand sides of all Eqs. (EC.34) through (EC.37) are low-degree polynomials of $\overline{L}_d,\overline{L}_p,\overline{p},1/|p^o-p^c|,M,1/\sigma$ and $\log T$, with degrees depending only on α',β' and γ'.

EC.6.3. Proof of Theorem 4.

The key step in this proof is to establish the first epoch ζ after which the estimates $\widehat{p}^c,\widehat{p}^o$ are (with high probability) consistent. By consistency, we mean that $|\widehat{p}^c-p^c|+|\widehat{p}^o-p^o|\leq 0.2|p^c-p^o|$, the consequence of Corollary EC.1 established in the previous section. It is easy to verify that, with this inequality, $p^c<p^o$ implies $\widehat{p}^c<\widehat{p}^o$ and vice versa. Furthermore, $\underline{p}^o=(\widehat{p}^c-\widehat{p}^o)/2$ satisfies $p^c\leq\underline{p}^o\leq p^o$. With these conditions, the regret upper bounds proved in Theorems 1 and 3 can be directly applied.

To show $|\widehat{p}^c-p^c|+|\widehat{p}^o-p^o|\leq 0.2|p^c-p^o|$ with probability $1-O(T^{-2})$, we only need to prove the conditions in Corollary EC.1 are satisfied. Consider an arbitrary epoch ζ in Algorithm 5. By definition, $\mathbb{E}[|\mathcal{G}(\zeta)|]=\sqrt{T(\zeta)}$ and $|\mathcal{G}(\zeta)|$ is the sum of $T(\zeta)$ i.i.d. binary random variables. By multiplicative Chernoff bound and the union bound, if $T(\zeta)\geq T_0\geq 16\ln T$ then it holds with probability $1-O(T^{-1})$ uniformly over all ζ that $|\mathcal{G}(\zeta)|\geq\sqrt{T(\zeta)}/2$. Hence, $T_3\geq\sqrt{T(\zeta)}/2$ and $N_3\geq[T(\zeta)]^{1/4}/2$. Additionally, because there are at most εT selling periods being corrupted in epoch τ, and $\mathcal{G}(\zeta)$ are selected uniformly at random, we have with probability $1-O(T^{-2})$ that

$$Z=\frac{1}{|\mathcal{G}(\zeta)|}\sum_{t\in\mathcal{G}(\zeta)}\iota_t=\mathbb{E}[Z]+O(\sqrt{\mathbb{E}[Z^2]\log T}+\log T/T_3)\leq O(\varepsilon T/T(\zeta)).$$

To satisfy the condition $Z\leq\frac{1}{3}\max\{\frac{|p^o-p^c|}{40\overline{L}_d},\frac{1}{2\overline{p}}(\frac{\sigma^2|p^o-p^c|}{40\overline{L}_d})^2\}=:B_Z$, it suffices for ζ to be large enough such that $T(\zeta)\geq\varepsilon T/B_Z$. Because B_Z is a polynomial of problem parameters and is independent of T, it suffices that $T(\zeta)=\Omega(\varepsilon T)$. To satisfy the conditions that $T_3=T^{\beta'}$ and $N_3=T^{\gamma'}$ for some $0<\gamma'<\beta'$, we will simply set $\beta'=0.2$ and $\gamma'=0.1$. Subsequently, $T(\zeta)$ must satisfy $\sqrt{T(\zeta)}\geq T^{0.2}$, $[T(\zeta)]^{1/4}=T^{0.1}$, or more specifically $T(\zeta)\geq T^{0.4}$, to allow $T_3=T^{\beta'}$ and $N_3=T^{\gamma'}$ to hold. Since $T_0=\lceil\sqrt{T}\rceil$, we conclude that all ζ satisfies $T_3=T^{\beta'}$ and $N_3=T^{\gamma'}$.

Now let $\zeta^\sharp$ be the first epoch such that $T(\zeta^\sharp)=\Omega(\varepsilon T)$, such that $Z\leq\frac{1}{3}\max\{\frac{|p^o-p^c|}{40\overline{L}_d},\frac{1}{2\overline{p}}(\frac{\sigma^2|p^o-p^c|}{40\overline{L}_d})^2\}$ holds on and after epoch $\zeta^\sharp$. Per the above discussion, this implies that epochs later than $\zeta^\sharp$ will have consistent $\widehat{p}^c,\widehat{p}^o$ estimates, and therefore the regret of π^c or π^o selected by Algorithm 5 can be upper bounded by $\widetilde{O}(\varepsilon T+\sqrt{T})$. Let also ζ_0 be the last epoch, which must satisfy $|\mathcal{T}(\zeta_0)|\leq T$. The regret of Algorithm 5 can then be upper bounded as

$$\begin{aligned}\mathfrak{R}_{T,\varepsilon T}(\text{Alg.4}; f_0,x_0) &\leq \sum_{\zeta\leq\zeta^\sharp}T(\zeta)+\sum_{\zeta>\zeta^\sharp}\widetilde{O}(\varepsilon T+\sqrt{T}) &\text{(EC.38)}\\ &\leq O(T_0 2^{\zeta^\sharp})+\widetilde{O}(\varepsilon T\zeta_0)+\widetilde{O}(\sqrt{T}\zeta_0)\\ &\leq O(\varepsilon T)+\widetilde{O}(\varepsilon T)+\widetilde{O}(\sqrt{T})=\widetilde{O}(\varepsilon T+\sqrt{T}). &\text{(EC.39)}\end{aligned}$$

Here, in Eq. (EC.38) we apply the upper regret bounds for π^o and π^c, and Eq. (EC.39) holds because $|T(\zeta^\sharp)|=2^{\zeta^\sharp}T_0\asymp\varepsilon T$ and $\zeta_0=O(\log T)$. This proves Theorem 4.

EC.7. Proof of Theorem 5

It suffices to prove that $\mathfrak{R}_{T,\varepsilon T}(\pi; f_0,x_0)\geq C'\max\{\varepsilon T,\sqrt{T}\}$ for some constant $C'>0$. It is a standard result $\mathfrak{R}_{T,0}(\pi; f_0,x_0)\geq\Omega(\sqrt{T})$ when there are no outlier customers (Wang et al. 2014). On the other hand, by setting the first εT customers as outliers who never make any purchases, it is clear that $\mathfrak{R}_{T,\varepsilon T}(\pi; f_0,x_0)\geq\Omega(\varepsilon T)$ because the regret is defined as $Tr^*-\mathbb{E}^\pi[\sum_{t=1}^T r_t]$ where $r^*>0$ is the expected per-period revenue of typical customers. This complete the proof of Theorem 5.