Management Science

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

Dynamic Pricing and Inventory Control with Fixed Ordering Cost and Incomplete Demand Information

Boxiao Chen, David Simchi-Levi, Yining Wang, Yuan Zhou

To cite this article:
Boxiao Chen, David Simchi-Levi, Yining Wang, Yuan Zhou (2022) Dynamic Pricing and Inventory Control with Fixed Ordering Cost and Incomplete Demand Information. Management Science 68(8):5684-5703. https://doi.org/10.1287/mnsc.2021.4171

informs
http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 68, No. 8, August 2022, pp. 5684–5703
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Dynamic Pricing and Inventory Control with Fixed Ordering Cost and Incomplete Demand Information

Boxiao Chen,[a] David Simchi-Levi,[b] Yining Wang,[c] Yuan Zhou[d,e]

[a] College of Business Administration, University of Illinois, Chicago, Illinois 60607; [b] Institute for Data, Systems and Society, Operations Research Center, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139; [c] Warrington College of Business, University of Florida, Gainesville, Florida 32611; [d] Department of Industrial & Enterprise Systems Engineering, University of Illinois, Urbana-Champaign, Illinois 61801; [e] Yanqi Lake Beijing Institute of Mathematical Science and Applications, Beijing 101408, China

Contact: bbchen@uic.edu, https://orcid.org/0000-0002-5967-4822 (BC); dslevi@mit.edu, https://orcid.org/0000-0002-4650-1519 (DS-L); yining.wang@warrington.ufl.edu, https://orcid.org/0000-0001-9410-0392 (YW); yuanz@illinois.edu (YZ)

Received: July 2, 2020
Revised: January 14, 2021; June 27, 2021
Accepted: July 14, 2021
Published Online in Articles in Advance: December 14, 2021

https://doi.org/10.1287/mnsc.2021.4171

Abstract. We consider the periodic review dynamic pricing and inventory control problem with fixed ordering cost. Demand is random and price dependent, and unsatisfied demand is backlogged. With complete demand information, the celebrated $(s, S, \boldsymbol{p})$ policy is proved to be optimal, where s and S are the reorder point and order-up-to level for ordering strategy, and $\boldsymbol{p}$, a function of on-hand inventory level, characterizes the pricing strategy. In this paper, we consider incomplete demand information and develop online learning algorithms whose average profit approaches that of the optimal $(s, S, \boldsymbol{p})$ with a tight $\widetilde{O}(\sqrt{T})$ regret rate. A number of salient features differentiate our work from the existing online learning researches in the operations management (OM) literature. First, computing the optimal $(s, S, \boldsymbol{p})$ policy requires solving a dynamic programming (DP) over *multiple* periods involving unknown quantities, which is different from the majority of learning problems in OM that only require solving single-period optimization questions. It is hence challenging to establish stability results through DP recursions, which we accomplish by proving uniform convergence of the profit-to-go function. The necessity of analyzing action-dependent state transition over multiple periods resembles the *reinforcement learning* question, considerably more difficult than existing bandit learning algorithms. Second, the pricing function $\boldsymbol{p}$ is of infinite dimension, and approaching it is much more challenging than approaching a finite number of parameters as seen in existing researches. The demand-price relationship is estimated based on upper confidence bound, but the confidence interval cannot be explicitly calculated due to the complexity of the DP recursion. Finally, because of the multiperiod nature of $(s, S, \boldsymbol{p})$ policies the actual distribution of the randomness in demand plays an important role in determining the optimal pricing strategy $\boldsymbol{p}$, which is unknown to the learner a priori. In this paper, the demand randomness is approximated by an empirical distribution constructed using dependent samples, and a novel Wasserstein metric-based argument is employed to prove convergence of the empirical distribution.

History: Accepted by J. George Shanthikumar, big data analytics.
Funding: Y. Zhou was supported in part by the Division of Computing and Communication Foundations [Grant CCF-2006526].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2021.4171.

Keywords: dynamic pricing • inventory control • fixed ordering cost • online learning • asymptotic analysis

1. Introduction

The joint optimization of pricing and inventory control has received tremendous attention from both academia and practice. In the literature, there is a vast amount of research devoted to this topic under a variety of problem settings (Petruzzi and Dada 1999, Elmaghraby and Keskinocak 2003, Yano and Gilbert 2003, Chen and Simchi-Levi 2012).

One of the most classic settings is the dynamic pricing and inventory control problem with fixed ordering cost. The firm makes periodic pricing and inventory ordering decisions, and the ordering cost includes both a fixed component and a variable component that is proportional to the ordering quantity. Demand is random and price dependent, and the firm aims to maximize the total profit over all periods. The celebrated $(s, S, \boldsymbol{p})$ policy, first put forth by Thomas (1974), is proved to be optimal with only additive demand randomness for both finite and infinite planning horizons in Chen and Simchi-Levi (2004a, b).

Under the $(s, S, \boldsymbol{p})$ policy, if the inventory level at the beginning of a period is below the reorder point s, an order will be placed to bring the inventory to the order-up-to level S. Otherwise, no order is placed. Price depends on the initial inventory level of the same period, and is characterized by $\boldsymbol{p}$ as a function of inventory level.

Prior studies on this topic assume complete information of demand; that is, the firm knows both the demand-price relationship and the distribution for demand randomness, which is hardly satisfied in practice. In this paper, we explore the problem under incomplete demand information, in which case both the expected demand (as a function of price) and the distribution of additive noises are unknown a priori. We consider the additive demand model $D_0(p) + \beta$, where the demand-price curve $D_0(p) \equiv \mathfrak{D}(\eta(p)|\theta_0)$ is in the parametric from with unknown parameters $\theta_0 \in \Theta$ that can be estimated by a regression oracle $\mathcal{O}$, and β represents the random noise, for which the distribution is unknown in the nonparametric sense. Our proposed framework for the demand-price curve admits both the linear and the generalized linear model.

1.1. Main Results and Contributions

We shall summarize major results and contributions.

1.1.1. Online Learning Algorithm and Convergence Rate. Using the full-information optimal $(s, S, \boldsymbol{p})$ policy as a natural benchmark, we develop online learning algorithms for the joint pricing and inventory control problem with fixed ordering cost. Only taking historical data as input, we use upper confidence bound (UCB) to estimate the unknown parameters of the demand curve $D_0(p)$, and construct an empirical distribution using dependent samples to approximate the distribution of the random noise β. The expected profit of the algorithms converges to the profit of the full-information optimal $(s, S, \boldsymbol{p})$ policy with a cumulative regret on the order of $\widetilde{O}(\sqrt{T})$, which matches the theoretical lower bound (Theorems 1 and 2).

1.1.2. Stability Results of DP Recursion. With complete demand information, computing the optimal $(s, S, \boldsymbol{p})$ policy requires solving a dynamic programming (DP) over multiple periods. The states are inventory levels at the beginning of a period, and $\boldsymbol{p}$ is the optimal action (price) given the state (inventory level). This is in contrast to most problems in pricing, inventory control, and assortment planning that only require solving single-period optimization problems (with incomplete information).

The complexity of multiperiod DP recursion imposes significant challenges in establishing convergence of a learning algorithm, in which both the demand rate curve and the distribution of the demand noise are unknown. One key element to establish convergence is certain stability properties, which guarantees that a small estimation error (in the demand rate curve or the demand noise distribution) translates to a small profit loss. In the case of single-period optimization, such stability properties can be established via proving Lipschitz continuity of the action function and bounded second-order derivatives of the reward function (Broder and Rusmevichientong 2012; Chen et al. 2019b; Chen and Shi 2019a, b). However, in our problem, establishing stability properties of the multiperiod DP recursion requires much more delicate analysis and understanding into the structure of the DP problem.

In this paper, we prove that the empirical profit-to-go function (Lemma 8), surrogate profit-to-go function (Lemma 10), and long-run average profit function (Lemma 9 and Corollary 4) converge to their full-information counterpart uniformly over the state space. These convergence results all build up the stability property through the DP recursions. The necessity of analyzing state transitions of the DP recursion renders this problem a direct application of *reinforcement learning*, which is different than other learning algorithms in operations management (OM) that are based on bandits.

1.1.3. Learning an Infinite Dimensional Object. In an $(s, S, \boldsymbol{p})$ policy, (s, S) are two-dimensional scalars dictating when to order and the order-up-to level, but $\boldsymbol{p}$ is a function of (continuous) inventory levels and is of infinite dimension. The need to learn an infinite dimensional object has not been seen in online learning problems in the OM literature. For problems with multidimensional learning object, a common approach is to learn them in layers. For example, for joint pricing and inventory control without fixed ordering cost, one would like to approach the optimal inventory-price pair, (y, p), which is two dimensional. All works that learn (y, p) under nonparametric noise, including Burnetas and Smith (2000), Chen et al. (2019a, 2021), and Keskin et al. (2022), proposed two-layer learning algorithms, with each layer learning in one dimension. Other examples include Chen and Shi (2019a) learning a two-dimensional tailored base-surged (TBS) policy in dual sourcing inventory systems, and Yuan et al. (2019) learning a two-dimensional (s, S) policy for the pure inventory control problem with fixed ordering cost. For the latter, their two-layer algorithm estimates S in the first layer using stochastic gradient decent (SGD) and $\delta = S - s$ in the second layer by updating an active set after discretization. With an infinite dimensional object, we are definitely not able to adopt this approach.

1.1.4. Convergence Rate of Empirical Distributions with Dependent Samples. We use an empirical distribution to approximate the unknown distribution of the random noise β. There are two technical challenges arising from this approach:

1. The true values of $\{\beta_t\}$ are not directly observable, because $\beta_t = d_t - \mathfrak{D}(\eta(p_t)|\theta_0)$ involves unknown quantity θ_0. Instead, one can only estimate β_t via $\widehat{\beta}_t = d_t - \mathfrak{D}(\eta(p_t)|\widehat{\theta}_t)\rangle$, where $\widehat{\theta}_t$ is the estimate of θ_0 at time t. Hence, the quality of $\widehat{\beta}_t$ samples depends on the quality of the estimates $\widehat{\theta}_t$, which are unfortunately heterogeneous because of the nature of linear or generalized linear contextual bandit structures;

2. The $\{\widehat{\beta}_t\}$ samples are *dependent* because of the dependency of p_t and $\widehat{\theta}_t$ across periods. $\widehat{\beta}_t$ obtained from prior selling periods are statistically *correlated* with the $(s, S, \boldsymbol{p})$ policy implemented on later periods. This means that the impacts of the errors (from both $\widehat{\theta}_t$ and the empirical approximation itself) of β on the expected profit of optimized $(s, S, \boldsymbol{p})$ must be bounded in an *uniform* manner, adding another layer of sophistication to the convergence problem.

In this paper, we use the following strategies to address the two major challenges mentioned previously:

1. To overcome the heterogeneity of the qualities of estimation of $\widehat{\theta}_t$, we only utilize data points $\widehat{\beta}_t$ from certain time periods, when the corresponding $\widehat{\theta}_t$ has a high enough accuracy level. We can also rigorously prove that there are always sufficient number of time periods during which $\widehat{\theta}_t$ is an accurate estimate of θ_0;

2. To address the data dependency and uniform convergence issue, we adopt a novel Wasserstein metric-based argument to prove convergence rate of the empirical distribution (Lemma 6, Corollary 1, and Lemma 7), which is new in the literature. In contrast, the empirical distributions in Chen et al. (2019a, 2021) and Keskin et al. (2022) are all based on independent samples and their convergence can be argued by relatively simple probability tools such as the Hoeffding's inequality.

1.1.5. Applying UCB to Inventory Control Problems.

To our knowledge, this is the first paper that solves a data-driven inventory control problem using UCB, a technique successfully adopted in pricing and assortment planning problems. The necessity of an UCB-based approach, compared with explore-then-exploit or gradient descent strategies in the previous literature (Chen et al. 2019a, 2021; Yuan et al. 2019), is justified by the lack of *concavity* in the $(s, S, \boldsymbol{p})$-learning problem we considered, especially regarding the infinite-dimensional object $\boldsymbol{p}$. Hence, an explore-then-exploit type algorithm delivers suboptimal regret (e.g., $\widetilde{O}(T^{2/3})$) in our problem.

The fundamental challenge of applying UCB to inventory control problems lies in how UCB-type algorithms are analyzed. Virtually all analysis of UCB algorithms relies on the following principle: *Regret is upper bounded by the total lengths of the confidence intervals. For example, in multiarmed bandit the above principle trivially holds by concentration inequalities; in linear contextual bandit, the UCB principle is justified by the analysis of ordinary least squares estimators, which asserts that the estimation error of $\widehat{\theta}$ is characterized by the sample covariance. The total lengths of the confidence intervals are subsequently upper bounded by elliptical potential lemmas*.

In our problem, however, such a regret principle is highly nontrivial to prove. This is because, as multiperiod inventory control policies are implemented, the sum of lengths of the confidence intervals (bounding the estimation errors in θ_0 and the distribution of β) depend on the *trajectories* of the inventory levels, which in turn depend on the estimated pricing strategy $\boldsymbol{p}_b$ that can be *drastically different* from the optimal pricing function $\boldsymbol{p}^*$, despite that they may have similar expected per-period profit. Hence to adopt the UCB framework one must prove that the (expected) regret of the proposed UCB algorithm is upper bounded by the (expected) lengths of confidence intervals *on the trajectories being implemented*; see, for example, Corollary 4 with the right-hand side depending on $\boldsymbol{p}$ instead of the optimal $\boldsymbol{p}^*$. This subtle yet significant technical difficulty resembles the analysis of Q-learning algorithms in reinforcement learning (Jin et al. 2019) rather than single-period bandit learning problems.

1.2. Literature Review

Our work is closely related to the following two streams of literature on pricing and inventory control.

1.2.1. Literature on Joint Pricing and Inventory Control.

Since the seminal paper of Federgruen and Heching (1999), the joint pricing and inventory control problem has been studied extensively in the literature. Readers are referred to survey papers Petruzzi and Dada (1999), Elmaghraby and Keskinocak (2003), Yano and Gilbert (2003), and Chen and Simchi-Levi (2012) for a comprehensive overview of the field.

The joint optimization of pricing and inventory control with fixed ordering cost and stochastic demand is first introduced by Thomas (1974), who proposed the elegant $(s, S, \boldsymbol{p})$ policy and conjectured that it is optimal under fairly general conditions for the backlogging system. According to this policy, inventory is not replenished until it drops below the reorder point s, and then it is ordered up to the order-up-to level S. Price depends on the initial inventory level of the same period, and is characterized by the function $\boldsymbol{p}$, which is a function of the initial inventory level. For the backlogging system, in Chen and Simchi-Levi (2004a), the $(s, S, \boldsymbol{p})$ policy is proved to be optimal for additive demand under finite planning horizon, and in Chen and Simchi-Levi (2004b), it is proved to be optimal for general demand models under both the average and discounted profit criterion and infinite planning horizon. The $(s, S, \boldsymbol{p})$ policy is further studied in Polatoglu and Sahin (2000) for the lost-sales inventory system. Chen et al. (2006) proved that the $(s, S, \boldsymbol{p})$ policy is optimal for additive demand in the lost-sales system, and Song et al. (2009) proved the same result for multiplicative demand. Huh and Janakiraman (2008) reestablished the optimality of $(s, S, \boldsymbol{p})$ for both backlogging and lost-sales systems using an alternative approach. They characterized single-period conditions that are key to establish optimality of (s, S) type policies. The joint pricing and inventory control problem with fixed ordering cost is also studied for the continuous review inventory system in Feng and Chen (2003), Chao and Zhou (2006), and Chen and Simchi-Levi (2006). Results in Chen and Simchi-Levi (2004a, b) are then extended by Yin and Rajaram (2007) to Markovian environments. Hu et al. (2019) considered both convex and concave variable cost and developed heuristics for the problem.

Existing studies on the topic all assume the firm knows the demand-price relationship and distribution for demand noise, which is hardly satisfied in practice. Our work differs from the existing literature by not imposing this assumption, and instead, we develop learning algorithms that make pricing and inventory ordering decisions based on historical data.

1.2.2. Literature on Pricing and Inventory Control with Demand Learning. Existing models on pricing and inventory control with demand learning can be categorized into online and offline settings. For the offline setting, Levi et al. (2007, 2015), Cheung and Simchi-Levi (2019), and Ban (2020) implemented sample average approximation (SAA) to inventory control problems, Ban and Rudin (2018) applied the empirical risk minimization principle to solve the newsvendor problem with feature information, and Ban et al. (2020) tested the effect of model misspecification on the newsvendor problem by exploring different estimation methods. Bu et al. (2020) considered the pricing problem with offline learning under censored data, and Qin et al. (2019) studied the offline joint pricing and inventory control problem without fixed ordering cost, in which available samples are assumed to be independent.

There are considerable technical challenges arising from an online/dynamic learning perspective compared with offline learning of pricing and inventory control strategies. The major difference is that, in an online learning setting, the inventory control levels and the advertised prices (price functions) are adaptively chosen and therefore statistically correlated with the randomly realized demands in prior time periods. Such statistical correlation results in two major technical challenges. First, because of the adaptive nature of the inventory/pricing policies the data samples obtained in online learning may not be well conditioned, making classical statistical analysis much more difficult. Furthermore, the realized *noises* of the demands in prior periods are also correlated with the inventory or pricing strategies implemented in later periods, making regret analysis even more challenging.

For the pure pricing problem with online learning, Harrison et al. (2012) identified the true demand from a set of two candidate functions using Bayesian updating. Keskin and Zeevi (2014) estimated unknown parameters of a linear demand model by least-square regression. Broder and Rusmevichientong (2012) and den Boer and Zwart (2014) used maximum likelihood estimation (MLE) and maximum quasi-likelihood estimation (MQLE), respectively, to learn some general parametric models. Ban and Keskin (2021) studied the data-driven pricing problem under discontinuous demand, and den Boer and Keskin (2020) solved the personalized pricing problem using high-dimensional feature information. First moment estimation was used by Cheung et al. (2017) under the constraint of limited price changes, and linear approximation was used by Besbes and Zeevi (2015) to approximate a nonparametric demand model. For the pricing problem with limited initial inventory, see Besbes and Zeevi (2009) and Wang et al. (2014) for data-driven algorithms for single product and Besbes and Zeevi (2012), Ferreira et al. (2018), Chen et al. (2019b), and Chen and Shi (2019b) for multiple products. Wang et al. (2019) studied both the pure pricing problem and pricing with initial inventory. They considered a general nonparametric model that allows multimodality of the reward function and used UCB plus local polynomial approximation to approach the true optimal price.

For inventory control problems with online demand learning, see applications of SGD in Huh and Rusmevichientong (2009) for learning in the periodic lost-sale system, Shi et al. (2016) for considerations of warehouse capacity, Huh et al. (2009) and Zhang et al. (2020) for systems with positive lead time and lost sales, and Zhang et al. (2018) for learning in the perishable inventory system. Other approaches including Huh et al. (2011) implementing the Kaplan-Meier estimator, Chen and Plambeck (2008) using Bayesian updating, Agrawal and Jia (2019) following the bisection procedure, and Besbes and Muharremoglu (2013) leveraging the structure of the newsvendor quantile solution. Chen and Shi (2019a) learned the best tailored base-surge policy in dual sourcing inventory systems by combining bisection and SGD. Yuan et al. (2019) studied the periodic inventory system with lost sales and fixed ordering cost, and their algorithm relies on a combination of policy elimination and SGD.

Less research has been done on joint pricing and inventory control with online demand learning; most of them do not consider fixed ordering cost except one, and none have learned the best $(s, S, \boldsymbol{p})$ policy. For papers not considering fixed ordering cost, this stream of research originates from Burnetas and Smith (2000), which developed a gradient decent-type algorithm for the repeated single period problem without carry-over inventories. They showed the average profit converges to the true optimal profit under complete demand information but did not provide the convergence rate. Chen et al. (2019a) considered the joint pricing and inventory control model with backlogged demand. They developed a linear approximation based nonparametric learning algorithm that achieves a tight regret rate. Chen et al. (2021) considered the joint optimization problem with lost sales and censored demand, and their spline approximation-based learning algorithm achieves a regret that almost matches the theoretical lower bound. Keskin et al. (2022) considered the joint pricing and inventory control problem for perishable products in a changing environment and estimated demand parameters by MQLE. Chen et al. (2019a) adopted general parametric demand models and considered constraints of limited price changes in the lost-sale system. Yang (2020) is the only paper that considered fixed ordering cost. Instead of $(s, S, \boldsymbol{p})$, they adopted a one-price-(s, S) policy as benchmark, under which the price is fixed throughout the planning horizon and does not change based on inventory levels. This benchmark is suboptimal for the full-information problem without performance justification. They considered discrete price and discrete demand. When demand is unbounded, their regret is $O(\bar{i}^{2.193} + \bar{i}^{1.193} T^{0.839})$, and when demand is bounded, their regret is $\widetilde{O}(\bar{i}^{1.415} T^{0.708})$, where $\bar{i}$ is the number of candidate price levels, therefore, their result cannot be generalized to continuous price decisions.

1.2.3. Literature on Linear Contextual Bandit. In the linear contextual bandit question, at each time period t an action space $\mathcal{A}_t$ is given, with each action $a \in \mathcal{A}_t$ being associated with a context vector $x_{ta} \in \mathbb{R}^{\mathfrak{d}}$. The objective is to select actions $\{a_t\}$ so that the linear rewards $\langle x_{ta}, \theta_0 \rangle$ are maximized, where θ_0 is an unknown regression model.

There have been extensive prior studies on linear contextual bandit in both the operations research and computer science literature (Auer 2002, Dani et al. 2008, Filippi et al. 2010, Rusmevichientong and Tsitsiklis 2010, Abbasi-Yadkori et al. 2011, Li et al. 2019). Some tools from the linear contextual bandit literature, for example the elliptical potential lemma, are useful in our problem as well (Lemma 4 upper bounding the total lengths of confidence bounds). Nevertheless, the linear contextual bandit problem is a stateless single-period optimization problems with contexts supplied by the nature, whereas the inventory control problem is multiperiodic and relies on much more delicate analysis and insights into the structure of the DP solutions, as we have remarked in the previous section.

Our learning framework admits both linear models and generalized linear models, which have been studied in, for example, Keskin and Zeevi (2014), Nambiar et al. (2019), and Ban and Keskin (2021).

2. Problem Formulation

We consider a retailer selling one type of product over T time periods, conveniently labeled as $t = 1, 2, \ldots, T$. At the beginning of every period t, after observing the initial inventory level x_t, the retailer makes a pricing decision p_t, as well as an inventory order-up-to decision $y_t \geq x_t$, such that the inventory level reaches y_t after ordering. Demand is stochastic and price dependent and realizes to be d_t. If d_t is lower than y_t, there are leftover inventories that will be carried over to the next period. If d_t is higher than y_t, unsatisfied demands are backlogged.

Demand and system costs can be characterized using a model $\mathcal{M} = (D_0, \mu_0, k, c, h)$, where the parameters are explained as follows:

1. The term $D_0 : [0,1] \to [\underline{d}_0, \overline{d}_0]$ is the (expected) demand function that maps a price $p \in [0,1]$ to an expected demand $D_0(p) \in [\underline{d}_0, \overline{d}_0]$;

2. The term μ_0 is a probability measure that governs the process of the noise variables; more specifically, given an advertised price p, the realized demand is a random variable of the form $d_t = D_0(p) + \beta$ where $\mathbb{E}_{\mu_0}[\beta] = 0$;

3. The expression $k > 0$ is the fixed ordering cost, which is incurred if $y_t > x_t$;

4. The expression $c > 0$ is the variable ordering cost of ordering one unit of inventory; and

5. The term $h : \mathbb{R} \to \mathbb{R}^+$ is the *holding cost* (when the remaining inventory level is positive) or the *backlogging cost* (when the remaining inventory level is negative).

This is the joint inventory and pricing control model studied in Chen and Simchi-Levi (2004a, b) with fixed ordering cost. Departing from the existing literature, in our paper, we assume that the demand curve D_0 and the noise distribution μ_0 are *unknown* and must be learned or estimated on the fly as inventory and pricing decisions are sequentially made. All the other model parameters k, c, and h are assumed to be known.

An admissible policy is represented by a sequence of order-up-to levels and price levels, $\{y_t, p_t, t = 1, \ldots, T\}$, for which ($y_t$, p_t) only depend on historical information up to period $t - 1$, $(y_r, p_r, d_r : r = 1, \ldots, t-1)$, but not on future information. Given any admissible policy

π, the sequence of events for each period t is described as follows:

1. At the beginning of period t, the retailer observes the initial inventory level x_t.

2. The retailer decides the selling price p_t and the inventory order-up-to level $y_t \geq x_t$. New orderings, if there is any, arrive instantaneously.

3. Demand realizes to be $d_t = D_0(p_t) + \beta_t$ and is satisfied to the maximum extent using on-hand inventory. Unsatisfied demand is backlogged, and any leftover inventory is carried to the next period. The state transition is $x_{t+1} = y_t - d_t$.

4. At the end of period t, the retailer collects profit

$$\mathfrak{r}_t = -\underbrace{k \times \mathbf{1}\{y_t > x_t\}}_{\text{fixed ordeing cost}} - \underbrace{c(y_t - x_t)}_{\text{variable ordering cost}} + \underbrace{p_t(D_0(p_t) + \beta_t)}_{\text{sales revenue}} - \underbrace{h(y_t - D_0(p_t) - \beta_t)}_{\text{holding/backlogging cost}}. \quad (1)$$

In (1), $\mathfrak{r}_t, t = 1, \ldots, T$ are not independent, because $\mathfrak{r}_t$ depends on x_t, which correlates with inventory levels, demand realizations and rewards of previous periods.

We are interested in designing an admissible policy that maximizes the expected long-run average profit. More specifically, for a policy π, we are interested in

$$R_T(\pi) := \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mathfrak{r}_t] \quad \mathfrak{r}_1, \ldots, \mathfrak{r}_T \sim \pi, \mathcal{M} \quad (2)$$

for sufficiently large time horizon T. Equivalently, let $\pi^* \in \Pi$ be the optimal policy within a certain policy family Π *maximizing* the average profit defined in Equation (2). We are interested in designing a (dynamically changing) policy $\widehat{\pi}$ such that the cumulative *regret*

$$T \times [R_T(\pi^*) - R_T(\widehat{\pi})] \quad (3)$$

is minimized with high probability.

2.1. The $(s, S, \boldsymbol{p})$ Policies and Their Average Rewards

Under an $(s, S, \boldsymbol{p})$ policy, the retailer will only order new inventories when $x_t < s$, and after ordering the inventory level reaches $y_t = S$. The function $\boldsymbol{p}$ prescribes the pricing decision that depends on the initial inventory level of the same period. With known demand curve D_0 and noise distribution μ_0, the work of Chen and Simchi-Levi (2004b) proves that, under mild conditions (which we will detail in Section 2.2), there exists a stationary $(s, S, \boldsymbol{p})$ policy that is optimal for infinite horizon under both the average and discounted profit criterion.

We next detail, for a certain $\pi = (s, S, \boldsymbol{p})$ policy, how to calculate its long-run average per-period reward $R_T(\pi)$. For convenience of our presentation and analysis, we opt for a slightly more general notation that applies for an arbitrary noise distribution μ. The optimal $(s, S, \boldsymbol{p})$ policy under the full-information setting can be obtained by simply replacing all occurrences of μ with the true noise distribution μ_0 for the rest of this section.

Define $H_0(x, p; \mu)$ as the *expected* immediate reward of pricing decision p at inventory level x, without ordering new inventories. It is easy to verify that

$$H_0(x, p; \mu) = -\mathbb{E}_\mu[h(x - D_0(p) - \beta)] + pD_0(p) - cD_0(p). \quad (4)$$

For a certain $(s, S, \boldsymbol{p})$ policy, define quantities $I(s, x, \boldsymbol{p}; \mu)$ and $M(s, x, \boldsymbol{p}; \mu)$ as follows:

$$I(s, x, \boldsymbol{p}; \mu) := \begin{cases} H_0(x, \boldsymbol{p}(x); \mu) & \\ + \mathbb{E}_\mu[I(s, x - D_0(\boldsymbol{p}(x)) - \beta, \boldsymbol{p}; \mu)], & x \geq s, \\ 0, & x < s; \end{cases} \quad (5)$$

$$M(s, x, \boldsymbol{p}; \mu) := \begin{cases} 1 + \mathbb{E}_\mu[M(s, x - D_0(\boldsymbol{p}(x)) - \beta, \boldsymbol{p}; \mu)], & x \geq s, \\ 0, & x < s. \end{cases} \quad (6)$$

As indicated by the definitions in (5) and (6), $I(s, x, \boldsymbol{p}; \mu)$ represents the expected total reward collected over a time interval that starts with x initial inventory and ends as soon as the inventory level drops to below s, and $M(s, x, \boldsymbol{p}; \mu)$ represents the expected number of time periods for the inventory to drop from x to below s.

Define $r(s, S, \boldsymbol{p}; \mu)$ as

$$r(s, S, \boldsymbol{p}; \mu) := \frac{-k + I(s, S, \boldsymbol{p}; \mu)}{M(s, S, \boldsymbol{p}; \mu)}. \quad (7)$$

When $I(s, S, \boldsymbol{p}; \mu_0)$ and $M(s, S, \boldsymbol{p}; \mu_0)$ are bounded, lemma 2 from Chen and Simchi-Levi (2004b) shows that $\lim_{T \to \infty} R_T(\pi) = r(s, S, \boldsymbol{p}; \mu_0)$.

Throughout the rest of this paper, we use Π^{ssp} to denote the class of all $(s, S, \boldsymbol{p})$ policies. We also write $(s^*, S^*, \boldsymbol{p}^*)$ for the optimal policy $\pi^* \in \Pi^{\text{ssp}}$ that maximizes $r(s, S, \boldsymbol{p}; \mu_0)$.

2.2. Assumptions and Discussion

We make the following standard assumptions on model parameters D_0, μ_0, and $h(\cdot)$.

(A1) $D_0 : [0, 1] \longrightarrow [\underline{d}_0, \overline{d}_0]$ is continuous, monotonically decreasing, and satisfies $D_0(0) = \overline{d}_0$, $D_0(1) = \underline{d}_0$;

(A2) The noise distribution μ_0 satisfies $D_0(p) + \beta \in [\underline{d}, \overline{d}]$ almost surely for all p, and $\underline{d} > 0$; furthermore, μ_0 is continuous and is equipped with a probability density function (PDF) f_{μ_0} such that $\|f_{\mu_0}\|_\infty < A$ for some constant $A < \infty$; and

(A3) The holding/backlogging cost $h(\cdot)$ is convex and Lipschitz continuous with Lipschitz constant $L' \geq 1$. More specifically, for any $x, x' \in \mathbb{R}$ and $\lambda \in [0, 1]$, it holds that $|h(x)-h(x')| \le L'|x-x'|$ and $h(\lambda x + (1-\lambda)x') \le \lambda h(x) + (1-\lambda)h(x')$.

Assumptions A1 and A2 are very mild and are satisfied by most commonly used demand functions and bounded noise distributions. A3 is a conventional assumption in the inventory control literature (Chen and Simchi-Levi 2004b) and is satisfied, for example, when both the holding cost and backlogging cost are linear. This assumption is needed for the $(s, S, \boldsymbol{p})$ policy to be optimal and for regret analyses later.

We remark that the assumption that $D_0(p) + \beta \ge \underline{d} > 0$ in A2 can be relaxed to the following condition: consider the random process that starts with inventory level $\overline{S}$ (upper bound for order-up-to level, see definition in Assumption C1) and keeps offering the highest price ($p = 1$) without restocking, the inventory diminishes after a finite number of time periods with probability $1 - O(T^{-2})$. Indeed, this is the only property used by our analysis that is relevant to $\underline{d}$. This condition can be further relaxed such that the inventory diminishes after $O(\log T)$ number of time periods with probability $1 - O(T^{-2})$, and our regret convergence rate will be affected up to a logarithmic factor in this case.

In this paper we consider parametric demand rate functions with the following specific assumptions are imposed on D_0.

(B1) There exists a known feature map $\eta : [0,1] \to \mathbb{R}^{\mathfrak{d}}$ (e.g., $\eta(p) = (1, p)$ with $\mathfrak{d} = 2$) and a given parametric model $\{\mathfrak{D}(\cdot|\theta_0) : \theta_0 \in \Theta\}$ with *unknown* parameter θ_0 such that $D_0(p) \equiv \mathfrak{D}(\eta(p)|\theta_0)$;

(B2) There exists a known constant $L < \infty$ such that $\|\eta(p)\|_2 \le L$, $\|\eta(p) - \eta(p')\|_2 \le L|p - p'|$, and $|D_0(p) - D_0(p')| \le L^2|p - p'|$ for all $p, p' \in [0,1]$.

(B3) There exists a regression oracle $\mathcal{O}$ with the following guarantee. Given a set of time periods $\mathcal{H} = \{1, 2, \ldots, \tau\}$, and given p_t as the advertised price at each time period $t \in \mathcal{H}$ and $d_t = D_0(p_t) + \beta_t$ as the realized demand, the regression oracle $\mathcal{O}$ finds an *estimate* $\widehat{\theta}$ of the unknown regression parameter θ_0. Let $\Lambda = I_{\mathfrak{d}\times\mathfrak{d}} + \sum_{t\in\mathcal{H}} \eta(p_t)\eta(p_t)^\top$. There exists $\gamma = \gamma(\mathfrak{d}, \overline{d}, L, T)$ that only logarithmically depends on T, such that with probability $1 - O(T^{-2})$,[1] for every $p \in [0,1]$, it holds that

$$\left| D_0(p) - \mathfrak{D}(\eta(p)|\widehat{\theta}) \right| \le \gamma \sqrt{\eta(p)^\top \Lambda^{-1} \eta(p)}. \tag{8}$$

2.2.1. Linear Models. Our assumption about the regression model naturally admits the linear model where $\Theta \subseteq \mathbb{R}^{\mathfrak{d}}$ and $\mathfrak{D}(\eta(p)|\theta_0) = \eta(p)^\top \theta_0$. More specifically, for the linear model, we let the regression oracle $\mathcal{O}$ use the (regularized) least-squares estimation, that is, let

$$\widehat{\theta}_{\text{Linear}} := \arg\min_{\theta \in \mathbb{R}^{\mathfrak{d}}} \left\{ \frac{1}{2} \sum_{t\in\mathcal{H}} |d_t - \langle \eta(p_t), \theta \rangle|^2 + \frac{1}{2}\|\theta\|_2^2 \right\}. \tag{9}$$

Using standard self-normalized empirical process arguments, we can prove the following lemma (in the online appendix) showing that the constructed regression oracle satisfies Assumption B3.

Lemma 1. *Suppose that* $\gamma \ge \sqrt{2\overline{d}^2 \mathfrak{d} \ln(T^2 L)}$. *If we use* Equation (9) *to estimate the parameter for the linear model, then* Equation (8) *holds with probability* $1 - T^{-2}$.

2.2.2. Generalized Linear Models. Our assumption also admits the generalized linear model where $\Theta \subseteq \mathbb{R}^{\mathfrak{d}}$ and $\mathfrak{D}(\eta(p)|\theta_0) = \upsilon(\eta(p)^\top \theta_0)$ for $\upsilon(\cdot)$ as a given *link function*. If we assume $\upsilon(\cdot)$ is continuously differentiable, Lipschitz with constant k_υ and $c_\upsilon = \inf_{\theta_0 \in \Theta, p \in [0,1]} \upsilon' (\eta(p)^\top \theta_0) > 0$, the following regression oracle satisfies Assumption B3 with $\gamma = O(k_\upsilon c_\upsilon^{-1} \overline{d} \sqrt{\mathfrak{d} \ln^3(TL\mathfrak{d})})$,

$$\widehat{\theta}_{\text{GLM}} := \arg\min_{\theta\in\Theta} \left\| \sum_{t\in\mathcal{H}} (\upsilon(\eta(p_t)^\top \theta) - d_t)\eta(p_t) \right\|_{\Lambda^{-1}}. \tag{10}$$

Formally, we have the following lemma.

Lemma 2. *There exists a universal constant* $C_{\text{GLM}} > 0$ *such that if we suppose that* $\gamma \ge C_{\text{GLM}} \cdot k_\upsilon c_\upsilon^{-1} \overline{d} \sqrt{\mathfrak{d} \ln^3(TL\mathfrak{d})}$ *and use* Equation (10) *to estimate the model parameters, then* Equation (8) *holds with probability* $1 - T^{-2}$.

For the details about the proof of Lemma 2, we refer the readers to Proposition 1 of Filippi et al. (2010).

Finally, the following condition reflects prior knowledge and the potential ranges of the optimal inventory levels s^* and S^*. This assumption can be partially removed in Section 5 with a more complicated dynamic inventory management and pricing algorithm.

(C1) The algorithm has access to inventory level ranges $0 < \underline{s} < \overline{s} \le \underline{S} < \overline{S} < \infty$ such that $s^* \in [\underline{s}, \overline{s}]$ and $S^* \in [\underline{S}, \overline{S}]$; we also assume that $\overline{S} \ge \underline{d}$ because otherwise the problem trivializes.

3. Our Proposed Algorithm

In this section, we first explain our estimation approaches for the unknown demand function and noise distribution, based on which we then introduce our proposed algorithm. In Section 3.1, we develop upper confidence bounds for the demand function D_0, and in Section 3.2, we build an empirical approximation for the unknown distribution μ_0 using dependent and carefully selected samples. Section 3.3 presents a DP procedure explaining how to solve for the corresponding $(s, S, \boldsymbol{p})$ policy given estimators of D_0 and μ_0. Our proposed learning algorithm is presented in Section 3.4.

The algorithm proceeds in epochs, conveniently labeled as $\mathcal{B}_1, \mathcal{B}_2, \ldots$, and at the beginning of every epoch, it renews its estimations for D_0 and μ_0 following the approaches presented in Sections 3.1 and 3.2, respectively. Then based on the updated estimators and following the approach in Section 3.3, it recomputes the $(s, S, \boldsymbol{p})$ policy, which is then implemented for the new epoch.

3.1. Upper Condence Bounds for D_0

Let $b \in \{1, 2, \ldots\}$ be a particular epoch and $\mathcal{H}_{b-1} = \mathcal{B}_1 \cup \ldots \cup \mathcal{B}_{b-1}$ be the union of all epochs prior to b. For time period $t \in \mathcal{H}_{b-1}$, let p_t be the advertised price and $d_t = D_0(p_t) + \beta_t$ be the realized demand. Let the *estimate* $\widehat{\theta}_b$ of the unknown regression parameter θ_0 be computed by the regression oracle $\mathcal{O}$ specified in Assumption B3 given samples from $\mathcal{H}_{b-1}$. Define $\Lambda_b := I_{\mathfrak{d}\times\mathfrak{d}} + \sum_{t\in\mathcal{H}_{b-1}} \eta(p_t)\eta(p_t)^\top$. For every $p \in [0, 1]$, define $\Delta_b(p)$ as

$$\Delta_b(p) := \gamma\sqrt{\eta(p)^\top \Lambda_b^{-1} \eta(p)},$$

where $\gamma > 0$ is the oracle-specific parameter specified in Assumption B3. We then define an upper estimate of D_0, $\overline{D}_b$, as

$$\overline{D}_b(p) := \min\{\bar{d}_0, \underline{d}_0 + L^2(1-p), \mathfrak{D}(\eta(p)|\widehat{\theta}_b) + \Delta_b(p)\}, \tag{11}$$

where $\bar{d}_0, \underline{d}_0$ are maximum and minimum demands defined in Assumption A1, and L is the Lipschitz constant defined in Assumption B2. The Lipschitz continuity of $\eta(p)$ and $\Lambda_b \geq I$ imply the continuity of $\Delta_b(\cdot)$ in p, which further implies the continuity of $\overline{D}_b(\cdot)$ in p.

3.2. Empirical Distributional Approximation of μ_0

One key challenge in the learning-while-doing setting is the fact that all of the important quantities H_0, I, M and r involve expectational evaluated under the noise distribution μ_0, an object which we do not know a priori. In this section, we give details on how empirical distributions are used to approximate μ_0.

At the beginning of epoch b, let $\mathcal{E}_{<b} \subseteq \mathcal{B}_1 \cup \ldots \mathcal{B}_{b-1}$ be a *nonempty subset* of historical selling periods used to approximate the noise distribution μ_0. We define the empirical noise distribution $\widehat{\mu}_b$ as

$$\widehat{\mu}_b := \frac{1}{|\mathcal{E}_{<b}|}\sum_{t\in\mathcal{E}_{<b}} \mathbb{I}[d_t - \mathfrak{D}(\eta(p_t)|\widehat{\theta}_{b(t)})], \tag{12}$$

where $\mathbb{I}[\beta']$ is the point mass at β' and $b(t)$ denotes the epoch to which selling period t belongs. Samples in $\{d_t - \mathfrak{D}(\eta(p_t)|\widehat{\theta}_{b(t)})\}_t$ are *dependent* because both p_t and $\widehat{\theta}_{b(t)}$ are dependent across periods. For technical reasons, $\mathcal{E}_{<b}$ is *not* chosen to include all selling periods prior to epoch b. Instead, we construct $\mathcal{E}_{<b}$ such that all $t \in \mathcal{E}_{<b}$ have small estimation errors of D_0 on the advertised prices.

To further upper bound the deviation of $H_0(x, p; \widehat{\mu}_b)$ from $H_0(x, p; \mu_0)$, we need to demonstrate that the empirical distribution $\widehat{\mu}_b$ is close to the true noise distribution μ_0. Because such deviations must include the estimation errors of D_0 by $\overline{D}_{b(t)}$ themselves, it is crucial to select time periods $t \in \mathcal{B}_1 \cup \ldots \mathcal{B}_{b-1}$ during which the error $\Delta_{b(t)}(p_t)$ is small. To this end, we define $\mathcal{E}_{<b}$ as

$$\mathcal{E}_{<b} := \left\{t \in \mathcal{B}_1 \cup \ldots \cup \mathcal{B}_{b-1} : \Delta_{b(t)}(p_t) \leq \kappa/\sqrt{b}\right\}, \tag{13}$$

where $\kappa > 0$ is a scaling algorithm parameter, set as $\kappa = 2\underline{d}^{-3/2}\overline{d}\overline{S}^{3/2}\gamma\sqrt{\mathfrak{d}\ln(TL^2)}$. Note that κ will only depend logarithmically on T. As we show later in the proof of Lemma 6, our selection of κ leads to $|\mathcal{E}_{<b}| \geq b/2$, meaning that the set is nonempty, and therefore the definition in Equation (13) is proper. The idea of the construction of $\mathcal{E}_{<b}$ in Equation (13) is as follows. Note that $d_t - \mathfrak{D}(\eta(p_t)|\widehat{\theta}_{b(t)}) = \beta_t + (\mathfrak{D}(\eta(p_t)|\theta_0) - \mathfrak{D}(\eta(p_t)|\widehat{\theta}_{b(t)}))$. Although β_t is the desired sample from the noise distribution, $\mathfrak{D}(\eta(p_t)|\theta_0) - \mathfrak{D}(\eta(p_t)|\widehat{\theta}_{b(t)})$ is incurred because of the estimation error of $\widehat{\theta}_{b(t)}$, which may be very large. Also the absolute value of this estimation error is upper bounded by $\Delta_{b(t)}(p_t)$. Constructing $\mathcal{E}_{<b}$ as in Equation (13) allows us to only exploit selling periods during which the estimation errors are sufficiently small. This ensures that the obtained (approximate) noise samples $\{d_t - \mathfrak{D}(\eta(p_t)|\widehat{\theta}_{b(t)})\}_{t\in\mathcal{E}_{<b}}$ are of high quality.

3.3. Dynamic Optimization of $(s, S, \boldsymbol{p})$ Strategies

With the upper confidence bounds $\overline{D}_b$ and the approximate noise distribution $\widehat{\mu}_b$ constructed at the beginning of epoch b, we use the dynamic programming approach detailed in the work of Chen and Simchi-Levi (2004a) to obtain an approximately optimal strategy $(s_b, S_b, \boldsymbol{p}_b)$ to be carried out during epoch b.

First we define an upper bound estimate $\overline{H}_b(x, p; \widehat{\mu}_b)$ on $H_0(x, p; \widehat{\mu}_b)$ as

$$\begin{aligned}\overline{H}_b(x, p; \widehat{\mu}_b) := &-\mathbb{E}_{\widehat{\mu}_b}[h(x - \overline{D}_b(p) - \beta)] \\ &+ p\overline{D}_b(p) - c\overline{D}_b(p) + (c + L')\Delta_b(p),\end{aligned} \tag{14}$$

where the constant L' is defined in Assumption A3.

For any $s \in [\underline{s}, \bar{s}]$, $S \in [\underline{S}, \overline{S}]$, $r \in \mathbb{R}$, demand function $D : [0, 1] \to [\underline{d}, \infty)$, noise distribution μ and their associated $H : \mathbb{R} \times [0, 1] \to \mathbb{R}$, define

$$\begin{aligned}&\phi^{(s,S)}(x; D, r, \mu) \\ &:= \begin{cases} \sup\limits_{p\in[0,1]} H(x, p; \mu) - r + \mathbb{E}_\mu[\phi^{(s,S)}(x - D(p) - \beta; D, r, \mu)], & x \geq s; \\ 0, & x < s. \end{cases}\end{aligned} \tag{15}$$

With $D = \overline{D}_b$ and $H = \overline{H}_b(\cdot, \cdot; \widehat{\mu}_b)$, the functions $\phi^{(s,S)}(x; \overline{D}_b, r, \widehat{\mu}_b)$ can be computed for every $s \in [\underline{s}, \bar{s}]$, $S \in [\underline{S}, \overline{S}]$ and $r \in \mathbb{R}$, because both $H(\cdot, \cdot; \widehat{\mu}_b)$ and the expectation with respect to $\widehat{\mu}_b$ can be evaluated. For every $(s,\ S)$, define

$$\bar{r}_b(s,S) := \inf\{r \in \mathbb{R} : \phi^{(s,S)}(S;\overline{D}_b, r, \widehat{\mu}_b) = k\} \qquad (16)$$

and let the pricing strategy $\boldsymbol{p}$ (associated with inventory levels s, S) be the optimal solution to the $\phi^{(s,S)}(\cdot;\overline{D}_b,\bar{r}_b(s,S),\widehat{\mu}_b)$ dynamic programming; that is, $\boldsymbol{p}(x)$ is defined such that $\phi^{(s,S)}(x;\overline{D}_b,\bar{r}_b(s,S),\widehat{\mu}_b) = \overline{H}_b(x,\boldsymbol{p}(x);\widehat{\mu}_b) - \bar{r}_b(s,S) + \mathbb{E}_{\widehat{\mu}_b}[\phi^{(s,S)}(x - \overline{D}_b(\boldsymbol{p}(x)) - \beta;\overline{D}_b,\ \bar{r}_b(s,S),\widehat{\mu}_b)]$ for all x.

Comparing equations in (15)–(16) with those in (4)–(7), it is easy to observe connections between them; $r(s,S,\boldsymbol{p};\mu)$ in (7) represents the expected per-period profit, which includes both the immediate reward H and the fixed ordering cost k. On the other hand, $\phi^{(s,S)}(S;D,r,\mu)$ in (15) accumulates the immediate reward H over time and subtracts a constant r every period. If the constant r in (15) equals the expected per-period profit (after charging H every period and k every replenishment), intuitively one would expect $\phi^{(s,S)}(S;D,r,\mu)$ to be equal to k. Lemma 3 of Chen and Simchi-Levi (2004b) confirms this connection, which shows that $\phi^{(s,S)}(S;D,r^*(s,S),\mu) = k$, where $r^*(s,S) = \sup_{\boldsymbol{p}} r(s,S,\boldsymbol{p};\mu)$. Therefore, $\bar{r}_b(s,S)$ can be considered as an empirical approximation of $r^*(s,S)$.

We finally remark that in practice, one may discretize the choices of s, S, x, and p in the dynamic programming scheme described previously with granularity T^{-1}. This leads to a computationally efficient algorithm. Conversely, by the Lipschitz property of $\overline{H}_b(\cdot,\cdot;\widehat{\mu}_b)$, it can be shown that the error caused by discretization is at most $O(T^{-1})$, which does not affect the order of the overall regret.

3.4. The Algorithm

Our proposed algorithm is based on an $(s,S,\boldsymbol{p})$ policy with evolving inventory levels (s, S) and pricing strategies $\boldsymbol{p}$. As mentioned earlier, in our algorithm the T time periods are partitioned into *epochs*, labeled as $\mathcal{B}_1$, $\mathcal{B}_2, \ldots$. Restocking only occurs at the first time period of each epoch $\mathcal{B}_b$, $b \in \{1,2,\ldots\}$. Each epoch $\mathcal{B}_b$ is also associated with inventory levels (s_b, S_b) and pricing strategy $\boldsymbol{p}_b$, such that for the first time period $t_b \in \mathcal{B}_b$, the restocked inventory level is $y_{t_b} = S_b$; the epoch $\mathcal{B}_b$ terminates whenever $x_t < s_b$, and for all $t \in \mathcal{B}_b \backslash \{t_b\}$, $y_t = x_t$ and $p_t = \boldsymbol{p}_b(x_t)$. Algorithm 1 gives a pseudo-code description of our proposed algorithm.

Algorithm 1 (Main Algorithm: Dynamic Inventory Control and Pricing with Unknown Demand)

```
1: Input: problem parameters $k$, $c$, $h$, time horizon $T$, the regression-oracle-specific parameter $\gamma$.
2: Output: inventory and pricing decisions $y_t$, $p_t$ for each $t \in [T]$.
3: for epoch $b = 1,2,3,\ldots$ do
4:     Compute the model estimate $\widehat{\theta}_b$ using the regression oracle $\mathcal{O}$ and samples from $\mathcal{H}_{b-1}$;
5:     Construct upper-confidence bounds $\overline{D}_b$ as in Equation (11);
6:     Construct $\widehat{\mu}_b = \frac{1}{|\mathcal{E}_{<b}|}\sum_{t\in\mathcal{E}_{<b}} \mathbb{I}[d_t - \mathfrak{D}(\eta(p_t)|\widehat{\theta}_{b(t)})]$, where $\mathcal{E}_{<b}$ is constructed in Equation (13);
7:     Construct $\overline{H}_b$ as in Equation (14);
8:     For every $s \in [\underline{s},\bar{s}], S \in [\underline{S},\overline{S}]$ compute $\phi^{(s,S)}(S;\overline{D}_b,r,\widehat{\mu}_b)$ as in Equation (15) and find $\bar{r}_b(s,S) = \inf\{r \in \mathbb{R} : \phi^{(s,S)}(S;\overline{D}_b,r,\widehat{\mu}_b) = k\}$;
9:     Select $(s_b,S_b) = \arg\max_{s,S}\bar{r}_b(s,S)$ and let $\boldsymbol{p}_b$ be the optimal pricing decisions associated with dynamic programming $\phi^{(s_b,S_b)}(\cdot;\overline{D}_b,\bar{r}_b(s_b,S_b),\widehat{\mu}_b)$;
10:    For the first time period $t_b$ in epoch $\mathcal{B}_b$ set $y_{t_b} = S_b$ and $p_{t_b} = \boldsymbol{p}_b(S_b)$; for the rest of epoch $\mathcal{B}_b$ set $y_t = x_t$ and $p_t = \boldsymbol{p}_b(x_t)$; epoch $\mathcal{B}_b$ terminates once $x_t < s_b$;
11: end for
```

Updates of the $(s,S,\boldsymbol{p})$ policies being implemented occur at the beginning of each epoch, as detailed from Step 4 to Step 9 in Algorithm 1. More specifically, at the beginning of epoch b when policy update is due, we first collect all realized demand information from previous epochs to construct model estimate $\widehat{\theta}_b$ (of the demand-rate curve) and noise distribution $\widehat{\mu}_b$. With estimates $\widehat{\theta}_b$ and $\widehat{\mu}_b$, dynamic programming (reflected in $\phi^{(s_b,S_b)}(\cdot;\overline{D}_b,\bar{r}_b,\widehat{\mu}_b)$) is computed to obtain an approximately optimal pricing function $\boldsymbol{p}_b$, as well as the inventory levels s_b, S_b.

On the computational side, we remark that Algorithm 1 requires DP computation to be carried out every epoch. Because each epoch lasts for at most $(\overline{S} - \underline{S})/\underline{d} = O(1)$ time periods, Algorithm 1 requires $\Omega(T)$ DP computations which is rather intensive. Furthermore, the boundary between s_b and S_b must be known a priori (reflected in the $\bar{s} \le \underline{S}$ assumption) to make sure that the inventory level is *increased* to S_b at the beginning of each epoch. In Section 5, we present a variant of Algorithm 1 with infrequent policy changes to further reduce the number of DP calculations required to $O(\mathfrak{d}\log T)$, and to remove the $\bar{s} \le \underline{S}$ condition.

3.5. Regret Convergence

Now we provide the convergence rate for the regret of Algorithm 1.

Theorem 1. *Let π be the policy described in Algorithm 1. Suppose that all assumptions listed in Section 2.2 hold. Then with probability $1 - O(T^{-1})$, it holds that*

$$\begin{aligned} &T \times (R_T(\pi^*) - R_T(\pi)) \\ &\quad \le O((\overline{S}/\underline{d})^{7/2}(L' + A + 1)^2(C_1 + 1)(\bar{d} + 1)(c + L' + 1) \\ &\qquad \gamma\sqrt{\mathfrak{d}T\ln(TL)}), \end{aligned}$$

where π^* is the optimal policy in Π^{ssp} that maximizes $r(s,S,\boldsymbol{p};\mu_0)$, $C_1 > 0$ is a constant depending only on $\underline{d}, \bar{d}$, *and only a universal constant factor is hidden in the $O(\cdot)$ notation.*

The regret bound in Theorem 1 grows in the order of $\widetilde{O}(\sqrt{T})$ if we omit the dependencies on other parameters. With $\mathfrak{D}(\eta(p)|\theta_0) = \eta(p)^\top\theta_0$, $k = c = 0$ and $h(\cdot) \equiv 0$, the problem becomes a pure pricing problem with unknown linear demand functions. As long as $\mathfrak{d} > 1$, the works of Broder and Rusmevichientong (2012) and Keskin and Zeevi (2014) prove an $\Omega(\sqrt{T})$ lower bound for any admissible pricing policies. Therefore, the $\widetilde{O}(\sqrt{T})$ regret established in Theorem 1 is optimal.

To prove Theorem 1, there are two parts of critical analyses. First, estimation errors of $\overline{D}_b$, $\widehat{\theta}_b$ and $\overline{H}_b(x, p; \widehat{\mu}_b)$ need to be upper bounded as functions of data size, showing that as more data accumulates, estimations get more accurate at fast enough rates. Second, stability results through DP recursions need to be established, so that small estimation errors can guarantee small differences in the DP solution and profit. Detailed analyses for proving Theorem 1 will be presented in the next section.

Remark 1. If D_0 follows the linear model, we may choose $\gamma = \sqrt{2\overline{d}^2\mathfrak{d}\ln(T^2L)}$ by Lemma 1. Now if we only focus on the dependencies on $\mathfrak{d}$ and T, the regret of Algorithm 1 is upper bounded $O(\mathfrak{d}\sqrt{T}\ln T)$.

4. Regret Analysis

In this section, we present analyses to prove Theorem 1. In Section 4.1, we upper bound the estimation errors of $\overline{D}_b$, $\widehat{\theta}_b$ and $\overline{H}_b(x, p; \widehat{\mu}_b)$. In Section 4.2, we establish stability results of DP recursions. In Section 4.3, we present the final steps to prove Theorem 1 unifying results developed in the previous sections.

Before diving into the technical details, we first provide the following proposition showing that, for epoch based strategies such as the proposed Algorithm 1 or the stationary $(s^*, S^*, \boldsymbol{p}^*)$ strategy, its cumulative reward can be (with high probability) approximated by the $r(s, S, \boldsymbol{p}; \mu_0)$ function defined in Equation (7).

Proposition 1. *Let π be an epoch-based policy with epochs $\mathcal{B}_1 \cup \ldots \cup \mathcal{B}_B = \{1, \ldots, T\}$, such that an $(s_b, S_b, \boldsymbol{p}_b)$ policy is executed in epoch $\mathcal{B}_b$,[2] (with $s_b \in [\underline{s}, \overline{s}]$, $S_b \in [\underline{S}, \overline{S}]$), and that the choices of $(s_b, S_b, \boldsymbol{p}_b)$ only depend on the filtration of prior epochs $\mathcal{H}_{b-1}$. Then with probability $1 - O(T^{-1})$, it holds that*

$$\left|\sum_{t=1}^{T}\mathfrak{r}_t - \sum_{b=1}^{B}\sum_{t\in\mathcal{B}_b} r(s_b, S_b, \boldsymbol{p}_b; \mu_0)\right| \le O(C_3\overline{d}\sqrt{T\log T}), \quad (17)$$

where $C_3 := \lceil(\overline{S} - \underline{s})/\underline{d}\rceil$ and only a universal constant factor is hidden in the $O(\cdot)$ notation.

With Proposition 1, to upper bound the cumulative regret of Algorithm 1 defined in (3), it suffices to upper bound $\sum_b r(s^*, S^*, \boldsymbol{p}^*; \mu_0) - r(s_b, S_b, \boldsymbol{p}_b; \mu_0)$ with high probability (i.e., probability $1 - O(T^{-1})$), where $(s^*, S^*, \boldsymbol{p}^*)$ is the optimal policy in Π^{ssp} maximizing $r(s, S, \boldsymbol{p})$.

4.1. Upper Bounds for Estimation Errors

In this section, we provide upper bounds for the distances between $\overline{D}_b(p)$ and $D_0(p)$, $\widehat{\mu}_b$ and μ_0, $\overline{H}_b(x, p; \widehat{\mu}_b)$ and $H_0(x, p; \mu_0)$. In particular, because only dependent samples are accessible to construct the empirical distribution $\widehat{\mu}_b$, commonly used approaches cannot be applied to prove its convergence. We therefore use a novel Wasserstein metric-based argument, which is new in the literature.

4.1.1. Comparing $\overline{D}_b(p)$ and $D_0(p)$. To compare $\overline{D}_b(p)$ constructed in (11) with the true demand function $D_0(p)$, note that by Assumption B2, we have that $D_0(p) \le d_0 + L^2(1 - p)$ holds for all $p \in [0, 1]$. By Assumption B3 and a union bound over all (at most T) epochs, we have with probability $1 - O(T^{-1})$ that $|\mathfrak{D}(\eta(p)|\widehat{\theta}_b) - D_0(p)| \le \Delta_b(p)$ holds uniformly over all epochs $b \in \{1, 2, \ldots\}$ and all $p \in [0, 1]$. Combining these two facts, we deduce the following lemma.

Lemma 3. *With probability $1 - O(T^{-1})$ uniformly over all epochs $b \in \{1, 2, \ldots\}$, it holds that $\overline{D}_b(p) \ge D_0(p)$ and $\overline{D}_b(p) - D_0(p) \le 2\Delta_b(p)$ for all $p \in [0, 1]$.*

Lemma 3 also implies that $\overline{D}_b(1) = D_0(1) = \underline{d}_0$ with high probability, because $\mathfrak{D}(\eta(1)|\widehat{\theta}_b) + \Delta_b(1)$ must exceed $D_0(1) = \underline{d}_0$ with high probability.

Although the lengths of the confidence intervals $\Delta_b(p)$ may vary wildly for different epochs b and on different advertised prices p, the following lemma, also famously known in the linear contextual bandit literature as the *elliptical potential lemma* (Auer 2002, Filippi et al. 2010, Rusmevichientong and Tsitsiklis 2010, Abbasi-Yadkori et al. 2011), upper bounds the *total sum* of the lengths of the confidence intervals for the first B epochs. For completeness we also include the proof of Lemma 4 in the online appendix.

Lemma 4. *Consider the first B epochs. With probability 1, it holds that $\sum_{b=1}^{B}\sum_{t\in\mathcal{B}_b}\Delta_b(\boldsymbol{p}_b(x_t)) \le \underline{d}^{-1}\overline{d}\,\overline{S}\gamma\sqrt{\mathfrak{d}B\ln(2BL^2)}$.*

4.1.2. Comparing $\widehat{\mu}_b$ and μ_0. In this section, we establish performance guarantees on the approximations $\widehat{\mu}_b$, constructed in (12). Because samples in $\{d_t - \mathfrak{D}(\eta(p_t)|\widehat{\theta}_{b(t)})\}_{t\in\mathcal{E}_{<b}}$ that are used to construct $\widehat{\mu}_b$ are dependent (see discussions in Section 3.2), simply applying concentration inequalities cannot prove that $\widehat{\mu}_b$ is close to μ_0, and this is significantly different than Chen et al. (2019a, 2021), Keskin et al. (2022), and Qin et al. (2019) that are able to access independent

samples. Instead, we adopt a creative argument based on the Wasserstein metric.

We first introduce the Wasserstein metric, based on which we will establish convergence of empirical measures. Let μ, ν be two probability measures supported on a compact set $\mathcal{I} \subseteq \mathbb{R}$. Let $\Xi(\mu, \nu)$ be the set of all distributions supported on $\mathcal{I} \times \mathcal{I}$ such that the marginalization are μ and ν, respectively. For any $\mathrm{p} \in [1, \infty)$, the ℓ_{p}-Wasserstein distance between μ and ν is defined as

$$\mathcal{W}_{\mathrm{p}}(\mu, \nu) := \left[\inf_{\xi \in \Xi(\mu,\nu)} \left\{ \int_{\mathcal{I}\times\mathcal{I}} |x-y|^{\mathrm{p}} \mathrm{d}\xi(x,y) \right\} \right]^{1/\mathrm{p}}. \tag{18}$$

It is a standard result that $\mathcal{W}_{\mathrm{p}}$ is a *metric* and therefore satisfies basic properties for metric distances such as symmetry and subadditivity.

In this paper, we exclusively use the ℓ_1 version of the $\mathcal{W}_{\mathrm{p}}$ metric. It is a famous result (Kantorovich and Rubinstein 1958) that the $\mathcal{W}_1$ distance is equivalent to the discrepancy measured by 1 − Lipschitz functions. More specifically, let $\mathcal{F} = \{f : \mathcal{I} \to \mathbb{R}; |f(x) - f(y)| \le |x-y|,\ \forall x, y \in \mathcal{I}\}$ be the set of all functions that are 1 − Lipschitz continuous. It then holds that

$$\begin{aligned} \mathcal{W}_1(\mu,\nu) &= \sup_{f\in\mathcal{F}} \left| \int_{\mathcal{I}} f(x) \mathrm{d}(\mu(x) - \nu(x)) \right| \\ &= \sup_{f\in\mathcal{F}} |\mathbb{E}_{\mu}[f(x)] - \mathbb{E}_{\nu}[f(x)]|. \end{aligned} \tag{19}$$

Hence, $\mathcal{W}_1$ can be conveniently used to (uniformly) upper bound deviation on functions that are Lipschitz continuous, which will be helpful in our later analysis.

The final part of this section is on how empirical measures converge with respect to the $\mathcal{W}_1$ distance metric. Let μ be a probability measure supported on $\mathcal{I}$ and $x_1, \ldots, x_n \overset{i.i.d}{\sim} \mu$ be n independent samples. Denote $\widehat{\mu}_n = 1/n \sum_{i=1}^{n} \mathbb{I}[x_i]$ as the empirical measure, where $\mathbb{I}[x_i]$ is the point mass on x_i. The following result is cited from theorem 2 of Fournier and Guillin (2015).

Lemma 5. *There exists a constant $c > 0$ depending only on $|\mathcal{I}|$, such that for any $\epsilon \in (0,1)$,*

$$\Pr\left[\mathcal{W}_1(\widehat{\mu}_n, \mu) \ge \epsilon\right] \le e^{-cn\epsilon^2}.$$

Next, we develop upper bounds for the distance between the empirical distribution $\widehat{\mu}_b$ constructed in (12) and the true noise distribution μ_0. The following lemma, built on Lemma 5, upper bounds the $\mathcal{W}_1$-distance between $\widehat{\mu}_b$ and μ_0. It is proved in the online appendix.

Lemma 6. *Suppose κ in Equation (13) satisfies $\kappa \ge 2\underline{d}^{-3/2}\overline{d}\overline{S}^{3/2}\gamma\sqrt{\mathfrak{d}\ln(TL^2)}$. Then there exists a constant $C_1 > 0$ depending only on $\underline{d}, \overline{d}$, such that for any $\delta \in (0,1)$, with probability $1 - 2\delta$ uniformly over all epochs b that*

$$\mathcal{W}_1(\widehat{\mu}_b, \mu_0) \le C_1\sqrt{\frac{\log(T/\delta)}{b}} + \frac{\kappa}{\sqrt{b}}.$$

With Lemma 6, the following corollary immediately follows Equation (19), using the equivalence between the $\mathcal{W}_1$-distance and uniform concentration over Lipschitz continuous functions.

Corollary 1. *Conditioned on the event that $\mathcal{W}_1(\widehat{\mu}_b, \mu_0) \le C_1\sqrt{2\log T/b} + \kappa/\sqrt{b}$ for all b, it holds for any L-Lipschitz continuous function f and any epoch b that*

$$|\mathbb{E}_{\widehat{\mu}_b}[f(\beta)] - \mathbb{E}_{\mu_0}[f(\beta)]| \le L\left[C_1\sqrt{\frac{\log(T/\delta)}{b}} + \frac{\kappa}{\sqrt{b}}\right].$$

Finally, we also show that $\widehat{\mu}_b$ and μ_0 are close in the Kolmogorov-Smirnov statistic. The proof is based on the Dvoretzky-Kiefer-Wolfowitz inequality (Dvoretzky et al. 1956, Massart 1990) and is deferred to the online appendix.

Lemma 7. *Let $F_{\widehat{\mu}_n}$ and F_{μ_0} be the cumulative density functions (CDFs) of $\widehat{\mu}_n$ and μ_0. With probability $1 - 2\delta$, it holds uniformly over all epochs b that*

$$\sup_{\beta} |F_{\widehat{\mu}_n}(\beta) - F_{\mu_0}(\beta)| \le \sqrt{\frac{\log(2T/\delta)}{b}} + \frac{A\kappa}{\sqrt{b}}.$$

4.1.3. Comparing $\overline{H}_b(x, p; \widehat{\mu}_b)$ and $H_0(x, p; \mu_0)$.

Now we compare $\overline{H}_b(x, p; \widehat{\mu}_b)$ constructed in (14) with the true immediate reward $H_0(x, p; \mu_0)$. Note that $\overline{H}_b(x, p; \widehat{\mu}_b)$ and $H_0(x, p; \mu_0)$ are mainly different in the demand function D and the noise distribution μ. Proposition 2 upper bounds the difference resulting from D, whereas μ is held the same. Proposition 3 upper bounds the difference stemming from μ, whereas D is held the same. Both propositions are proved in the online appendix.

Proposition 2. *Conditioned on the event that $\overline{D}_b(p) \ge D_0(p)$ and $\overline{D}_b(p) - D_0(p) \le 3(c + L')\Delta_b(p)$ for all $p \in [0,1]$, it holds with probability 1 that $\overline{H}_b(x, p; \widehat{\mu}_b) \ge H_0(x, p; \widehat{\mu}_b)$ and $\overline{H}_b(x, p; \widehat{\mu}_b) - H_0(x, p; \widehat{\mu}_b) \le 2(c + L')\Delta_b(p)$ for all x and $p \in [0,1]$.*

Proposition 3. *Conditioned on the event that $\mathcal{W}_1(\widehat{\mu}_b, \mu_0) \le C_1\sqrt{\frac{2\log T}{b}} + \frac{\kappa}{\sqrt{b}}$, it holds with probability 1 that $|H_0(x, p; \widehat{\mu}_b) - H_0(x, p; \mu_0)| \le L'\left(C_1\sqrt{\frac{2\log T}{b}} + \frac{\kappa}{\sqrt{b}}\right)$.*

4.2. Stability Results Through DP Recursions

In this section, we analyze DP recursions to prove convergences of the empirical profit-to-go function, surrogate profit-to-go function, and long-run average profit function. The necessity of analyzing action-dependent state transition over multiple periods resembles questions in *reinforcement learning*, which is significantly more challenging than bandit learning algorithms in the existing operations management literature.

Alongside our proof, we prove several technical results that might be of independent interest as well. To facilitate our proof, for an $(s, S, \boldsymbol{p})$ policy, $r \in \mathbb{R}$, demand function D, noise distribution μ, and the associated $H : \mathbb{R} \times [0, 1] \to \mathbb{R}$, we define

$$\psi^{(s,S,\boldsymbol{p})}(x; D, r, \mu) := \begin{cases} H(x, \boldsymbol{p}(x); \mu) - r \\ + \mathbb{E}_{\mu}[\psi^{(s,S,\boldsymbol{p})}(x - D(\boldsymbol{p}(x)) - \beta; D, r, \mu)], & x \ge s; \\ 0, & x < s. \end{cases} \tag{20}$$

Intuitively, the $\psi^{(s,S,\boldsymbol{p})}$ function is defined recursively to capture the expected excessive per-period reward of a *fixed* pricing function $\boldsymbol{p}$ not necessarily being the optimal solution to the DP problem defined in Equation (15). Clearly, if $\boldsymbol{p}$ is the optimal pricing policy implied by $\phi^{(s,S)}(\cdot; D, r, \mu)$, we have that

$$\psi^{(s,S,\boldsymbol{p})}(\cdot; D, r, \mu) \equiv \phi^{(s,S)}(\cdot; D, r, \mu). \tag{21}$$

However, for other (suboptimal) pricing policies, such equalities might not hold.

Recall that $C_3 := \lceil (\overline{S} - \underline{s}) / \underline{d} \rceil < \infty$. Every epoch lasts for at most C_3 time periods almost surely. This quantity will be repeatedly used in the following analysis.

4.2.1. Properties of $\boldsymbol{\phi^{(s,S)}}(\cdot)$ and $\boldsymbol{\bar{r}_b}(\cdot)$.

We first establish several properties of the $\phi^{(s,S)}(\cdot)$ function defined in Equation (15). In the following lemma, the demand function $D(\cdot)$ is general and not necessarily D_0 or $\overline{D}_b$, and the same for μ and H.

Lemma 8. *For any $s \in [\underline{s}, \bar{s}]$, $S \in [\underline{S}, \overline{S}]$, demand function $D : [0, 1] \to [\underline{d}, \bar{d}]$, approximate noise measure μ and immediate reward function H (built on D and μ), the following properties hold (with probability 1):*

1. *For every $r > r'$, if $x < s$ then it holds that $\phi^{(s,S)}(x; D, r, \mu) \le \phi^{(s,S)}(x; D, r', \mu)$, if $x \in [s, S]$, then $\phi^{(s,S)}(x; D, r, \mu) < \phi^{(s,S)}(x; D, r', \mu)$; for every $r > r'$ and $x \le S$, it also holds that $\phi^{(s,S)}(x; D, r', \mu) - \phi^{(s,S)}(x; D, r, \mu) \le C_3(r - r')$;*
2. *There exists $r \in \mathbb{R}$ such that $\phi^{(s,S)}(S; D, r, \mu) = k$;*
3. *If $\sup_{\beta} |F_{\mu}(\beta) - F_{\mu_0}(\beta)| \le \epsilon$ for some $\epsilon \in (0, 1)$, then for every $s \le x \le x' \le S$ and $\boldsymbol{p}$, $|\phi^{(s,S)}(x; D, r, \mu) - \phi^{(s,S)}(x'; D, r, \mu)| \le (C_3 L' + A)|x - x'| + 2\epsilon$;*
4. *Consider $\overline{D}, \overline{H}$ such that $\overline{D}(p) \ge D(p)$ and $\overline{H}(x, p; \mu) \ge H(x, p; \mu)$ for every $p \in [0, 1]$ and $x \le S$, and that $\overline{D}$ are continuous with $\overline{D}(1) = \underline{d}_0$. $\phi^{(s,S)}(x; \overline{D}, r, \mu) \ge \phi^{(s,S)}(x; D, r, \mu)$ for every r; and*
5. *If $\sup_{\beta} |F_{\mu}(\beta) - F_{\mu_0}(\beta)| \le \epsilon$ and $\mathcal{W}_1(\mu, \mu_0) \le \epsilon$ hold for some $\epsilon \in (0, 1)$, then for every $s \le x \le S$ and $\boldsymbol{p}$, it holds that $|\phi^{(s,S)}(x; D, r, \mu) - \phi^{(s,S)}(x; D, r, \mu_0)| \le C_3(C_3 L' + A + 2)\epsilon$.*

The proofs of the properties in Lemma 8 are rather technical and we defer the complete proof to the online appendix. However, we remark, at a higher level, on the intuitive meanings of the properties here to help the readers get a big picture of the properties listed.

Property 1 shows that the $\phi^{(s,S)}(\cdot)$ function is monotonically decreasing in r, holding the other variables fixed, because of the $-r$ term in the definition of $\phi^{(s,S)}(\cdot)$ in Equation (15). Property 2 shows the existence of a unique r at which $\phi^{(s,S)}$ at $x = S$ evaluates to k (where the uniqueness is because of Property 1). Property 4 shows that, by replacing D, H with their uniform upper bounds $\overline{D}$, $\overline{H}$, the $\phi^{(s,S)}(\cdot)$ function does not decrease. All these three properties are quite intuitive and can be proved by simply following the definitions.

Property 3 shows that, if μ is close to μ_0 whose PDF is uniformly upper bounded, then $\phi^{(s,S)}(\cdot)$ is also smooth in the inventory level x. The upper bound on $\sup_{\beta} |F_{\mu}(\beta) - F_{\mu_0}(\beta)|$ is essential for this property to hold, because if μ is not (approximately) smooth then a small change in x could potentially lead to significant changes in probability mass in the recursion formula.

Property 5 shows that, if μ is close to μ_0 then the function $\phi^{(s,S)}(\cdot)$ does not change much by replacing μ_0 with μ. Importantly, the distance between μ and μ_0 are measured in the $\mathcal{W}_1$-distance, which implies uniform concentration of Lipschitz functionals.

With Lemma 8, the following corollary shows that $\bar{r}_b(s, S)$ computed in Algorithm 1 are legitimate upper bounds of per-period rewards of *any* pricing function $\boldsymbol{p}$.

Corollary 2. *Conditioned on the event that $\overline{D}_b(p) \ge D_b(p)$ for all p, it holds that $\bar{r}_b(s, S) \ge \sup_{\boldsymbol{p}} r(s, S, \boldsymbol{p}; \widehat{\mu}_b)$ for all $s \in [\underline{s}, \bar{s}]$, $S \in [\underline{S}, \overline{S}]$.*

4.2.2. Properties of $\boldsymbol{\psi^{(s,S,p)}}(\cdot)$ and $\boldsymbol{\bar{r}_b}(\cdot)$.

Lemmas 9 and 10 are the two key lemmas of this analysis.

Lemma 9. *For any $s \in [\underline{s}, \bar{s}]$, $S \in [\underline{S}, \overline{S}]$, noise distribution μ and pricing strategy $\boldsymbol{p}$, if $\psi^{(s,S,\boldsymbol{p})}(S; D_0, r, \mu) \ge k - \epsilon$ for some $\epsilon > 0$ and $r \in \mathbb{R}$, then $r(s, S, \boldsymbol{p}; \mu) \ge r - \epsilon$, where $r(s, S, \boldsymbol{p}; \mu)$ is the average-reward of policy $(s, S, \boldsymbol{p})$ under noise distribution μ defined in Equation (7).*

At a higher level, Lemma 9 shows that if a pricing function $\boldsymbol{p}$ has a ψ-value not too smaller than k with respect to certain profit hypothesis r and noise distribution μ, then the expected per-period profit of the policy $(s, S, \boldsymbol{p})$ cannot be much smaller than r (measured with respect to μ). The purpose of this lemma is to reduce the question of lower bounding the per-period reward of an $(s, S, \boldsymbol{p})$ policy by establishing the stability of ψ (with D_0 and μ_0 being replaced by their empirical surrogates), a task completed later by Lemma 10.

Corollary 3. *Conditioned on the upper bounds in Lemmas 6 and 7 on $\mathcal{W}_1(\widehat{\mu}_b,\mu_0)$ and $\sup_\beta|F_{\widehat{\mu}_b}(\beta)-F_{\mu_0}(\beta)|$, it holds that*

$$\sup r(s,S,\boldsymbol{p};\mu_0)\ge\sup r(s,S,\boldsymbol{p};\widehat{\mu}_b)$$
$$-C_3(C_3L'+A+2)\left[\sqrt{\frac{\log(2T/\delta)}{b}}+\frac{(A+1)\kappa}{\sqrt{b}}\right].$$

This corollary shows that if two noise distributions are close to each other, the optimal expected per-period profit under one distribution cannot be much worse than under the other one.

For a particular $(s,S,\boldsymbol{p})$ policy, consider the stochastic process $\{z_\tau\}_{\tau\in\mathbb{N}}$ defined as

$$z_\tau=\begin{cases}S, & \tau=0;\\ z_{\tau-1}-D_0(p(z_{\tau-1}))-\beta, & \tau>0;\end{cases}\quad\text{where }\beta\overset{i.i.d.}{\sim}\mu_0. \tag{22}$$

Define also $\tau_0\in\mathbb{N}$ as the stopping time of the first z_τ such that $z_\tau<s$, or more specifically

$$\tau_0:=\min\{\tau\in\mathbb{N}:z_\tau<s\}. \tag{23}$$

Intuitively, $\{z_\tau\}_{\tau\le\tau_0}$ are the (random) inventory levels starting at S, priced by $\boldsymbol{p}$ until it falls below s, with demand noises generated by μ_0.

Lemma 10. *For any s, S, consider policy $\boldsymbol{p}$ that is an optimal solution to $\phi^{(s,S)}(\cdot;\overline{D}_b,r,\widehat{\mu}_b)$. Conditioned on the event that $\overline{D}_b(p)\ge D_0(p)$ and $\overline{D}_b(p)-D_0(p)\le 2\Delta_b(p)$ for all $p\in[0,1]$, and the results in Lemmas 6 and 7, it holds for all $r\in\mathbb{R}$ that*

$$\left|\psi^{(s,S,\boldsymbol{p})}(S;\overline{D}_b,r,\widehat{\mu}_b)-\psi^{(s,S,\boldsymbol{p})}(S;D_0,r,\mu_0)\right|$$
$$\le 3(c+L')\mathbb{E}\left[\sum_{\tau<\tau_0}\Delta_b(\boldsymbol{p}(z_\tau))\right]$$
$$+C_3((4C_3L'+2A)C_1+4)\left(\sqrt{\frac{\log(2T/\delta)}{b}}+\frac{(A+1)\kappa}{\sqrt{b}}\right), \tag{24}$$

where the stochastic process $\{z_\tau\}_{\tau\in\mathbb{N}}$ and stopping time τ_0 are defined in Equations (22) and (23), respectively.

Lemma 10 shows that, the difference in ψ values by replacing μ_0,D_0 with their empirical estimates, can be effectively upper bounded by the aggregated lengths of confidence bands *on the expected inventory level trajectory priced by $\boldsymbol{p}$*. The fact that the right-hand side of Equation (24) is independent from the actual optimal price function $\boldsymbol{p}^*$ is of vital importance in the analysis of UCB-type policies, as we have remarked and emphasized in the introduction.

Corollary 4. *Conditioned on the event described in Lemma 10, for every $s\in[\underline{s},\overline{s}]$, $S\in[\underline{S},\overline{S}]$ and pricing policy $\boldsymbol{p}$ being the optimal pricing policy solved by $\phi^{(s,S)}(\cdot;\overline{D}_b,\overline{r}_b(s,S),\widehat{\mu}_b)$ it holds that*

$$\overline{r}_b(s,S)-r(s,S,\boldsymbol{p};\mu_0)\le 3(c+L')\mathbb{E}\sum_{\tau<\tau_0}\Delta_b(\boldsymbol{p}(z_\tau))$$
$$+C_3((4C_3L'+2A)C_1+4)\left(\sqrt{\frac{\log(2T/\delta)}{b}}+\frac{(A+1)\kappa}{\sqrt{b}}\right).$$

Corollary 4 can be proved by combining results from Lemmas 9 and 10. It considers the pricing policy developed by the DP procedure under parameter estimators and shows that the performance of the pricing policy under the true per-period profit $r(s,S,\boldsymbol{p};\mu_0)$ is not far away from the empirical reward $\overline{r}_b(s,S)$ if the parameter estimators are close to the true estimators.

4.3. Putting It Together

We are now ready to put all results in previous sections together and complete our proof of Theorem 1. The rest of the proof is conditioned on the event that $\overline{D}_b(p)\ge D_0(p)$ and $\overline{D}_b(p)-D_0(p)\le 2\Delta_b(p)$ for all $p\in[0,1]$ and b, and that $\mathcal{W}_1(\widehat{\mu}_b,\mu_0)\le C_1\sqrt{\log(T/\delta)/b}+\kappa/\sqrt{b}$, $\sup_\beta|F_{\widehat{\mu}_b}(\beta)-F_{\mu_0}(\beta)|\le\sqrt{\log(2T/\delta)/b}+A\kappa/\sqrt{b}$. With $\delta=1/T$, the success event occurs with probability at least $1-O(T^{-1})$, because of Lemmas 3, 6, and 7.

Recall the definition that $(s^*,S^*,\boldsymbol{p}^*)$ is the maximizer of $r(\cdot;\mu_0)$. Let also $(\widetilde{s}_b^*,\widetilde{S}_b^*,\widetilde{\boldsymbol{p}}_b^*)$ be the maximizer of $r(\cdot;\widehat{\mu}_b)$. With Proposition 1, we need to upper bound $\sum_b r(s^*,S^*,\boldsymbol{p}^*;\mu_0)-r(s_b,S_b,\boldsymbol{p}_b;\mu_0)$. Using the standard argument of UCB-type analysis, we have that

$$\sum_b r(s^*,S^*,\boldsymbol{p}^*;\mu_0)-r(s_b,S_b,\boldsymbol{p}_b;\mu_0)$$
$$\le\sum_b\Big\{|r(s^*,S^*,\boldsymbol{p}^*;\mu_0)-r(\widetilde{s}_b^*,\widetilde{S}_b^*,\widetilde{\boldsymbol{p}}_b^*;\widehat{\mu}_b)|+r(\widetilde{s}_b^*,\widetilde{S}_b^*,\widetilde{\boldsymbol{p}}_b^*;\widehat{\mu}_b)$$
$$-\overline{r}_b(\widetilde{s}_b^*,\widetilde{S}_b^*)+\overline{r}_b(\widetilde{s}_b^*,\widetilde{S}_b^*)-\overline{r}_b(s_b,S_b)+\overline{r}_b(s_b,S_b)$$
$$-r(s_b,S_b,\boldsymbol{p}_b;\mu_0)\Big\}$$
$$\le\sum_b\left\{C_3(C_3L'+A+2)\left[\sqrt{\frac{\log(2T/\delta)}{b}}+\frac{(A+1)\kappa}{\sqrt{b}}\right]\right.$$
$$+r(\widetilde{s}_b^*,\widetilde{S}_b^*,\widetilde{\boldsymbol{p}}_b^*;\widehat{\mu}_b)-\overline{r}_b(\widetilde{s}_b^*,\widetilde{S}_b^*)+\overline{r}_b(\widetilde{s}_b^*,\widetilde{S}_b^*)$$
$$\left.-\overline{r}_b(s_b,S_b)+\overline{r}_b(s_b,S_b)-r(s_b,S_b,\boldsymbol{p}_b;\mu_0)\right\}, \tag{25}$$

$$\le\sum_b\left\{C_3(C_3L'+A+2)\left[\sqrt{\frac{\log(2T/\delta)}{b}}+\frac{(A+1)\kappa}{\sqrt{b}}\right]\right.$$
$$\left.+\overline{r}_b(\widetilde{s}_b^*,\widetilde{S}_b^*)-\overline{r}_b(s_b,S_b)+\overline{r}_b(s_b,S_b)-r(s_b,S_b,\boldsymbol{p}_b;\mu_0)\right\}, \tag{26}$$

$$\le \sum_b \left\{ C_3(C_3 L' + A + 2)\left[\sqrt{\frac{\log(2T/\delta)}{b}} + \frac{(A+1)\kappa}{\sqrt{b}}\right] + \bar{r}_b(s_b, S_b) - r(s_b, S_b, \boldsymbol{p}_b; \mu_0)\right\}, \tag{27}$$

$$\le \sum_b \left\{ 20C_3^2(L' + A + 1)(C_1 + 1)\left(\sqrt{\frac{\log(2T/\delta)}{b}} + \frac{(A+1)\kappa}{\sqrt{b}}\right) + 3(c + L')\mathbb{E}\sum_{\tau<\tau_0^{(b)}} \Delta_b(\boldsymbol{p}_b(z_\tau^{(b)}))\right\}, \tag{28}$$

where $\{z_\tau^{(b)}\}_{\tau\in\mathbb{N}}$ and $\tau_0^{(b)}$ are the stochastic process and stopping time defined in Equations (22) and (23) with respect to the $(s_b, S_b, \boldsymbol{p}_b)$ policy. In the previous chain of derivations, Equation (25) holds because of Corollary 3 and the upper bounds on $\sup_\beta |F_{\widehat{\mu}_b}(\beta) - F_{\mu_0}(\beta)|$ and $\mathcal{W}_1(\widehat{\mu}_b, \mu_0)$ (by Lemmas 6 and 7). Equation (26) holds because of Corollary 2, which asserts that $\bar{r}_b(\widetilde{s}_b^*, \widetilde{S}_b^*) \ge \sup_{\boldsymbol{p}} r(\widetilde{s}_b^*, \widetilde{S}_b^*, \boldsymbol{p}; \widehat{\mu}_b) = r(\widetilde{s}_b^*, \widetilde{S}_b^*, \widetilde{\boldsymbol{p}}_b^*; \widehat{\mu}_b)$ for all b. Equation (27) holds because $(s_b,\ S_b)$ is the maximizer of $\bar{r}_b(\cdot,\cdot)$ according to Step 9 of Algorithm 1. Equation (28) holds by invoking Corollary 4.

Invoking Lemma 4, setting $\delta = T^{-1}$, and recalling that $\kappa = 2\underline{d}^{-3/2}\overline{d}\overline{S}^{3/2}\gamma\sqrt{\mathfrak{d}\ln(TL^2)}$ and $C_3 = \lceil(\overline{S} - \underline{s})/\underline{d}\rceil \le \overline{S}/\underline{d}$, we have that Equation (28) further implies that

$$\begin{aligned}&\sum_b r(s^*, S^*, \boldsymbol{p}^*; \mu_0) - r(s_b, S_b, \boldsymbol{p}_b; \mu_0)\\ &\le O(1) \times C_3^2(L' + A + 1)(C_1 + 1)\left(\overline{S}/\underline{d}\right)^{3/2}\\ &\quad \left(\overline{d} + 1\right)\gamma\sqrt{\mathfrak{d}T\ln(TL^2)} + (c + L')\underline{d}^{-1}\\ &\quad \overline{d}\overline{S}\gamma\sqrt{\mathfrak{d}T\ln(2TL^2)}\\ &\le O\left(\left(\overline{S}/\underline{d}\right)^{7/2}(L' + A + 1)^2(C_1 + 1)\left(\overline{d} + 1\right)\right.\\ &\quad \left.(c + L' + 1)\gamma\sqrt{\mathfrak{d}T\ln(TL)}\right),\end{aligned} \tag{29}$$

where only universal constant factors are hidden in the $O(\cdot)$ notation. This completes the proof of Theorem 1.

5. Improved Algorithm: Infrequent DP Updates

In this section, we present an improved algorithm (with provable $\widetilde{O}(\sqrt{T})$-regret guarantees) that achieves the following two objectives:

1. In Algorithm 1, a dynamic programming needs to be carried out after each epoch b to obtain a new policy $(s_b, S_b, \boldsymbol{p}_b)$. Because each epoch lasts at most $\overline{S}/\underline{d} = O(1)$ selling periods, the algorithm requires $\Omega(T)$ DP calculations which can be computationally expensive. In the improved algorithm, only $O(\mathfrak{d}\log T)$ DP calculations are needed to achieve virtually the same regret, which is much more computationally efficient.

2. In Assumption C1, it is assumed that *prior knowledge* is available on *disjoint* ranges of s^*, S^*, or more specifically $s^* \le \bar{s} \le \underline{S} \le S^*$. Although a lower bound on s^* (and similarly an upper bound on S^*) is most of the time available, $\bar{s} \le \underline{S}$ could be a strong condition. Without this condition, Algorithm 1 may not attain the desired regret bound because S_{b+1} might be potentially *smaller* than s_{b+1}, making selling inventories at the beginning of epoch b + 1 not feasible. This limitation can be fully removed by our improved algorithm with infrequent DP updates.

We will impose the following assumption replacing Assumption C1 earlier:

(C1') The algorithm has access to inventory level ranges $0 < \underline{s} < \overline{S} < \infty$ such that $s^*, S^* \in [\underline{s}, \overline{S}]$; we also assume that $\overline{S} \ge \underline{d}$ because otherwise the problem trivializes.

Algorithm 2 (Dynamic Inventory Control and Pricing with Infrequent DP Solutions)

```
1: Input: problem parameters $k$, $c$, $h$, time horizon $T$, the regression-oracle-specific parameter $\gamma$.
2: Output: inventory and pricing decisions $y_t$, $p_t$ for each $t \in [T]$.
3: Initialize: $\widehat{\theta}_0 = 0^{\mathfrak{d}}$, $\Lambda_1 = I_{\mathfrak{d}\times\mathfrak{d}}$ and $\zeta_1 = 1$;
4: for epoch $b = 1, 2, 3, \ldots$ do
5:     if $\det(\Lambda_b) \ge 2\zeta_b$ or $b = 2^\iota$ for some $\iota \in \mathbb{N}$ then
6:         Update $\zeta_{b+1} = \det(\Lambda_b)$ and compute the model estimate $\widehat{\theta}_b$ using the regression oracle $\mathcal{O}$ and samples from $\mathcal{H}_{b-1}$;
7:         Construct upper-confidence bounds $\overline{D}_b$ as in Equations (11) and (14);
8:         Construct $\widehat{\mu}_b = \frac{1}{|\mathcal{E}_{<b}|}\sum_{t\in\mathcal{E}_{<b}}\mathbb{I}[d_t - \mathfrak{D}(\eta(p_t)|\widehat{\theta}_{b(t)})]$, where $\mathcal{E}_{<b}$ is constructed in Equation (13);
9:         For every $s, S \in [\underline{s}, \overline{S}]$ compute $\phi^{(s,S)}(S; \overline{D}_b, r, \widehat{\mu}_b)$ as in Equation (15) and find $\bar{r}_b(s, S) = \inf\{r \in \mathbb{R} : \phi^{(s,S)}(S; \overline{D}_b, r, \widehat{\mu}_b) = k\}$;
10:        Select $(s_b, S_b) = \arg\max_{s,S}\bar{r}_b(s, S)$ and let $\boldsymbol{p}_b$ be the optimal pricing decisions associated with dynamic programming $\phi^{(s_b,S_b)}(\cdot; \overline{D}_b, \bar{r}_b(s_b, S_b), \widehat{\mu}_b)$;
11:    else
12:        Set $\widehat{\theta}_b = \widehat{\theta}_{b-1}$, $\zeta_{b+1} = \zeta_b$, $\overline{D}_b = \overline{D}_{b-1}$, $\widehat{\mu}_b = \widehat{\mu}_{b-1}$, $s_b = s_{b-1}$, $S_b = S_{b-1}$ and $\boldsymbol{p}_b = \boldsymbol{p}_{b-1}$;
13:    end if
14:    If the current inventory level exceeds $S_b$, set $p_t = 0$ until inventory level falls below $S_b$; *
15:    For the first time period $t_b$ in epoch $\mathcal{B}_b$ set $y_{t_b} = S_b$ and $p_{t_b} = \boldsymbol{p}_b(S_b)$; for the rest of epoch $\mathcal{B}_b$ set $y_t = x_t$ and $p_t = \boldsymbol{p}_b(x_t)$; epoch $\mathcal{B}_b$ terminates once $x_t < s_b$;
16:    Update $\Lambda_{b+1} = \Lambda_b + \sum_{t\in\mathcal{B}_b}\eta(p_t)\eta(p_t)^\top$;
17: end for
```

*This step may only happen when the policy changes. It does not belong to any epoch; and because it happens very infrequently, its incurred regret can be bounded separately.

Theorem 2. *Let π' be the policy described in Algorithm 2. Suppose that all assumptions listed in Section 2.2 hold, except for Assumption C1, which is replaced by Assumption C1′. Then with probability $1 - O(T^{-1})$ it holds that*

$$\begin{aligned} &T \times (R_T(\pi^*) - R_T(\pi')) \\ &\quad \le O((\overline{S}/\underline{d})^{7/2}(L' + A + 1)^2(C_1 + 1)(\overline{d} + 1)(c + L' + 1) \\ &\qquad \gamma\sqrt{\mathfrak{d} T \log(TL)}), \end{aligned}$$

where π^* is the optimal policy in Π^{ssp} that maximizes $r(s, S, \boldsymbol{p}; \mu_0)$. Furthermore, with probability 1, Steps 6–10 are executed for at most $O(\mathfrak{d} \log(TL))$ times.

In the previous $O(\cdot)$ notations, we only hide universal constant factors.

The regret bound in Theorem 2 grows in the order of $\widetilde{O}(\sqrt{T})$ if dependencies on other parameters are omitted, which matches the theoretical lower bound (see discussion after Theorem 1). Remark 1 also applies to the regret bound in Theorem 2.

Comparing Algorithm 2 with Algorithm 1, which updates the $(s, S, \boldsymbol{p})$ policy to be implemented at the beginning of *every* epoch, the new algorithm only updates/changes policies infrequently. More specifically, a new $(s, S, \boldsymbol{p})$ policy is computed only if 2^ι, $\iota \in \{1, 2, \ldots,\}$ epochs are met, or the determinant of the sample covariance Λ_b doubles. This greatly reduces the number of DP calculations from $O(T)$ to $O(\mathfrak{d}\log T)$. It also removes the $\overline{s} \le \underline{S}$ condition in Assumption C3, because Step 14 in Algorithm 2, which suffers constant regret per period, is carried out for at most $O(\mathfrak{d}\log T)$ periods because of the infrequent policy changes.

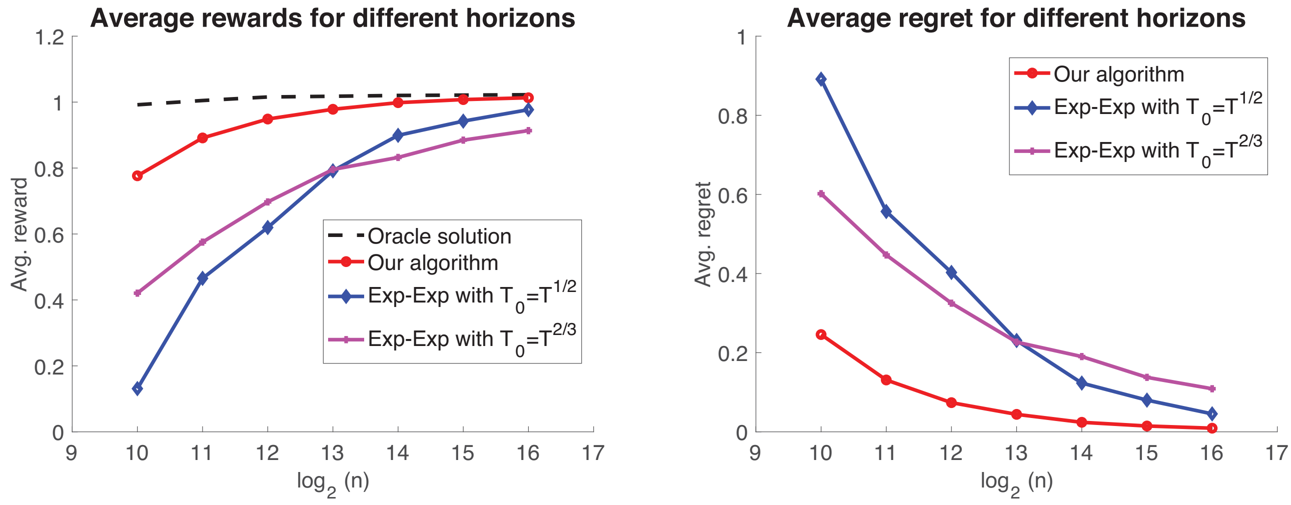
The proof of Theorem 2 is given in the online appendix.

6. Numerical Results

To corroborate our theoretical analysis, we report some numerical results comparing the performances of our proposed algorithms with some baseline methods on synthetically generated data. The first model we used in our numerical studies is two dimensional with feature map $\eta(p) = (1, p)$, demand model $\theta_0 = (18, -15)$, fixed ordering cost $k = 10$, variable ordering cost $c = 0.25$, and noise distribution μ_0 being the uniform distribution on $[-1, 1]$. The parametric demand model is the simple linear model, that is, $\mathfrak{D}(\eta(p)|\theta) = \langle \eta(p), \theta \rangle$. The holding/backlogging cost function h is a piecewise linear function $h(x) = \overline{h}\max\{x, 0\} - \overline{b}\min\{x, 0\}$, with holding cost $\overline{h} = 0.05$ and backlogging cost $\overline{b} = 1$. Numerical results and comparisons for this model are reported in Figures 1 and 3. Figure 2 reports the results for a three-dimensional model with $\eta(p) = (1, p, p^2)$ and $\theta_0 = (18, -12, -3)$, and the other problem parameters remain the same.

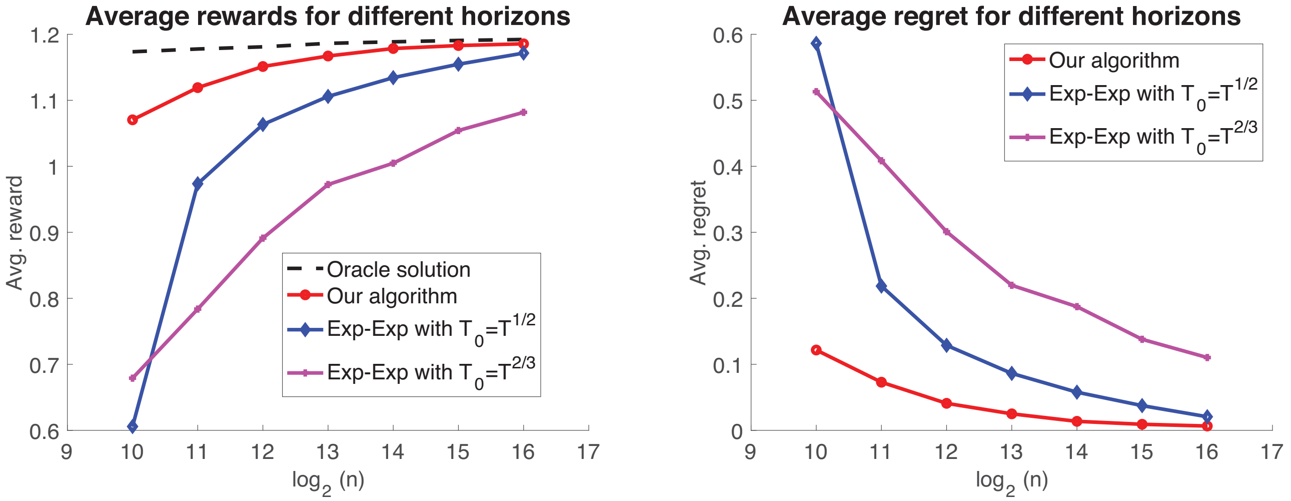
We compare our algorithm (with infrequent changes) to a baseline algorithm using exploration-exploitation, with $T_0 = \sqrt{T}$ and $T_0 = T^{2/3}$ pure exploration phases. More specifically, the baseline algorithm collects data using a random policy during the exploration phase, estimates the unknown demand curve and noise distribution based on the collected demand observations, and computes an empirical optimal $(s, S, \boldsymbol{p})$ that is implemented during the exploitation phase (detailed pseudo codes for the baseline algorithm can be found in the online appendix). Figures 1 to 3 report the average rewards and regret compared with the oracle solution with access to full model information. As we can see, our proposed algorithm significantly outperforms explore-then-exploit baselines, especially in

Figure 1. (Color online) Average Rewards and Regret for Our Algorithm and the Baseline Exploration-Exploitation Algorithm with Varying Time Horizons (x-Axis) on the Two-Dimensional Problem Instance



Note. The oracle solution is based on the solution with full information available.

Figure 2. (Color online) Average Rewards and Regret for Our Algorithm and the Baseline Exploration-Exploitation Algorithm with Varying Time Horizons (x-Axis) on the Three-Dimensional Problem Instance

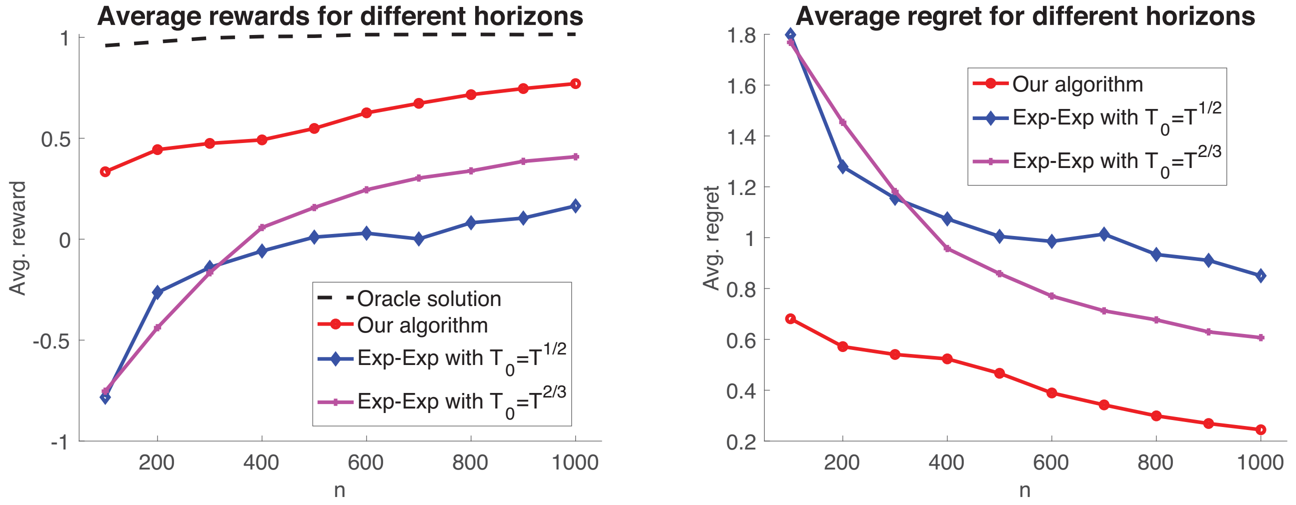


Note. The oracle solution is based on the solution with full information available.

cases where the time horizon is short (e.g., $T = 2^{10} = 1024$ periods). Comparing the convergence rate of our algorithm with the baseline algorithms, one observes the advantage of iterative updating. For small T, the gap between our algorithm and the baseline algorithms is even larger, and this is because the proportion of exploration length is large for small T, yielding a larger percentage of profit loss for baseline algorithms. One also observes that the algorithm with $T_0 = \sqrt{T}$ outperforms that with $T_0 = T^{2/3}$, because the exploration length under $T_0 = T^{2/3}$ is too long, and the profit loss incurred by exploration outweighs the benefit of generating more data.

Table 1 also reports how the $(s, S, \boldsymbol{p})$ policies computed by our algorithm converge to the optimal solution with full information (because the $\boldsymbol{p}$ function is infinite dimensional, we only report the values of $\boldsymbol{p}$ on inventory levels 10, 20, 30, and 40) on the three-dimensional problem instance. We also report the difference in the per-period rewards (Δreward) between the optimal solution and the current solution computed by the bandit algorithm. As we can see, with as few as 500 time periods the policies computed by our algorithm are already very close to the optimal policy in hindsight.

Figure 3. (Color online) Average Rewards and Regret for Our Algorithm and the Baseline Exploration-Exploitation Algorithm with Varying Time Horizons (x-Axis) on the Two-Dimensional Problem Instance, with Short Time Horizons



Note. The oracle solution is based on the solution with full information available.

Table 1. Convergence of the $(s, S, \boldsymbol{p})$ Solution Obtained by Our Proposed Algorithm over $T = 10{,}000$ Selling Periods

	s	S	$\boldsymbol{p}(10)$	$\boldsymbol{p}(20)$	$\boldsymbol{p}(30)$	$\boldsymbol{p}(40)$	Δreward
Optimal solution	2.0	49.8	0.89	0.85	0.78	0.77	0
$t = 50$	2.7	49.1	1.00	1.00	1.00	0.73	0.36
$t = 100$	3.0	52.1	1.00	1.00	0.79	0.75	0.28
$t = 500$	2.3	52.1	0.98	0.92	0.79	0.77	0.06
$t = 1{,}000$	2.3	49.2	0.94	0.87	0.81	0.77	0.05
$t = 5{,}000$	2.0	50.0	0.89	0.85	0.79	0.77	0.02
$t = 10{,}000$	2.0	52.6	0.87	0.85	0.79	0.77	0.02

6.1. Sensitivity Analysis

In this section, we report additional numerical results on the sensitivity of the performance of our algorithm with respect to several model and algorithm parameters.

In the first set of experiment we report the performance of our algorithm with different settings of the upper inventory order-up-to level $\overline{S}$ (the lower limit $\underline{s}$ on the other hand can be conveniently set to zero in most cases), shown in Figure 4. As we can see, when the time horizon is short the performance of our algorithm is insensitive to the aggressive or conservative choices of $\overline{S}$. For slightly longer time horizons, the performance of our algorithm remains stable as long as $\overline{S}$ is not too large.

Next, we report the performance of our algorithm with different values of k (the fixed ordering set-up cost) and c (the variable ordering cost) in Figure 5. As we can see, the regret of our algorithm remains stable under various settings of fixed and variable ordering costs setups, and the regret decreases as the time horizon increases, indicating the effectiveness of our proposed algorithm under different model parameter settings.

7. Conclusion and Future Research

In this paper, we study the problem of coordinating inventory control and pricing with fixed ordering cost and incomplete demand information. By developing an epoch-based UCB approach, our proposed algorithm achieves the optimal $\widetilde{O}(\sqrt{T})$ cumulative regret.

Going beyond this paper, we believe the next important question along this research direction is to handle *censored demand* in the fixed ordering cost inventory model (with pricing components). In a censored demand model, the realized demand d_t per every period is not directly observable; instead only *censored* demands $\max\{d_t, y_t\}$ are observable. The censorship of realized demands impose significant challenges to a UCB-type method because the maximum likelihood models are truncated, and therefore clean confidence bounds are much more difficult to construct.

Another important future research direction is to consider nonparametric demand functions. Our current techniques do not easily extend to the nonparametric regime as the problem becomes significantly different. Indeed, recently Chen et al. (2020) showed

Figure 4. (Color online) Performance Sensitivity with Respect to $\overline{S}$ on Two-Dimensional and Three-Dimensional Problem Instances with Time Horizon Ranging from 200 to 10^4

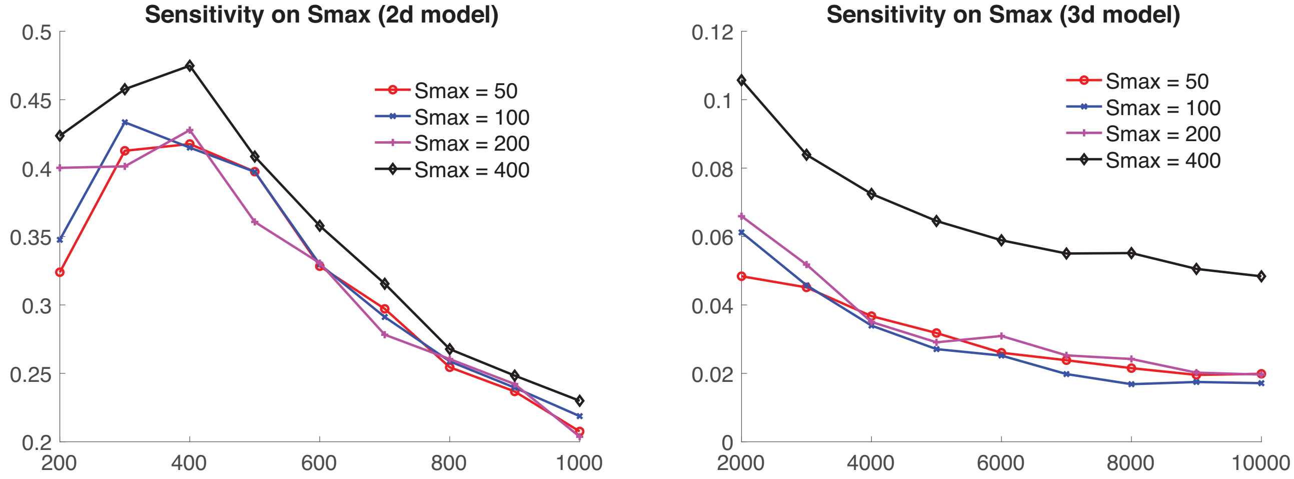
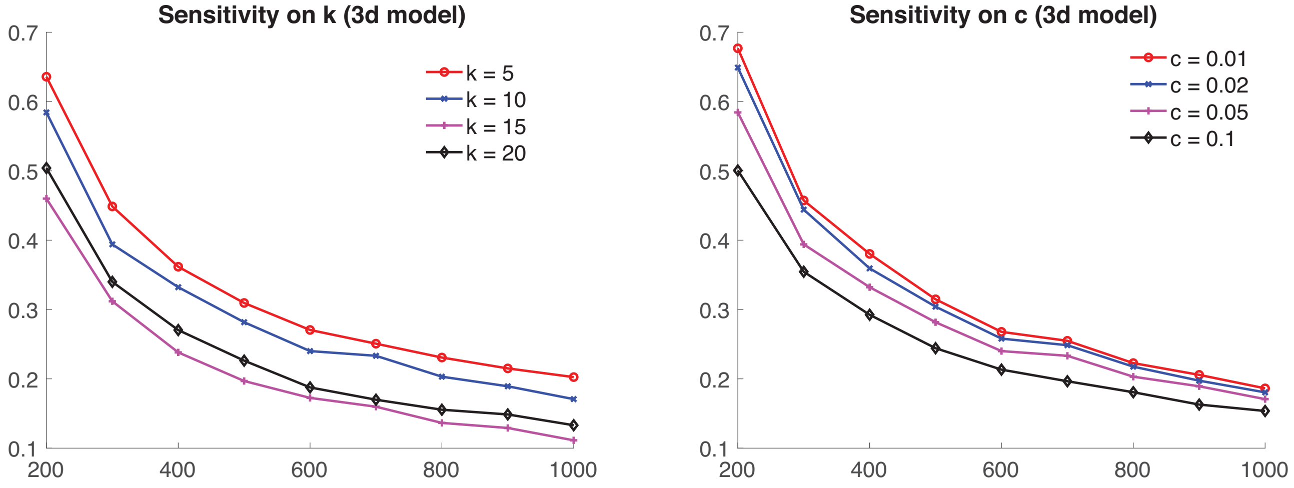


Figure 5. (Color online) Performance Sensitivity with Respect to k (Left) and c (Right) on the Three-Dimensional Problem Instance, with Time Horizon Ranging from 200 to 1,000



an $\widetilde{\Omega}(T^{3/5})$ regret lower bound for the nonparametric setting even when there is no fixed ordering cost, meaning that the problem is inherently more difficult than the parametric setting.

Acknowledgments

The authors thank the department editor, associate editor, and three anonymous referees for detailed and constructive comments that considerably improved the quality of this paper.

Endnotes

[1] The randomness is taken over $\{\beta_t\}_{t\in\mathcal{H}}$ and the internal randomness of $\mathcal{O}$.

[2] This means that re-stocking only occurs at the first time period of each epoch, and an epoch $\mathcal{B}_b$ terminates when $x_t < s_b$ for the first time.

References

Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ, eds. *Adv. Neural Information Processing Systems*, vol. 24 (Curran Associates, Red Hook, NY), 2312–2320.

Agrawal S, Jia R (2019) Learning in structured mdps with convex cost functions: Improved regret bounds for inventory management. Karlin A, ed. *Proc. ACM Conf. Econom. Comput.* (ACM, New York), 743–744.

Auer P (2002) Using confidence bounds for exploitation-exploration trade-offs. *J. Machine Learning Res.* 3(Nov):397–422.

Ban G-Y (2020) Confidence intervals for data-driven inventory policies with demand censoring. *Oper. Res.* 68(2):309–326.

Ban G-Y, Keskin NB (2021) Personalized dynamic pricing with machine learning: High dimensional features and heterogeneous elasticity. *Management Sci.* 67(9):5549–5568.

Ban G-Y, Rudin C (2018) The big data newsvendor: Practical insights from machine learning. *Oper. Res.* 67(1):90–108.

Ban G-Y, Gao Z, Taigel F (2020) Model mis-specification in newsvendor decisions: A comparison of frequentist parametric, bayesian parametric and nonparametric approaches. Preprint, posted January 1, https://papers.ssrn.com/sol3/papers.cfm?abstract_id= 3495733.

Besbes O, Muharremoglu A (2013) On implications of demand censoring in the newsvendor problem. *Management Sci.* 59(6): 1407–1424.

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 (2012) Blind network revenue management. *Oper. Res.* 60(6):1537–1550.

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

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

Bu J, Simchi-Levi D, Wang L (2020) Offline pricing and demand learning with censored data. Preprint, posted July 1, https://dx.doi.org/10.2139/ssrn.3619625.

Burnetas AN, Smith CE (2000) Adaptive ordering and pricing for perishable products. *Oper. Res.* 48(3):436–443.

Chao X, Zhou SX (2006) Joint inventory-and-pricing strategy for a stochastic continuous-review system. *IIE Trans.* 38(5): 401–408.

Chen B, Shi C (2019a) Tailored base-surge policies in dual-sourcing inventory systems with demand learning. Preprint, posted September 27, https://dx.doi.org/10.2139/ssrn.3456834.

Chen B, Chao X, Ahn H-S (2019a) Coordinating pricing and inventory replenishment with nonparametric demand learning. *Oper. Res.* 67(4):1035–1052.

Chen B, Chao X, Shi C (2021) Nonparametric algorithms for joint pricing and inventory control with lost-sales and censored demand. *Math. Oper. Res.* 46(2):726–756.

Chen B, Chao X, Wang Y (2020a) Data-based dynamic pricing and inventory control with censored demand and limited price changes. *Oper. Res.* 68(5):1445–1456.

Chen B, Wang Y, Zhou Y (2020b) Optimal policies for dynamic pricing and inventory control with nonparametric censored demands. Preprint, Posted December 17, https://dx.doi.org/10.2139/ssrn.3750413.

Chen L, Plambeck EL (2008) Dynamic inventory management with learning about the demand distribution and substitution probability. *Manufacturing Service Oper. Management* 10(2):236–256.

Chen Q, Jasin S, Duenyas I (2019b) Nonparametric self-adjusting control for joint learning and optimization of multiproduct pricing with finite resource capacity. *Math. Oper. Res.* 44(2): 601–631.

Chen X, Simchi-Levi D (2004a) Coordinating inventory control and pricing strategies with random demand and fixed ordering cost: The finite horizon case. *Oper. Res.* 52(6):887–896.

Chen X, Simchi-Levi D (2004b) Coordinating inventory control and pricing strategies with random demand and fixed ordering cost: The infinite horizon case. *Math. Oper. Res.* 29 (3):698–723.

Chen X, Simchi-Levi D (2006) Coordinating inventory control and pricing strategies: The continuous review model. *Oper. Res. Lett.* 34(3):323–332.

Chen X, Simchi-Levi D, Philips R, Ozer O, eds. (2012) *Pricing and Inventory Management* (Oxford University Press, Oxford, UK).

Chen Y, Shi C (2019b) Network revenue management with online inverse batch gradient descent method. Preprint, posted February 26, https://dx.doi.org/10.2139/ssrn.3331939.

Chen Y, Ray S, Song Y (2006) Optimal pricing and inventory control policy in periodic-review systems with fixed ordering cost and lost sales. *Naval Res. Logist.* 53(2):117–136.

Cheung WC, Simchi-Levi D (2019) Sampling-based approximation schemes for capacitated stochastic inventory control models. *Math. Oper. Res.* 44(2):668–692.

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

Dani V, Hayes TP, Kakade SM (2008) Stochastic linear optimization under bandit feedback. Servedio RA, Zhang T, eds. *Proc. Conf. Learn. Theory* (Omnipress, Madison, WI), 355–366.

den Boer AV, Keskin NB (2020) Discontinuous demand functions: Estimation and pricing. *Management Sci.* 66(10):4516–4534.

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

Dvoretzky A, Kiefer J, Wolfowitz J (1956) Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* 27(3):642–669.

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.

Federgruen A, Heching A (1999) Combined pricing and inventory control under uncertainty. *Oper. Res.* 47(3):454–475.

Feng Y, Chen FY (2003) *Joint Pricing and Inventory Control with Setup Costs and Demand Uncertainty* (The Chinese University of Hong Kong, Hong Kong).

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

Filippi S, Cappe O, Garivier A, Szepesvári C (2010) Parametric bandits: The generalized linear case. Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A, eds. *Adv. Neural Inform. Processing Systems*, vol. 23 (Curran Associates, Red Hook, NY), 586–594.

Fournier N, Guillin A (2015) On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory Related Fields* 162(3-4):707–738.

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.

Hu P, Lu Y, Song M (2019) Joint pricing and inventory control with fixed and convex/concave variable production costs. *Production Oper. Management* 28(4):847–877.

Huh WT, Janakiraman G (2008) (s, s) optimality in joint inventory-pricing control: An alternate approach. *Oper. Res.* 56(3):783–790.

Huh WT, Rusmevichientong P (2009) A nonparametric asymptotic analysis of inventory planning with censored demand. *Math. Oper. Res.* 34(1):103–123.

Huh WT, Janakiraman G, Muckstadt JA, Rusmevichientong P (2009) An adaptive algorithm for finding the optimal base-stock policy in lost sales inventory systems with censored demand. *Math. Oper. Res.* 34(2):397–416.

Huh WT, Levi R, Rusmevichientong P, Orlin JB (2011) Adaptive data-driven inventory control with censored demand based on kaplan-meier estimator. *Oper. Res.* 59(4):929–941.

Jin C, Yang Z, Wang Z, Jordan MI (2019) Provably efficient reinforcement learning with linear function approximation. Preprint, July 11, https://arxiv.org/abs/1907.05388.

Kantorovich LV, Rubinstein GS (1958) On a space of completely additive functions. *Vestnik Leningradskogo Universiteta* 13(7): 52–59.

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, Li Y, Song J-SJ (2022) Data-driven dynamic pricing and ordering with perishable inventory in a changing environment. *Management Sci.* Forthcoming.

Levi R, Perakis G, Uichanco J (2015) The data-driven newsvendor problem: New bounds and insights. *Oper. Res.* 63(6): 1294–1306.

Levi R, Roundy RO, Shmoys DB (2007) Provably near-optimal sampling-based policies for stochastic inventory control models. *Math. Oper. Res.* 32(4):821–839.

Li Y, Wang Y, Zhou Y (2019) Nearly minimax-optimal regret for linearly parameterized bandits. Preprint, March 30, https://arxiv.org/abs/1904.00242.

Massart P (1990) The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probabilities* 18(3):1269–1283.

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

Petruzzi NC, Dada M (1999) Pricing and the newsvendor problem: A review with extensions. *Oper. Res.* 47(2):183–194.

Polatoglu H, Sahin I (2000) Optimal procurement policies under price dependent demand. *Internat. J. Production Econom.* 65: 141–171.

Qin H, Simchi-Levi D, Wang L (2019) Data-driven approximation schemes for joint pricing and inventory control models. Preprint, submitted March 25, https://dx.doi.org/10.2139/ssrn.3354358.

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

Shi C, Chen W, Duenyas I (2016) Nonparametric data-driven algorithms for multiproduct inventory systems with censored demand. *Oper. Res.* 64(2):362–370.

Song Y, Ray S, Boyaci T (2009) Optimal dynamic joint inventory-pricing control for multiplicative demand with fixed order costs and lost sales. *Oper. Res.* 57(1):245–250.

Thomas LJ (1974) Price and production decisions with random demand. *Oper. Res.* 22(3):513–518.

Wang Y, Chen B, Simchi-Levi D (2019) Multi-modal dynamic pricing. *Management Sci.* 67(10):5969–6627.

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.

Yang J (2020) Learning the best price and ordering policy under fixed costs and ambiguous demand. Preprint, posted April 9, https://dx.doi.org/10.2139/ssrn.3554042.

Yano CA, Gilbert SM, Eliashberg J, Chakravarty A, eds. (2003) *Coordinated Pricing and Production/Procurement Decisions: A Review* (Kluwer, Norwell, MA).

Yin R, Rajaram K (2007) Joint pricing and inventory control with a Markovian demand model. *Eur. J. Oper. Res.* 182(1):113–126.

Yuan H, Luo Q, Shi C (2019) Marrying stochastic gradient descent with bandits: Learning algorithms for inventory systems with fixed costs. *Management Sci.* 67(10):6089–6115.

Zhang H, Chao X, Shi C (2018) Perishable inventory systems: Convexity results for base-stock policies and learning algorithms under censored demand. *Oper. Res.* 66(5):1276–1286.

Zhang H, Chao X, Shi C (2020) Closing the gap: A learning algorithm for lost-sales inventory systems with lead times. *Management Sci.* 66(5):1962–1980.