# A Primal-dual Learning Algorithm for Personalized Dynamic Pricing with an Inventory Constraint

Ningyuan Chen*[1] and Guillermo Gallego†[2]

[1]Rotman School of Management, University of Toronto
[2]Department of Industrial Engineering & Decision Analytics
The Hong Kong University of Science and Technology

**Abstract**

We consider the problem of a firm seeking to use personalized pricing to sell an exogenously given stock of a product over a finite selling horizon to different consumer types. We assume that the type of an arriving consumer can be observed but the demand function associated with each type is initially unknown. The firm sets personalized prices dynamically for each type and attempts to maximize the revenue over the season. We provide a learning algorithm that is near-optimal when the demand and capacity scale in proportion. The algorithm utilizes the primal-dual formulation of the problem and learns the dual optimal solution explicitly. It allows the algorithm to overcome the curse of dimensionality (the rate of regret is independent of the number of types) and sheds light on novel algorithmic designs for learning problems with resource constraints.

**Keywords:** network revenue management, multi-armed bandit, learning and earning, dynamic pricing, online retailing

## 1. Introduction

Dynamic pricing is practiced in many industries including travel, entertainment, and retail. When the capacity cannot be adjusted within the sales horizon, dynamic pricing can increase revenues significantly by adjusting prices in response to the changes in the marginal value of capacity that are driven by the demand and pricing process. The presence of online channels has enabled sellers to use *personalized dynamic pricing* to different consumer types resulting in potentially higher firm profits. The types may correspond to the features of a consumer, such as age, gender and address, which can be observed or inferred through membership programs and browser cookies. Along with opportunities come challenges. The aggregate demand forecasts from historical data, which usually reflects the price sensitivity of the entire market, is of little use. Instead, the firm has to form accurate demand estimates for each type of consumers.

In this paper, we consider a firm selling a product over a finite sales horizon. The inventory is given at the beginning of the horizon and not allowed to be replenished. The inability to order additional

*ningyuan.chen@utoronto.ca
†ggallego@ust.hk

inventory is a hard constraint in the travel industry as it is nearly possible to add capacity to a plane or to a hotel in the short run. In fashion retailing, this is also a hard constraint as production and distribution lead-times may be larger than the sales horizon. We assume there are $M$ different consumer types. The types may be determined in advance by clustering algorithms, which can be applied to each consumer to label its type. We refer the reader to Chapter 8 in Gallego and Topaloglu (2019) for ideas on how to cluster consumer types into a reasonable number of types. Although the type of each consumer is observed by the firm, the demand functions associated with each type are not known initially. Therefore, the firm has to experiment different prices for each type of consumers to learn the demand functions and find the optimal prices. Therefore, it features the exploration/exploitation (learning/earning) trade-off.

We propose a learning algorithm for the personalized dynamic pricing problem described above. Compared to the literature, our algorithm explicitly learns the dual solution, in addition to the optimal prices in the primal space. This allows the algorithm to achieve the near-optimal regret, a measure commonly used to assess learning algorithms.

### 1.1. Contributions and Insights

This paper makes two contributions to the literature. By regarding the demand from the $M$ consumer types as demands for $M$ different products, then our problem is a special case of a network dynamic pricing problem with $M$ products and a single resource constraint. To the best of our knowledge, no algorithm has achieved the optimal rate of regret under the general assumptions.[1] See Section 1.2 and Section 5.1 for more details.

From the perspective of algorithmic design, we demonstrate the feasibility of integrating the primal-dual formulation and learning. The dual variable is not a typical target to learn in the learning literature, because unlike the primal variables, it cannot be experimented directly. In our algorithm, we empirically estimate the Lagrangian function and sequentially form interval estimators for the *dual optimal* solution. This approach may provide novel algorithmic architectures for other learning problems with resource constraints.[2]

This paper provides the following qualitative insights:

- It pays off to explicitly learn the dual optimal solution. The pricing decisions for $M$ types of consumers are coupled through the inventory constraint. However, having an accurate estimator for the dual optimal solution helps to decouple them into $M$ independent learning problems. This is the key reason why our algorithm can achieve the near-optimal regret.

- The learning complexity depends on not only the number of primal decision variables, but also the number of dual variables. As shown by Besbes and Zeevi (2012); Slivkins (2014), a high-dimensional decision vector ($M$ in this case) usually significantly complicates learning, reflecting the curse of dimensionality. Slivkins (2014) shows that the best achievable regret for a generic learning problem without resource constraints is $n^{-1/(2+M)}$ where $M$ is the dimension of the

[1] The algorithm in Chen et al. (2019) achieves the same regret assuming the objective function is infinitely smooth with uniformly bounded derivatives. However, such assumption are often too restrictive. For example, the $k$th derivative of $d(p) = \exp(-ap)$ is not uniformly bounded for $a > 1$ as $k \to \infty$.

[2] In other learning algorithms involving primal-dual explorations such as Badanidiyuru et al. (2013), the dual optimal solution is not learned explicitly. So the design of their algorithm is fundamentally different from ours.

decision vector[3]. In contrast, we are able to obtain the rate $n^{-1/2}$ whose exponent is independent of $M$. This is because the $M$ decision variables can be decoupled if the value of the dual optimal solution is given, and thus the *effective dimension* of the problem is no more than the number of dual variables, which is one in our case.

### 1.2. Literature Review

There is a stream of rapidly growing literature on a firm's pricing problem when the demand function is unknown (e.g. Besbes and Zeevi, 2009; Araman and Caldentey, 2009; Farias and Van Roy, 2010; Broder and Rusmevichientong, 2012; den Boer and Zwart, 2014, 2015; Keskin and Zeevi, 2014; Cheung et al., 2017; Keskin and Zeevi, 2018; den Boer and Keskin, 2020). See den Boer (2015) for a comprehensive survey. Since the firm does not know the optimal price, it has to experiment different (suboptimal) prices and update its belief about the underlying demand function. Therefore, the firm has to balance the exploration/exploitation trade-off, which is usually referred to as the learning-and-earning problem in this line of literature. Among them, our paper is related to those with nonparametric formulations and inventory constraints (Besbes and Zeevi, 2009; Wang et al., 2014; Lei et al., 2017). In addition, we consider personalized dynamic pricing for multiple types of consumers, while most of the above papers consider a single type.

Personalized dynamic pricing can be regarded as a special case of learning with contextual information (Qiang and Bayati, 2016; Javanmard and Nazerzadeh, 2016; Cohen et al., 2016; Ban and Keskin, 2017; Chen and Gallego, 2019; Keskin and Birge, 2019). The main difference of our paper from this stream of literature is summarized below. First, instead of representing the contextual information by a feature vector, we choose to use discrete types to categorize consumers. This could be the outcome of a clustering procedure that pre-processes consumer data. Since the number of types is arbitrary, our setup is merely a technical simplification without losing too much practical generality. Second, we use a nonparametric formulation for the objective function. That is, the demand functions of each type of consumers are only required to satisfy some basic assumptions such as continuity without any specific forms. Third, unlike this literature, we consider an inventory constraint and thus the pricing decision made over time has intertemporal dependence.

The problem studied in this paper is a special case of the multi-product dynamic pricing problem over a network (Gallego and Van Ryzin, 1997) and thus closely related to the literature on demand learning in that setting. Besbes and Zeevi (2012) study the multi-product network revenue management problem with unknown demand functions when the price for each product are chosen from a discrete set. (Hereafter we use network revenue management to highlight the setup of a price menu, in contrast to network dynamic pricing that allows for continuous prices.) The proposed algorithm achieves diminishing regret when the inventory and demand are scaled in proportion. Ferreira et al. (2018) study the same problem as Besbes and Zeevi (2012) and show that Thompson sampling can achieve the rate of regret, $n^{-1/2}$, which is the best one can hope for with even one product and one resource. For continuous prices, however, Besbes and Zeevi (2012) demonstrate that learning may suffer from the curse of dimensionality. The incurred regret may grow at rate $n^{-1/(d+3)}$ with $d$ products (which is equivalent to the number of types in our problem). This is consistent with the result in

[3]The rate of regret usually involves logarithmic terms. When there is no ambiguity, we omit those terms because they are dominated by the polynomial terms.

Slivkins (2014), which studies a generic learning problem without inventory constraints. The *tight* regret bound derived in Slivkins (2014) grows at $n^{-1/(d+2)}$ for $d$ continuous decision variables. Sufficient smoothness may relieve the curse of dimensionality, as argued by Besbes and Zeevi (2012); Chen et al. (2019). In particular, with an infinite degree of smoothness with bounded derivatives, Chen et al. (2019) design an algorithm that almost achieves rate $n^{-1/2}$. Global convexity helps as well, as Chen and Shi (2019) propose a gradient-based algorithm that achieves rate $n^{-1/5}$. In this paper, we present a learning algorithm that achieves the optimal rate $n^{-1/2}$ with one resource constraint and arbitrary number of products (consumer types), without imposing smoothness conditions.

This paper is also related to the vast literature studying multi-armed bandit problems. See Cesa-Bianchi and Lugosi (2006); Bubeck and Cesa-Bianchi (2012) for a comprehensive survey. The classic multi-armed bandit problem involves finite arms. There is a stream of literature studying the so-called continuum-armed bandit problems (Kleinberg, 2005; Auer et al., 2007; Kleinberg et al., 2008; Bubeck et al., 2011), in which there are infinite number of arms (decisions). Slivkins (2014) provides a tight regret bound on a generic learning problem with multiple continuous decision variables; the regret deteriorates exponentially as the number of decisions increases, demonstrating the curse of dimensionality. This line of literature does not consider resource constraints. Recently, there are studies combining multi-armed bandit problems with resource constraints, which is referred to as bandits with knapsacks (BwK) Badanidiyuru et al. (2013, 2014); Agrawal and Devanur (2014). The regret derived in those papers is not directly comparable to ours, because the decisions are discrete in their setting. For discrete decisions, the curse of dimensionality does not emerge. This paper is also related to online convex optimization (OCO); see Shalev-Shwartz et al. (2012) for a review. It is worth pointing out that the duality approach has also been used in BwK and OCO to implement the algorithm and prove the regret bound. However, the dual optimal solution is typically not learned explicitly. One exception is Mahdavi et al. (2013), whose algorithm learns the dual solutions by gradient descent in the primal/dual space. In their OCO setting, the function is given in each period and the gradient can be evaluated, which does not apply to our problem.

## 2. Problem Formulation

Consider a monopolistic firm selling a single product in a finite selling season $T$, with $c$ units of initial inventory. The product cannot be replenished and perishes at the end of the horizon with zero salvage values. There are $M$ types of consumers. Consumers with the same type have similar features such as education backgrounds, ages, and addresses. The firm observes the type of each arriving consumer, and is allowed to price-discriminate according to the type. This is referred to as personalized pricing, which is increasingly popular in online retailing due to the observation that the demand function differs dramatically across types. Therefore, we model the arrival of type-$m$ consumers by a Poisson process with instantaneous rate $d_m(p_m(t))$, where $p_m(t)$ is the price charged for type-$m$ consumers at time $t$ and $d_m(\cdot)$ is the demand function of type-$m$ consumers.

We focus on the case that the information of the demand function associated with each type and the type distribution among the population is absent at the beginning of the season. The objective of the firm is to maximize the expected revenue collected over the horizon, subject to the inventory constraint. To achieve the goal, the firm has to learn the demand function $d_m(\cdot)$ for $m = 1, \ldots, M$ and

the associated optimal prices in the process. We first characterize the problem when all the information is available.

## 2.1. The full-information Benchmark

When $d_m(\cdot)$ is known to the firm, the firm's objective is to maximize

$$\begin{aligned} J(T,c) = \max_{\boldsymbol{p}(t)\in\mathcal{F}_t} \quad & \mathsf{E}\left[\sum_{m=1}^{M}\int_{t=0}^{T} p_m(t)dN_{m,t}(d_m(p_m(t)))\right] \\ \text{subject to} \quad & \sum_{m=1}^{M}\int_{t=0}^{T} dN_{m,t}(d_m(p_m(t))) \le c, \end{aligned} \tag{1}$$

where $\boldsymbol{p}(t) = \{p_1(t),\ldots,p_M(t)\}$ is a pricing policy that is adapted to the filtration $\mathcal{F}_t$ associated with the sales process, and $N_{m,t}(\lambda_t)$ is an independent Poisson process with instantaneous rate $\lambda_t$. When the inventory is depleted, then $\boldsymbol{p}(t)$ is forced to be $\boldsymbol{p}_\infty$, which is a menu of choke prices at which future demand from all types is turned off.

A classic approach in revenue management (e.g., see Gallego and Van Ryzin 1997) to this problem is to consider the fluid approximation of (1). That is

$$\begin{aligned} J^D(T,c) = \max_{\boldsymbol{p}(t)} \quad & \sum_{m=1}^{M}\int_{t=0}^{T} p_m(t)d_m(p_m(t))dt \\ \text{subject to} \quad & \sum_{m=1}^{M}\int_{t=0}^{T} d_m(p_m(t))dt \le c, \end{aligned} \tag{2}$$

where we have replaced the Poisson process $N_{m,t}(d_m(p_m(t))$ in the original formulation by the intensity $d_m(p_m(t))$. Note that the fluid approximation (2) is a deterministic optimization problem. Before presenting the primal-dual formulation of (2), we make the following standard assumption:

**Assumption 1.** For a price domain $p \in [0, p_\infty]$ and $m = 1,\ldots,M$, we assume

1. The demand $d_m(p)$ is strictly decreasing with an inverse function $d_m^{-1}(\cdot)$ and bounded with $d_m(p) \le M_1$.
2. Define the revenue rate as a function of the demand rate $\lambda$, $r_m(\lambda) \triangleq \lambda d_m^{-1}(\lambda)$. The functions $r_m(\lambda)$, $d_m(p)$ and $d_m^{-1}(\lambda)$ are Lipschitz continuous with factor $M_2$.
3. $r(\lambda)$ is twice-differentiable and strictly concave, $0 < M_3 \le -r_m''(\lambda) \le M_4$.

*Remark* 1. Assumption 1 implies that both $d_m(p)$ and $d_m^{-1}(\lambda)$ are differentiable. The derivatives are bounded by the interval $[-M_2, -1/M_2]$.

It is easy to verify that this mild assumption is satisfied by most demand functions, such as the exponential demand $d(p) = a\exp(-bp)$ and the linear demand $d(p) = a - bp$. The exponential demand, on the other hand, does not have uniformly bounded derivatives of any orders, required by Chen et al. (2019).

### Primal-dual Formulation

Consider the Lagrangian function for the fluid approximation (2)

$$L(\boldsymbol{p}(t), z) = cz + \sum_{m=1}^{M} \int_{t=0}^{T} (p_m(t) - z) d_m(p_m(t)) dt \tag{3}$$

and the dual function

$$g(z) = \max_{\boldsymbol{p}(t)} \{L(\boldsymbol{p}(t), z)\} = cz + \max_{\boldsymbol{p}(t)} \left\{ \sum_{m=1}^{M} \int_{t=0}^{T} (p_m(t) - z) d_m(p_m(t)) dt \right\}. \tag{4}$$

Under Assumption 1, the following quantities are well defined

$$\mathcal{R}_m(z) \triangleq \max_{p} \{d_m(p)(p - z)\} \quad \text{and} \quad \mathcal{P}_m(z) \triangleq \operatorname*{argmax}_{p} \{d_m(p)(p - z)\}. \tag{5}$$

$\mathcal{R}_m(z)$ and $\mathcal{P}_m(z)$ can be interpreted as the optimal value and optimal solution of the profit-maximization problem for type-$m$ consumers when the unit cost is $z$. They are closely related to the dual function (4) as $g(z) = cz + T \sum_{m=1}^{M} \mathcal{R}_m(z)$; provided with a dual variable $z$, the optimal $\boldsymbol{p}(t)$ in (4) is time-invariant: $p_m(t) \equiv \mathcal{P}_m(z)$. The properties are summarized below (see also Gallego and Topaloglu 2019):

**Proposition 1.** *Under Assumption 1, we have*

1. *$\mathcal{P}_m(z)$ is increasing in $z$ and $\mathcal{P}'_m(z)$ is bounded.*
2. *$\mathcal{R}_m(z)$ is decreasing and convex in $z$; $\mathcal{R}'_m(z) = -\sum_{m=1}^{M} d_m(\mathcal{P}_m(z))$.*
3. *$g(z)$ is twice differentiable and strictly convex.*
4. *Let $z^* \triangleq \operatorname{argmin}_{z \geq 0}\{g(z)\}$. The optimal solution to (2) is $p_m(t) \equiv p_m^* \triangleq \mathcal{P}_m(z^*)$. Moreover, complementary slackness holds: $z^*(c - T \sum_{m=1}^{M} d_m(\mathcal{P}_m(z^*))) = 0$.*

*Remark* 2. To simplify the notation, we use the same set of constants as in Assumption 1 and assume that $0 < M_3 \leq g''(z) \leq M_4$; $\mathcal{P}_m(z)$ is Lipschitz continuous with factor $M_2$.

Proposition 1 states that the fluid approximation (2) admits a time-invariant pricing policy $p_m^*$. Moreover, the optimal solution is closely related to the dual optimal solution $z^*$, which is usually interpreted as the shadow cost of inventory. When $z^* > 0$, the initial capacity is insufficient and thus the inventory constraint is binding by complementary slackness: $c = T \sum_{m=1}^{M} d_m(\mathcal{P}_m(z^*))$. When $z^* = 0$, the inventory is sufficient and the optimal price $p_m^* = \mathcal{P}_m(0)$ maximizes the revenue rate $pd_m(p)$ as if there is no inventory constraint.

### Scaled Demand and Capacity

The connection between $J^D(T, c)$ and $J(T, c)$ has been studied extensively in the revenue management literature. In particular, Gallego and Van Ryzin (1997) find that the revenue from the fluid approximation is an upper bound for the stochastic problem, i.e., $J^D(T, c) \geq J(T, c)$. The tie becomes closer when the demand and capacity scale in proportion: if we index a sequence of systems by $n$ and let $d_{m,n}(\cdot) = nd_m(\cdot)$ and $c_n = nc$ in the $n$th system, then the revenues satisfy $J_n^D(T, c) - J_n(T, c) = O(\sqrt{n})$.

Since $J_n^D(T,c)$ scales linearly in $n$, the gap between $J_n^D(T,c)$ and $J_n(T,c)$ diminishes relative to the earned revenue as $n$ grows. More importantly, the optimal solution to (2), $\{p_m^*\}_{m=1}^M$, which also maximizes the fluid approximation for the scaled system $J_n^D(T,c)$, performs well in the stochastic problem (1) as a special suboptimal pricing policy (it is constant and thus adapted to $\mathcal{F}_t$). More precisely, the expected revenue for $\{p_m^*\}_{m=1}^M$ in the $n$th stochastic system satisfies

$$J_n^D(T,c) - \mathsf{E}\left[\sum_{m=1}^M \int_{t=0}^{\tau} p_m^* dN_{m,t}(nd_m(p_m^*))\right] = O(\sqrt{n}),$$

where $\tau$ is the minimum of $T$ and the stopping time when the inventory reaches zero. Combined with the fact that $J_n^D \geq J_n$, $\{p_m^*\}_{m=1}^M$ is near-optimal in the $n$th stochastic system. Therefore, the prices $\{p_m^*\}_{m=1}^M$ are the goal of our learning policy when $d_m(\cdot)$ is not known to the firm.

## 2.2. The Learning Policy and the Target Regret

Suppose the firm does not know $d_m(\cdot)$ at the beginning of the horizon. To earn high expected revenues over the horizon, the firm adopts an $\mathcal{F}_t$-predictable pricing policy $\pi$. That is, at time $t$, $\pi_t$ only depends on the adopted prices and the observed sales for each type of consumers prior to $t$. It then outputs a price vector $\{P_1(t),\ldots,P_m(t)\}$ for each type of consumers at time $t$. We denote the expected revenue associated with a policy $\pi$ by $J^\pi(T,c)$. Clearly, the unavailability of the information regarding $d_m(\cdot)$ incurs a cost to the firm, and thus $J^\pi(T,c) \leq J(T,c) \leq J^D(T,c)$.

The objective of this study is to design a policy so that $J(T,c) - J^\pi(T,c)$ is small, especially when demand and capacity are scaled. Therefore, similar to Besbes and Zeevi (2009), we consider the following criterion, referred to as the regret, of a policy $\pi$:

$$R_n^\pi(T,c) = 1 - \frac{J_n^\pi(T,c)}{J_n^D(T,c)}, \tag{6}$$

where $J_n^\pi(T,c)$ is the expected revenue $\pi$ generates in the $n$th stochastic system. Note that $\pi$ may depend on $n$, and we suppress the dependence to simplify notations. The regret measures the revenue loss $J_n^D(T,c) - J_n^\pi(T,c)$ relative to $J_n^D(T,c)$. The goal of the policy $\pi$ is to ensure $\lim_{n\to\infty} R_n^\pi(T,c) = 0$. That is, the learning incurs no significant cost for large systems.

It has been shown in Besbes and Zeevi (2009); Wang et al. (2014) that for $M = 1$, any learning policy incurs regret whose rate is no less than $n^{-1/2}$ for some problem instances. Indeed, even if we replace $J_n^\pi(T,c)$ by $J_n(T,c)$, which is the full-information upper bound for $J_n^\pi(T,c)$, the quantity (6) is of order $n^{-1/2}$ by the discussion in Section 2.1. In other words, one cannot expect to design a learning policy whose regret grows slower than $n^{-1/2}$. Thus, $n^{-1/2}$ (possibly with logarithmic terms in $n$) is the target regret of our proposed learning policy. It is also worth noticing that we adopt a nonparametric formulation. In parametric formulations, one may estimate the parameter using historical data and conduct a policy that maximizes the objective based on the estimator, without deliberate exploration. The so-called certainty-equivalence control may or may not lead to incomplete learning as shown by Keskin and Zeevi (2018). In the nonparametric setting, a wider range of exploration seems mandatory because no global information is learned through local experimentation.

## 3. The Primal-dual Learning Algorithm

In this section, we introduce an algorithm (learning policy) based on the primal-dual formulation. We first explain the steps of the algorithm, and then analyze its regret. Combined with the lower bound for regret in Besbes and Zeevi (2009); Wang et al. (2014), the regret of the algorithm achieves near optimality for the problem considered in Section 2.

Before proceeding, we state what the firm has information of initially. The firm knows $M$, $T$, $n$, $c$, and the constants specified by Assumption 1. Moreover, we impose a mild assumption in addition to Assumption 1.

**Assumption 2.** There exist intervals $[\underline{p}, \overline{p}]$ and $[0, \overline{z}]$ such that $p_m^* \in (\underline{p}, \overline{p})$ and $z^* \in (0, \overline{z})$ for all $m$. Moreover, $\{\mathcal{P}_m(z) : z \in [0, \overline{z}]\} \subset [\underline{p}, \overline{p}]$. The intervals $[\underline{p}, \overline{p}]$ and $[0, \overline{z}]$ are known to the firm.

Note that $p_m^*$ and $z^*$ are the primal/dual optimal solutions to the fluid approximation (2). Assumption 2 states that although the firm does not know the optimal solutions, it does have the information of their ranges. Since $\underline{p}, \overline{p}$ and $\overline{z}$ can be arbitrary finite numbers, this assumption is not restrictive.

### 3.1. The Intuition

We first explain the intuition behind the algorithm. If the firm knew the full information, then it would have found the pricing policy $p_m^*$ through the primal-dual formulation

$$\begin{aligned} z^* &= \operatorname*{argmin}_{z \geq 0} \{g(z)\} = \operatorname*{argmin}_{z \geq 0} \left\{ cz + T \sum_{m=1}^{M} \mathcal{R}_m(z) \right\} \\ p_m^* &= \mathcal{P}_m(z^*) \quad \forall m = 1, \ldots, M. \end{aligned} \tag{7}$$

When the information of $d_m(p)$ is not available, both optimization problems are unsolvable. However, the firm can experiment with different prices and use the observed sales as a noisy but unbiased estimator for $d_m(p)$ at those prices. The noisy estimator is a Poisson random variable. Then, the firm could plug the estimators into (7) to solve them "empirically", obtaining estimators for $z^*$ and $p_m^*$ for $m = 1, \ldots, M$. One would imagine that the estimators for $z^*$ and $p_m^*$ are not necessarily accurate. Indeed, the accuracy of such a procedure depends crucially on two aspects:

1. The length of the period during which the price is experimented. The longer the period, the less noisy the estimators for $d_m(p)$ are.

2. The granularity of the experimented prices. The estimator for $d_m(p)$ is based on a discrete set of prices. It inevitably incurs discretization error in order to solve a continuous optimization problem (7).

Ideally, to obtain accurate estimators for $z^*$ and $p_m^*$, the firm would set a refined grid of prices for each type of consumers and try each price for a long period during the season. Those suboptimal prices, however, lead to substantial revenue loss.

To solve the exploration/exploitation dilemma, we divide $[0, T]$ into multiple phases. After each phase, the sales for each type of consumers during the phase are observed at a set of prices to form estimators for $d_m(p)$. Then (7) is solved empirically to obtain point estimators for $z^*$ and $p_m^*$. In the

next phase, those point estimators are used to form *interval estimators* for $z^*$ and $p_m^*$. The interval estimators help to narrow down the range of prices to experiment. Therefore, as the algorithm enters new phases, the burden to explore is gradually relieved and it can afford to try a more refined price grid for a longer period of time. The revenue loss is also limited because the experimented prices fall into a narrow interval containing $p_m^*$ with high probability.

### 3.2. Description of the Algorithm

Next we explain the details of the algorithm. Let $\{P_1(t), \ldots, P_M(t)\}$ be the stochastic pricing policy associated with $\pi$. Without further mention, we always suppose that when the inventory is depleted, $P_m(t)$ is automatically switched to the choke price $p_\infty$ for all $m$. Given $n$, the algorithm divides $[0, T]$ into consecutive phases $k = 1, 2, \ldots, K$. The length of phase $k$ is $\tau^{(k)}$. We also denote the beginning of phase $k$ by $t_k$. Thus $t_k = \sum_{i=1}^{k-1} \tau^{(i)}$. Let $\epsilon > 0$ be a small constant.

At the beginning of phase $k$, the firm has interval estimators for $p_m^*$ and $z^*$, $[\underline{p}_m^{(k)}, \overline{p}_m^{(k)}]$ and $[\underline{z}^{(k)}, \overline{z}^{(k)}]$, obtained from the last phase, which ensure that $p_m^* \in [\underline{p}_m^{(k)}, \overline{p}_m^{(k)}]$ and $z^* \in [\underline{z}^{(k)}, \overline{z}^{(k)}]$ with high probability. During phase $k$, the price interval $[\underline{p}_m^{(k)}, \overline{p}_m^{(k)}]$, whose length is denoted $\Delta_m^{(k)}$, is discretized to $N^{(k)}+1$ equally spaced grid points, i.e., $p_{m,j}^{(k)} \triangleq \underline{p}_m^{(k)} + j\delta_m^{(k)}$ for $j = 0, \ldots, N^{(k)}$ and $\delta_m^{(k)} \triangleq \Delta_m^{(k)}/N^{(k)}$. During phase $k$, the algorithm sets price $p_{m,j}^{(k)}$ for type-$m$ consumers for a period of length $\tau^{(k)}/(N^{(k)}+1)$.

At the end of phase $k$, the observed sales, $D_{m,j}^{(k)}$, from type-$m$ consumers at price $p_{m,j}^{(k)}$ is a Poisson random variable with mean $nd_m(p_{m,j}^{(k)})\tau^{(k)}/(N^{(k)}+1)$. Therefore, an unbiased estimate for $d_m(p_{m,j}^{(k)})$ is

$$\hat{d}_{m,j}^{(k)} \triangleq \frac{N^{(k)}+1}{n\tau^{(k)}} D_{m,j}^{(k)}.$$

To form a point estimator for $z^*$, the firm substitutes $\hat{d}_{m,j}^{(k)}$ into $g(z)$ (the right-hand side of the first equation of (7)), i.e.,

$$g(z) = cz + T\sum_{m=1}^{M} \max_{p_m} d_m(p_m)(p_m - z) \approx cz + T\sum_{m=1}^{M} \max_{j_m=0,\ldots,N^{(k)}} \hat{d}_{m,j_m}^{(k)}(p_{m,j_m}^{(k)} - z).$$

To find $z \in [\underline{z}^{(k)}, \overline{z}^{(k)}]$ that maximizes the above expression, the firm divides $[\underline{z}^{(k)}, \overline{z}^{(k)}]$, whose length is denoted $\Delta_z^{(k)}$, into $N_z^{(k)}$ equally spaced grid points, $\underline{z}^{(k)} + j\delta_z^{(k)}$ for $j = 0, \ldots, N_z^{(k)}$ and $\delta_z^{(k)} \triangleq \Delta_z^{(k)}/(N_z^{(k)}+1)$. Therefore, a point estimator for $z^*$ can be obtained as follows[4]:

$$z^{(k)*} \triangleq \operatorname*{argmin}_{z \in \left\{\underline{z}^{(k)} + i\delta_z^{(k)}\right\}_{i=0}^{N_z^{(k)}}} \left\{ cz + T\sum_{m=1}^{M} \max_{j_m=0,\ldots,N^{(k)}} \hat{d}_{m,j_m}^{(k)}(p_{m,j_m}^{(k)} - z) \right\}. \tag{8}$$

Based on $z^{(k)*}$, the firm can obtain point estimators for $p_m^*$ by the second equation in (7):

$$p_m^{(k)*} \triangleq p_{m,j_m^*}^{(k)}, \quad \text{where} \quad j_m^* = \operatorname*{argmax}_{j_m=0,\ldots,N^{(k)}} \hat{d}_{m,j_m}^{(k)}(p_{m,j_m}^{(k)} - z^{(k)*}). \tag{9}$$

[4]Alternatively, the firm can find $z^{(k)*}$ by solving the first-order condition for $g(z)$, $c - T\sum_{m=1}^{M} d_m(\mathcal{P}_m(z)) = 0$, using the empirical version of $d_m$ and $\mathcal{P}_m$. The regret analysis holds for this case. Also note that in theory we can find the optimal $z^{(k)*}$ exactly without discretization, as the equation is solved offline. The discretization is for practical purposes.

This completes the procedure in phase $k$.

At the beginning of phase $k$+1, the firm constructs interval estimators $[\underline{p}_m^{(k+1)}, \overline{p}_m^{(k+1)}]$ ($[\underline{z}^{(k+1)}, \overline{z}^{(k+1)}]$) based on the point estimators $p_m^{(k)*}$ ($z^{(k)^*}$) and pre-specified width $\bar{\Delta}^{(k+1)}$ ($\bar{\Delta}_z^{(k+1)}$) for all $m$:

$$\begin{aligned}
\underline{p}_m^{(k+1)} &= \max\left\{\underline{p}, p_m^{(k)*} - \frac{\bar{\Delta}^{(k+1)}}{2}\right\}, \quad \overline{p}_m^{(k+1)} = \min\left\{\overline{p}, p_m^{(k)*} + \frac{\bar{\Delta}^{(k+1)}}{2}\right\} \\
\underline{z}^{(k+1)} &= \max\left\{0, z^{(k)*} - \frac{\bar{\Delta}_z^{(k+1)}}{2}\right\}, \quad \overline{z}^{(k+1)} = \min\left\{\overline{z}, z^{(k)*} + \frac{\bar{\Delta}_z^{(k+1)}}{2}\right\}.
\end{aligned} \tag{10}$$

Note that the intervals are properly truncated by $[\underline{p}, \overline{p}]$ and $[0, \overline{z}]$, and this is the only reason why $\bar{\Delta}^{(k+1)}$ ($\bar{\Delta}_z^{(k+1)}$) can potentially be different from $\Delta_m^{(k+1)}$ ($\Delta_z^{(k+1)}$). Then the procedure is repeated for phase $k+1$.

In the last phase, phase $K$, the algorithm behaves differently after forming the interval estimators $[\underline{p}_m^{(K)}, \overline{p}_m^{(K)}]$ and $[\underline{z}^{(K)}, \overline{z}^{(K)}]$. At the beginning of phase $K$, the firm checks whether $0 \in [\underline{z}^{(K)}, \overline{z}^{(K)}]$. If so, then with high probability $z^* = 0$ and the capacity is sufficient. Therefore, the price $p_m^*$ is the unconstrained maximizer of $pd_m(p)$, i.e., $\mathcal{P}_m(0)$, for all $m$. As we will show in Section 3.3, the width of the interval estimator $[\underline{p}_m^{(K)}, \overline{p}_m^{(K)}]$ is roughly $\bar{\Delta}^{(K)} \sim n^{-1/4}$. Therefore, if the firm adheres to a constant price $p_m \in [\underline{p}_m^{(K)}, \overline{p}_m^{(K)}]$ for type-$m$ consumers, the relative revenue loss for type-$m$ consumers in phase $K$ (ignoring the random fluctuation of Poisson arrivals) is approximately

$$|p_m^* d_m(p_m^*) - pd_m(p)| \sim (d_m(p_m^*) - d_m(p_m))^2 \sim (p_m^* - p_m)^2 \sim (\bar{\Delta}^{(K)})^2 \sim n^{-1/2},$$

where we rely on the concavity in Assumption 1 and the fact that $p_m^* \in [\underline{p}_m^{(K)}, \overline{p}_m^{(K)}]$ with high probability. This meets the target regret in Section 2.2. Motivated by the argument above, the algorithm simply charges a constant price $p_m^{(K)} = \overline{p}_m^{(K)} + \alpha$ for type-$m$ consumers until the end of the season for a pre-specified parameter $\alpha$. Note that we slightly mark up the prices by a small adjustment $\alpha \sim n^{-1/4}$ to guarantee that the inventory is sufficient when the inventory just meets the unconstrained optimal prices, i.e., $c = T\sum_{m=1}^M d_m(p_m^*)$.

If the firm finds $0 \notin [\underline{z}^{(K)}, \overline{z}^{(K)}]$ at the beginning of phase $K$, which implies that $z^* > 0$ and the capacity is insufficient with high probability, then a different procedure has to be used in phase $K$. The method for the case of $z^* = 0$ no longer works: because $p_m^*$ is no longer the unconstrained maximizer of $pd_m(p)$, even for $p_m, p_m^* \in [\underline{p}_m^{(K)}, \overline{p}_m^{(K)}]$, we have

$$|p_m^* d_m(p_m^*) - pd_m(p)| \sim |p_m^* - p_m| \sim |\bar{\Delta}^{(K)}| \sim n^{-1/4}. \tag{11}$$

This implies that a constant price in $[\underline{p}_m^{(K)}, \overline{p}_m^{(K)}]$ for type-$m$ consumers will fail to meet the target regret. To address the problem, let $p_m^l = \underline{p}_m^{(K)} - \alpha$ and $p_m^u = \overline{p}_m^{(K)} + \alpha$ be conservative lower and upper bounds for $p_m^*$. It makes sure that $p_m^* \in [p_m^l, p_m^u]$ along with buffers around the boundary. The buffer guarantees that the solution to the linear interpolation (14) below is stable. Let $S(t)$ be the cumulative sales aggregated from all types of consumers up to time $t$. The algorithm in phase $K$ is divided into the following four steps:

1. For $t \in (t_K, t_K + (\log n)^{-\epsilon}]$, apply $p_m^l$ to type-$m$ consumers. Record the aggregate sales rate by

$D_l^{(K)}$. That is

$$D_l^{(K)} \triangleq (\log n)^{\epsilon} \sum_{m=1}^{M} \int_{t_K}^{t_K+(\log n)^{-\epsilon}} dN_{m,t}(nd_m(p_m^l)). \tag{12}$$

Clearly, $D_l^{(K)}$ is an unbiased estimator for $n\sum_{m=1}^{M} d_m(p_m^l)$.

2. For $t \in (t_K + (\log n)^{-\epsilon}, t_K + 2(\log n)^{-\epsilon}]$, apply $p_m^u$ to type-$m$ consumers. Record the aggregate sales rate $D_u^{(K)}$:

$$D_u^{(K)} \triangleq (\log n)^{\epsilon} \sum_{m=1}^{M} \int_{t_K+(\log n)^{-\epsilon}}^{t_K+2(\log n)^{-\epsilon}} dN_{m,t}(nd_m(p_m^u)), \tag{13}$$

which is an unbiased estimator for $n\sum_{m=1}^{M} d_m(p_m^u)$.

3. At $t = t_K + 2(\log n)^{-\epsilon}$, solve $\theta \in [0,1]$ from

$$(T - t_K)(\theta D_l^{(K)} + (1-\theta)D_u^{(K)}) = nc - S(t_K). \tag{14}$$

If the solution $\theta \notin [0,1]$ (which will be shown to have negligible probability), then we project it to $[0,1]$. To interpret $\theta$, note that $nc - S(t_K)$ is the remaining inventory at $t_K$, the beginning of phase $K$. If $D_l^{(K)}$ and $D_u^{(K)}$ were equal to their means, $n\sum_{m=1}^{M} d_m(p_m^l)$ and $n\sum_{m=1}^{M} d_m(p_m^u)$, then in a fluid system starting from $t_K$ with inventory $nc - S(t_K)$, applying $p_m^l$ for type-$m$ consumers for a period of length $\theta(T - t_K)$ and $p_m^u$ for a period of length $(1-\theta)(T - t_K)$ would make the inventory reach zero right at $T$, according to (14).

4. For $t \in (t_K + 2(\log n)^{-\epsilon}, T]$, apply $p_m^l$ for a period of length $\theta(T - t_K) - (\log n)^{-\epsilon}$, and $p_m^u$ for a period of length $(1-\theta)(T - t_K) - (\log n)^{-\epsilon}$ until $T$.

The goal of the above steps is to ensure the deviation of $S(T)$ from $nc$ is relatively small. In particular, from Lemma 8 in Section 4, the steps guarantee $|S(T) - nc| \sim n^{-1/2}$. Without further exploring the price space[5] the algorithm can still meet the target regret with a little exploration on the aggregate demand rate and by controlling the aggregate sales at $T$. We will discuss this point in Section 5. The notations are summarized in Table 1 in the appendix. The detailed steps of the algorithm are demonstrated in Algorithm 1. Note that both "Input" and "Constant" are known to the firm, while "Parameters" are computed in Section 3.3.

[5]Recall that the interval estimators for $p_m^*$, $\bar{\Delta}^{(K)} \sim n^{-1/4}$, are still too wide to meet the target regret.

**Algorithm 1** The Primal-dual Learning Algorithm

Input: $n$, $c$, $T$, $M$

Constants: $M_1$, $M_2$, $M_3$, $M_4$, $\underline{p}$, $\overline{p}$, $\overline{z}$

Parameters: $\epsilon$, $\alpha$, $K$, $\{\tau^{(k)}\}_{k=1}^{K-1}$, $\{\bar{\Delta}^{(k)}, \bar{\Delta}_z^{(k)}, N^{(k)}, N_z^{(k)}\}_{k=1}^{K}$

Initialize: $\underline{p}_m^{(1)} = \underline{p}$, $\overline{p}_m^{(1)} = \overline{p}$, $\underline{z}^{(1)} = 0$, $\overline{z}^{(1)} = \overline{z}$

**for** $k = 1$ to $K - 1$ **do**

  $t_k \leftarrow \sum_{i=1}^{k-1} \tau^{(i)}$ ▷ The start of phase $k$

  $\Delta_m^{(k)} \rightarrow \overline{p}_m^{(k)} - \underline{p}_m^{(k)}$ and $\delta_m^{(k)} \leftarrow \Delta_m^{(k)}/(N^{(k)} + 1)$ for $m = 1, \ldots, M$

  **for** $i = 0$ to $N^{(k)}$ **do**

    **for** $t = t_k + i\frac{\tau^{(k)}}{N^{(k)}+1}$ to $t_k + (i+1)\frac{\tau^{(k)}}{N^{(k)}+1}$ **do**

      Charge price $p_{m,i}^{(k)} \leftarrow \underline{p}_m^{(k)} + i\delta_m^{(k)}$ to type-$m$ consumers

      Record the observed sales $D_{m,i}^{(k)}$ for $p_{m,i}^{(k)}$ for $m = 1, \ldots, M$

    **end for**

    $\hat{d}_{m,i}^{(k)} \leftarrow \frac{N^{(k)}+1}{n\tau^{(k)}} D_{m,i}^{(k)}$ for $m = 1, \ldots, M$ ▷ Empirical estimate for $d_m(p_{m,i}^{(k)})$

  **end for**

  $\Delta_z^{(k)} \leftarrow \overline{z}^{(k)} - \underline{z}^{(k)}$ and $\delta_z^{(k)} = \Delta_z^{(k)}/(N_z^{(k)} + 1)$

  Obtain $z^{(k)*}$ according to (8) ▷ Estimate the dual optimal solution

  Obtain $p_m^{(k)*}$ according to (9) for $m = 1, \ldots, M$ ▷ Estimate the primal optimal solution

  Obtain $\underline{z}^{(k+1)}$, $\overline{z}^{(k+1)}$, $\underline{p}_m^{(k+1)}$, $\overline{p}_m^{(k+1)}$ according to (10) for $m = 1, \ldots, M$ ▷ Obtain the interval estimators

**end for**

$t_K \leftarrow \sum_{i=1}^{K-1} \tau^{(i)}$ ▷ The beginning of phase $K$

**if** $0 \in [\underline{z}^{(K)}, \overline{z}^{(K)}]$ **then** ▷ Sufficient capacity

  **for** $t = t_K$ to $T$ **do**

    Charge price $p_m^{(K)} \leftarrow \overline{p}_m^{(K)} + \alpha$ to type-$m$ consumers

  **end for**

**else** ▷ Insufficient capacity

  $p_m^l \leftarrow \underline{p}_m^{(K)} - \alpha$ and $p_m^u \leftarrow \overline{p}_m^{(K)} + \alpha$

  **for** $t = t_K$ to $t_K + (\log n)^{-\epsilon}$ **do**

    Charge price $p_m^l$ to type-$m$ consumers

    Record the aggregated sales rate $D_l^{(K)}$ according to (12)

  **end for**

  **for** $t = t_K + (\log n)^{-\epsilon}$ to $t_K + 2(\log n)^{-\epsilon}$ **do**

    Charge price $p_m^u$ to type-$m$ consumers

    Record the aggregated sales rate $D_u^{(K)}$ according to (13)

  **end for**

  Let $\theta$ be the projection of the solution to (14) to $[0, 1]$

  **for** $t = t_K + 2(\log n)^{-\epsilon}$ to $t_K + \theta(T - t_K) + (\log n)^{-\epsilon}$ **do**

    Charge price $p_m^l$ to type-$m$ consumers

  **end for**

  **for** $t = t_K + \theta(T - t_K) + (\log n)^{-\epsilon}$ to $T$ **do**

    Charge price $p_m^u$ to type-$m$ consumers

  **end for**

**end if**

### 3.3. Choice of Parameters

Let $\epsilon$ be a sufficiently small constant (independent of $n$). We set the following parameter values in Step 3:

$$\begin{aligned}
\alpha &= (\log n)^{1+9\epsilon} n^{-1/4} \\
\tau^{(k)} &= n^{-(1/2)(3/5)^{k-1}} (\log n)^{1+15\epsilon} \quad \text{for } k \le K-1 \\
\bar{\Delta}^{(k)} &= n^{-(1/4)(1-(3/5)^{k-1})} \\
\bar{\Delta}_z^{(k)} &= n^{-(1/4)(1-(3/5)^{k-1})} (\log n)^{-2\epsilon} \\
N^{(k)} &= n^{(1/10)(3/5)^{k-1}} (\log n)^{3\epsilon} \\
N_z^{(k)} &= n^{(1/10)(3/5)^{k-1}} (\log n)^{\epsilon} \\
K &= \min\left\{k : (\bar{\Delta}^{(k)})^2 \le n^{-1/2} (\log n)^{2+16\epsilon}\right\}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\delta_m^{(k)} &\le \bar{\Delta}^{(k)} / N^{(k)} = n^{-(1/4)(1-(3/5)^{k})} (\log n)^{-3\epsilon} \\
\delta_z^{(k)} &\le \bar{\Delta}_z^{(k)} / N_z^{(k)} = n^{-(1/4)(1-(3/5)^{k})} (\log n)^{-3\epsilon}
\end{aligned}$$

The choice of parameters guarantees that $p_m^* \in [\underline{p}_m^{(k)}, \overline{p}_m^{(k)}]$ and $z^* \in [\underline{z}^{(k)}, \overline{z}^{(k)}]$ occur with high probability. Moreover, at the beginning of phase $K$, the precision of $[\underline{p}_m^{(K)}, \overline{p}_m^{(K)}]$ and $[\underline{z}^{(K)}, \overline{z}^{(K)}]$ is $\bar{\Delta}^{(K)} \sim \bar{\Delta}_z^{(K)} \sim n^{-1/4}$.

Next we point out the connections to the algorithm in Wang et al. (2014). When the capacity is sufficient, then this algorithm is closely related to that in Wang et al. (2014): both target the precision $n^{-1/4}$ of the interval estimators for $p_m^*$, and our problem becomes $M$ independent learning problems in Wang et al. (2014). This is why the choices of $\tau^{(k)}$, $N^{(k)}$, $\bar{\Delta}^{(k)}$ and $K$ are almost identical to that of Wang et al. (2014) except for logarithmic terms. The design and analysis diverge for insufficient capacity. In order to track the dual variable, we construct interval estimators for $z$, which is not needed in Wang et al. (2014). In the last phase, based on the estimation of the dual variable, we solve for the optimal prices of all the types simultaneously. This is how we overcome the curse of dimensionality. Note that the dimensionality issue doesn't arise in Wang et al. (2014).

## 4. Analysis

In this section, we analyze the regret of the primal-dual learning algorithm. To simplify the notation, we sometimes resort to a less rigorous expression, such as $\mathrm{P}(A) = 1 - O(n^{-1})$; its equivalence to $\mathrm{P}(A^c) = O(n^{-1})$ should be clear in the context.

Before proceeding, we introduce a modified stochastic system. Technically, if the inventory is depleted at $t$, then $P_m(t)$ must be switched to $p_\infty$, a choke price at which the demand of type-$m$ consumers is turned off, for all $m$. We use a similar simplification to Lei et al. (2017) and consider a slightly different problem. When the inventory is depleted, instead of forced to set $p_\infty$ for all types of consumers, the firm can still use prices between $[\underline{p}, \overline{p}]$. To accommodate the extra demand, it outsources the extra demand at a unit cost $\overline{p}$. Denote the expected revenue of this modified system by $\tilde{J}_n^\pi$. Because $\overline{p}$ is

higher than the price charged, we must have $\tilde{J}_n^\pi \le J_n^\pi$. To bound $J_n^D - J_n^\pi$, it suffices to bound $J_n^D - \tilde{J}_n^\pi$. Therefore, from now on, we investigate the pricing policy $P_m(t)$ associated with the algorithm without switching to $p_\infty$ once the inventory is depleted.

*Remark* 3. The benefit of studying $\tilde{J}_n^\pi$ instead of $J_n^\pi$ is that the pricing policy $\pi$ can be implemented for $t \in [0, T]$ without having to switch to $p_\infty$ at the stopping time at which the inventory is depleted. This simplifies the analysis.

We first show that the number of phases is growing slowly in $n$.

**Lemma 1.** *For $n \ge 3$, the total number of phases $K \le 3 \log n + 3$.*

We next show that the last phase takes the majority of the season.

**Lemma 2.** *The total length of phases prior to phase $K$, $\sum_{k=1}^{K-1} \tau^{(k)}$ is less than or equal to $T/2$ for $n \ge \exp((8/T)^{1/\epsilon})$.*

Consider the following events which are measurable with respect to $\mathcal{F}_{t_k}$:

$$A_k = \cap_{m=1}^M \left\{ p_m^* \in [\underline{p}_m^{(k)}, \overline{p}_m^{(k)}] \right\}$$
$$B_k = \left\{ z^* \in [\underline{z}^{(k)}, \overline{z}^{(k)}] \right\}$$
$$C_k = \cap_{m=1}^M \left\{ \mathcal{P}_m(z) \in [\underline{p}_m^{(k)}, \overline{p}_m^{(k)}] \ \forall z \in [\underline{z}^{(k)}, \overline{z}^{(k)}] \right\}$$

By design, $A_k$ and $B_k$ are the key to the success of the algorithm. If in some phase $k$, the interval estimators $[\underline{p}_m^{(k)}, \overline{p}_m^{(k)}]$ and $[\underline{z}^{(k)}, \overline{z}^{(k)}]$ do not contain $p_m^*$ and $z^*$, then (8) and (9) do not make sense any more. To make things worse, the optimal primal/dual pair $(p_m^*, z^*)$ cannot be recovered in subsequent phases and the learning policy is doomed to fail. Therefore, we want to show that $A_k \cap B_k$ occurs with high probability. The event $C_k$ is also crucial. Note that to estimate $z^*$, the algorithm solves a discrete and empirical version of the dual function, i.e., (8). If $\mathcal{P}_m(z)$ does not fall into the interval estimator for some $z$, then $\max_{j_m=0,\dots,N^{(k)}} \hat{d}_{m,j_m}^{(k)}(p_{m,j_m}^{(k)} - z)$ in (8) may provide a negatively biased estimator for $\mathcal{R}_m(z)$. As a result, the minimization in (8) may not find the correct value.

From the definitions, it is easy to see that $A_k \supseteq B_k \cap C_k$. The following two lemmas show that $B_k$ and $C_k$ occur with high probability.

**Lemma 3.** *For $k = 1, \dots, K-1$, conditional on $B_k \cap C_k$, $\mathsf{P}(B_{k+1} | B_k \cap C_k) = 1 - O(1/n)$.*

**Lemma 4.** *Conditional on $B_k \cap C_k \cap B_{k+1}$, $\mathsf{P}(C_{k+1} | B_k \cap C_k \cap B_{k+1}) = 1 - O(1/n)$.*

Combining Lemma 3 and Lemma 4, we have the following lemma, which states that at the beginning of phase $K$, the probability that the algorithm "goes wrong" is negligible.

**Lemma 5.**

$$\mathsf{P}\left( \cap_{k=1}^K \{A_k \cap B_k \cap C_k\} \right) = 1 - O((\log n)^2 n^{-1}).$$

The following two lemmas characterize the cumulative sales $S(t)$ at $t = t_K$, the beginning of phase $K$. Recall that the cumulative sales $S(t_K)$ can be expressed as $\int_0^{t_K} \sum_{m=1}^M dN_{m,t}(n d_m(P_m(t)))$. Therefore, $\mathsf{E}[S(t_K)] = n\mathsf{E}[\int_0^{t_K} \sum_{m=1}^M d_m(P_m(t))dt]$. Also note that in the fluid model, the inventory level at $t_K$ is $n \sum_{k=1}^{K-1} \tau^{(k)} \sum_{m=1}^M d_m(p_m^*)$. Therefore, the next two lemma state that the cumulative sales process does not deviate too much from that in the fluid system at the beginning of phase $K$.

**Lemma 6.** *At the beginning of phase $K$, the conditional expectation of $S(t_K)$ given the pricing policy $P_m(t)$ satisfies*

$$\mathsf{P}\left(\left|\int_0^{t_K}\sum_{m=1}^{M} d_m(P_m(t))dt - \sum_{k=1}^{K-1}\tau^{(k)}\sum_{m=1}^{M} d_m(p_m^*)\right| > n^{-1/4}(\log n)^{1+8\epsilon}\right) = O((\log n)^2 n^{-1}).$$

**Lemma 7.** *At the beginning of phase $K$, $S(t_K)$ satisfies*

$$\mathsf{P}\left(\left|S(t_K) - n\sum_{k=1}^{K-1}\tau^{(k)}\sum_{m=1}^{M} d_m(p_m^*)\right| > 2n^{3/4}(\log n)^{1+8\epsilon}\right) = O((\log n)^{-2} n^{-1/2}).$$

Roughly speaking (ignoring the logarithmic terms), Lemma 6 and 7 show that the inventory level at the beginning of phase $K$ misses the target inventory level in the fluid system by $n^{3/4}$. This is consistent with the precision of the price interval at $t_K$, which satisfies $\bar{\Delta}^{(K)} \sim n^{-1/4}$.

### 4.1. Sufficient Capacity

We next bound the regret when $z^* = 0$, i.e., when the capacity is not constrained.

**Proposition 2.** *When $z^* \le 0$, we have $J_n^D - \mathsf{E}[\tilde{J}_n^\pi] = O((\log n)^{2+16\epsilon} n^{1/2})$. Therefore,*

$$R_n^\pi(T, c) = O((\log n)^{2+16\epsilon} n^{-1/2}).$$

The major steps of the proof are sketched below. We first express $\mathsf{E}[\tilde{J}_n^\pi]$ as

$$n\sum_{k=1}^{K}\mathsf{E}\left[\int_{t_k}^{t_{k+1}} P_m(t) d_m(P_m(t))dt\right] - \overline{p}\mathsf{E}\left[\left(\sum_{m=1}^{M}\int_0^T dN_{m,t}(n d_m(P_m(t))) - nc\right)^+\right]. \tag{15}$$

The first term is the expected revenue generated in each phase, and the second term accounts for the outsourcing cost explained in Remark 3. Moreover, $J_n^D$ can be expressed as $\sum_{k=1}^{K}\mathsf{E}\left[\int_{t_k}^{t_{k+1}} p_m^* d_m(p_m^*)dt\right]$. Thus, the difference between $J_n^D$ and the first term of (15) can thus be bounded by

$$\begin{aligned}
&n\sum_{k=1}^{K}\mathsf{E}\left[\int_{t_k}^{t_{k+1}}\left(p_m^* d_m(p_m^*) - P_m(t)d_m(P_m(t))\right)dt\right] \\
\sim& n\sum_{k=1}^{K}\mathsf{E}\left[\int_{t_k}^{t_{k+1}}\left(d_m(p_m^*) - d_m(P_m(t))\right)^2 dt\right] \sim n\sum_{k=1}^{K}\mathsf{E}\left[\int_{t_k}^{t_{k+1}}\left(p_m^* - P_m(t)\right)^2 dt\right] \\
\sim& n\sum_{k=1}^{K}\tau^{(k)}(\bar{\Delta}^{(K)})^2.
\end{aligned} \tag{16}$$

Because $p_m^*$ is the unconstrained maximizer of $p d_m(p)$, we can apply a quadratic bound in the second line by Assumption 1. The last line follows from the high-probability event $A_k$, which implies that $|P_m(t) - p_m^*| \le \bar{\Delta}^{(k)}$ in phase $k \le K-1$. At the beginning of phase $K$, we can show that Step 21 is triggered with high probability. In this case, $|P_m(t) - p_m^*| = |p_m^{(K)} - p_m^*| = |\overline{p}_m^{(K)} - p_m^* + \alpha| \le \bar{\Delta}^{(K)} + \alpha$ is not necessarily bounded by $\bar{\Delta}^{(K)}$. However, $\alpha$ is chosen carefully to match the order of $\bar{\Delta}^{(K)}$. By the choice of the parameters, the order of $n\sum_{k=1}^{K}\tau^{(k)}(\bar{\Delta}^{(K)})^2$ is $O((\log n)^{2+16\epsilon} n^{1/2})$.

To bound the second term of (15), we show that the mean of the total sales over the horizon, $n\mathsf{E}\left[\sum_{m=1}^{M}\int_0^T d_m(P_m(t))dt\right]$, does not exceed $nc$. This is achieved by charging a markup $p_m^{(K)} = \overline{p}_m^{(K)} + \alpha$ in phase $K$ and this is exactly the purpose of introducing $\alpha$. Since the random fluctuation of Poisson arrivals is bounded by $O(n^{1/2})$, we obtain the regret stated in Proposition 2.

### 4.2. Insufficient Capacity

When $z^* > 0$, the inventory is depleted at $T$ in the fluid system. We first show that the cumulative sales at the end of the horizon under the algorithm misses the target $nc$ by $O(n^{1/2})$.

**Lemma 8.** *When $z^* > 0$, we have*

$$\mathsf{E}[|S(T) - nc|] = O((\log n)^{\epsilon} n^{1/2}) \tag{17}$$

*and*

$$\mathsf{E}\left[\left|\int_0^T \sum_{m=1}^{M} d_m(P_m(t))dt - c\right|\right] = O((\log n)^{\epsilon} n^{-1/2}). \tag{18}$$

The intuition behind the proof is explained below. We first show that $D_l^{(K)}$ and $D_u^{(K)}$ estimate $n\sum_{m=1}^{M} d_m(p_m^l)$ and $n\sum_{m=1}^{M} d_m(p_m^u)$ with precision $n^{1/2}$ (ignoring the logarithmic terms), because they are Poisson random variables with means of order $n$ and thus the standard deviations are $O(n^{1/2})$. With such precision, $\theta$ (Step 35) approximately solves

$$n(T - t_K)\sum_{m=1}^{M}(\theta d_m(p_m^l) + (1-\theta)d_m(p_m^u)) \approx nc - S(t_K).$$

Note that regardless of the first $K-1$ phases, the remaining inventory is $nc - S(t_K)$ at $t_K$. Using $p_m^l$ ($p_m^u$) for a fraction $\theta$ ($1-\theta$) of phase $K$ serves as a corrective force to ensure the aggregate sales over the horizon to be close to $nc$ (the error bound $(\log n)^{\epsilon} n^{1/2}$ is caused by the random fluctuation of the Poisson arrivals).

Algorithm 1 does not explore the price space in phase $K$, and thus the precision of $p_m^l$ and $p_m^u$ in Step 26 is of order $n^{-1/4}$. Therefore, one would expect the sales to type-$m$ consumers miss the target $nTd_m(p_m^*)$ by $n^{-1/4} \times n = n^{3/4}$. This is indeed the case. However, Lemma 8 guarantees that a simple exploration (Step 28, 32 and 35) is effective and leads to higher precision ($O(n^{1/2})$ for the aggregate sales of all types of consumers. This is crucial to the proof of the next proposition.

**Proposition 3.** *When $z^* > 0$, we have $J_n^D - \mathsf{E}[\tilde{J}_n^{\pi}] = O((\log n)^{2+18\epsilon} n^{1/2})$. Therefore,*

$$R_n^{\pi}(T, c) = O((\log n)^{2+18\epsilon} n^{-1/2}).$$

Different from Proposition 2, $p_m^*$ is not the unconstrained maximizer of $p_m^* d_m(p_m^*)$ when $z^* > 0$.

Therefore, if we follow (16), the difference is approximately

$$\begin{aligned}
&n\sum_{k=1}^{K}\mathsf{E}\left[\int_{t_k}^{t_{k+1}}\left(p_m^* d_m(p_m^*)-P_m(t)d_m(P_m(t))\right)dt\right]\\
\sim& n\sum_{k=1}^{K}\mathsf{E}\left[\int_{t_k}^{t_{k+1}}\left|d_m(p_m^*)-d_m(P_m(t))\right|dt\right]\sim n\sum_{k=1}^{K}\tau^{(k)}\bar{\Delta}^{(K)}\sim n^{3/4},
\end{aligned}\tag{19}$$

which clearly does not meet our target $n^{1/2}$. The remedy to this situation is the following key observation. Let $r_m^{*\prime}$ and $r_m^{*\prime\prime}$ be the first- and second-order derivative of $r_m(\lambda)=\lambda d_m^{-1}(\lambda)$ at $\lambda=d_m(p_m^*)$. By Taylor's expansion, the difference in revenue rate can be expressed as

$$\begin{aligned}
&p_m^* d_m(p_m^*)-P_m(t)d_m(P_m(t))\\
\le& r_m^{*\prime}(d_m(p_m^*)-d_m(P_m(t)))+|r_m^{*\prime\prime}|(d_m(p_m^*)-d_m(P_m(t)))^2.
\end{aligned}$$

Because $\lambda=d_m(p_m^*)$ maximizes $r_m(\lambda)-\lambda z^*$ by the primal-dual formulation (Proposition 1), the first-order condition implies $r_1^{*\prime}=r_2^{*\prime}=\ldots=r_M^{*\prime}=z^*$. Therefore, an improved bound for (19) can be derived:

$$\begin{aligned}
&n\sum_{k=1}^{K}\mathsf{E}\left[\int_{t_k}^{t_{k+1}}\left(p_m^* d_m(p_m^*)-P_m(t)d_m(P_m(t))\right)dt\right]\\
\sim& nz^*\mathsf{E}\left[\int_0^T\sum_{m=1}^{M}(d_m(p_m^*)-d_m(P_m(t)))dt\right]+n\mathsf{E}\left[\int_0^T\sum_{m=1}^{M}|r_m^{*\prime\prime}|(d_m(p_m^*)-d_m(P_m(t)))^2dt\right]\\
\sim& nz^*\mathsf{E}\left[c-\int_0^T\sum_{m=1}^{M}d_m(P_m(t))dt\right]+M_4 n\mathsf{E}\left[\int_0^T\sum_{m=1}^{M}(d_m(p_m^*)-d_m(P_m(t)))^2dt\right].
\end{aligned}$$

The first term can be bounded by $(\log n)^{\epsilon}n^{1/2}$ by Lemma 8; this is the reason why we need to bound the aggregate sales. The second term is of the same order as $n\sum_{k=1}^{K}\tau^{(k)}(\bar{\Delta}^{(K)})^2$ with high probability, which has been shown to meet the target in the remarks following Proposition 2.

Combining Proposition 2 and Proposition 3 and recalling that $\epsilon$ can be an arbitrarily small constant, we obtain the main theorem.

**Theorem 1.** *Suppose Assumptions 1 and 2 hold. For any $\delta>0$, we can select $\epsilon$ such that the regret of the primal-dual learning algorithm satisfies*

$$R_n^{\pi}(T,c)\le C(\log n)^{2+\delta}n^{-1/2},$$

*where the constant $C$ is independent of $n$.*

By Besbes and Zeevi (2009), no learning policy can achieve regret that grows slower than $n^{-1/2}$ with $M=1$. Therefore, ignoring the logarithmic terms, the primal-dual learning algorithm achieves near-optimal regret.

# 5. Discussion

In this section, we discuss several salient features of the primal-dual learning algorithm and clarify our findings in comparison to the literature.

## 5.1. The Fundamental Limit of Online Learning

This paper sets out on a quest to identify the fundamental limit of online learning (the optimal rate of regret) in the multi-product dynamic pricing problem over a network with a single resource. The first benchmark is provided in Slivkins (2014): for a general Lipschitz-continuous objective function with a $d$-dimensional decision vector without resource constraints, the optimal rate of regret is $n^{-1/(2+d)}$. Recently, Chen and Gallego (2019) show that when the objective function is locally concave, then the optimal rate is slightly better, and may be adjusted to $n^{-2/(3+d)}$. The number of decisions $d$ is the number of products in network dynamic pricing, or $M$ in this paper. Note that setting $d = 1$ recovers the rate $n^{-1/2}$ in dynamic pricing (one product with one constraint) (Besbes and Zeevi, 2009; Wang et al., 2014). This is not good news: as the number of products $d$ increases, the rate deteriorates dramatically, which is referred to as the curse of dimensionality. It is also consistent with the finding in Besbes and Zeevi (2012). The quest to understand whether online learning is complicated by dimensionality in multi-product dynamic pricing problem is for continuous decision variables as is the case in dynamic pricing. The rate of regret from papers assuming discrete decisions (Badanidiyuru et al., 2013; Agrawal and Devanur, 2014; Ferreira et al., 2018) does not apply to our setting.

Some recent papers have shed new light on this fundamental problem. Chen and Shi (2019) design a learning algorithm for network dynamic pricing that achieves rate $n^{-1/5}$ regardless of the number of products or constraints. Although the rate doesn't seem to be optimal, it does not depend on the dimension of the problem (the number of products or the number of resources). One may then question the fundamental difference between a general learning problem (Slivkins, 2014) and network dynamic pricing what property makes the latter easier to learn? Li et al. (2019) and references there in reveal that it may be due to the intrinsic concavity structure of network dynamic pricing. In particular, Li et al. (2019) show that in the setting of Slivkins (2014), if the objective function is concave, then the regret is dimension-free ($n^{-1/2}$), because one can resort to gradient-based algorithms. In network dynamic pricing, Assumption 1 is commonly adopted to guarantee good behavior of the optimal policy. Since the objective function is concave after one transforms the decision variable from price to quantity, it somehow inherits the dimension-free nature.

This opens up a new research question: can the $n^{-1/2}$ rate be obtained for learning in network dynamic pricing? This paper takes a step forward to a positive answer of this question as we can confirm the rate $n^{-1/2}$ under two simplifications: The objective function in our setting can be expressed as the sum of the revenues collected from each of the $M$ types of consumers, and thus *separable* in terms of the decision variables (prices); there is a single resource constraint. The first simplification implies that, when the inventory constraint is not binding, then the optimal rate of regret should be the same as $M$ independent learning problems with one decision variable, i.e., $O(Mn^{-1/2})$. It doesn't lead to dimension-free regret automatically when the constraint is binding, as the $M$-dimensional decision variables can be regarded as a vector on a $(M-1)$-dimensional manifold, after a proper transformation. On the manifold, the objective function is not separable in terms of the transformed decision variables

any more. The second simplification can be relaxed as argued in Section 5.3. Therefore, our result confirms that $n^{-1/2}$ may be achievable. Moreover, we provide the first dimension-free algorithm that is not gradient-based in the context.

## 5.2. Learning in the Primal and Dual Spaces

The algorithm provides point and interval estimators for *both* the primal and dual optimal solutions, $p_m^*$ and $z^*$, in each phase. It turns out that such learning is necessary, as the regret of a policy that only learns the primal space incurs much higher regret (see, e.g., Section 4 of Besbes and Zeevi 2012). This is not surprising given the primal-dual formulation in Section 2.1 since the revenues collected from different types of consumers are only coupled through the inventory constraint. Therefore, having an accurate estimator for the dual optimal solution $z^*$ helps to decouple the problem into $M$ independent learning problems of the form $\max_p d_m(p)(p - z^*)$ for $m = 1, \ldots, M$. These independent learning problems are known to have regret $O(n^{-1/2})$ (the parametric version of such problems is solved in Keskin and Zeevi 2014; den Boer and Zwart 2014). However, $z^*$ is not given upfront and has to be learned. The key design of the algorithm is to nest the learning processes in the primal and dual spaces to narrow down the primal/dual optimal solutions sequentially.

## 5.3. Multiple Resource Constraints

The motivation of the study is personalized pricing, which can be recast as a multi-product dynamic pricing problem with a single constrained resource as mentioned in the introduction. We believe our algorithm may be applied to a generic multi-product dynamic pricing problemwith multiple resource constraints, as long as the demand is separable, as mentioned in Section 5.1. Next we briefly introduce the extension to multiple resources.

Suppose there are $L$ resources, with initial capacity $\boldsymbol{c} = (c_1, \ldots, c_L)$, and product $m$ consumes $a_{ml}$ units of resource $l$. The Lagrangian (3) can be reformulated as

$$L(\boldsymbol{p}(t), \boldsymbol{z}) = \sum_{l=1}^{L} c_l z_l + \sum_{m=1}^{M} \int_{t=0}^{T} (p_m(t) - \sum_{l=1}^{L} a_{ml} z_l) d_m(p_m(t)) dt.$$

For a fixed $\boldsymbol{z}$ we define

$$\mathcal{R}_m(\boldsymbol{z}) \triangleq \max_p \left\{ d_m(p)(p - \sum_{l=1}^{L} a_{ml} z_l) \right\}$$
$$\mathcal{P}_m(\boldsymbol{z}) \triangleq \operatorname{argmax}_p \left\{ d_m(p)(p - \sum_{l=1}^{L} a_{ml} z_l) \right\}.$$

and the dual function

$$g(\boldsymbol{z}) = \max_{\boldsymbol{p}(t)} \{L(\boldsymbol{p}(t), \boldsymbol{z})\}.$$

As for the implementation of the algorithm, the exploration of the first $K-1$ phases is essentially the same (Step 5 to Step 14). When finding the point estimator for $\boldsymbol{z}^{(k)*}$, Equation (8) is vectorized, and the optimal vector $\boldsymbol{z}^{(k)*}$ is obtained by testing the grid in the confidence set of $\boldsymbol{z}$, which is a hyper-rectangle since each entry of $\boldsymbol{z}$ has a confidence interval. Normally the grid size explodes exponentially in the

size of $\boldsymbol{z}$, i.e., $L$, and this is precisely causing the curse of dimensionality. However, in this algorithm (8) is solved offline, i.e., the computation is not counted toward the final regret. In the beginning of the last phase, the dual/primal variables may have been estimated with error $n^{-1/4}$, similar to (11), under properly chosen parameters. For products with sufficient capacity (a product $m$ enjoys sufficient capacity if $0 \in [\underline{z}_l^{(K)}, \bar{z}_l^{(K)}]$ for all $l$ such that $a_{ml} > 0$), this already leads to the optimal regret. For products with insufficient capacity, it is unclear how the linear interpolation (14) can be implemented in the high dimension. We thus leave the algorithm and regret analysis in this case for future research.

### 5.4. Controlling the Aggregate Sales

As explained in Section 3.2 and the remark following Proposition 3, when $z^* > 0$, the interval estimators for $p_m^*$ at the beginning of phase $K$ are still too wide ($n^{-1/4}$). Instead of exploring the price space further and attempting to narrow down the intervals in phase $K$, the algorithm simply controls the aggregate sales within an error margin of $n^{1/2}$ (Step 25 to 40 in Algorithm 1 and Lemma 8). Focusing on a single quantity (the aggregate sales) turns out much easier than controlling $M$ decision variables. In fact, we suspect that no learning policy could estimate all $p_m^*$ with precision $n^{-1/2}$ at the end of the horizon[6].

Surprisingly, controlling the aggregate sales is sufficient to meet the target regret, even though the estimators for the optimal prices are not precise enough. The reason is explained by the remarks following Proposition 3. In particular, the derivatives of the revenue rates $\lambda d_m^{-1}(\lambda)$ with respect to the demand rate $\lambda$ are all equal to $z^*$ at optimality $\lambda = d_m(p_m^*)$ for $m = 1, \ldots, M$. This allows the firm to control the regret by the deviation of the aggregate sales.

### 5.5. Data Reuse

Across different phases, the interval estimators for the optimal price $p_m^*$ may overlap for some $m$. In this case, the demand estimated at certain prices in the previous phases may be reused. For example, if $[\underline{p}_m^{(k)}, \overline{p}_m^{(k)}]$ and $[\underline{p}_m^{(k+1)}, \overline{p}_m^{(k+1)}]$ overlap, then for some $j_1$ and $j_2$ the price grid points $p_{m,j_1}^{(k)}$ and $p_{m,j_2}^{(k+1)}$ may be close, and the demand estimate for $p_{m,j_1}^{(k)}$ can provide useful information for $p_{m,j_2}^{(k+1)}$. In our algorithm, the data from previous phases are not reused mainly for the analysis, because data reuse introduces complex dependence between phases. In practice, we believe that data reuse may facilitate the learning of the demand function and increase the efficiency of the policy.

### 5.6. Discontinuous Demand

In a recent paper, den Boer and Keskin (2020) study discontinuous demand functions, which is by far the most general assumption on smoothness, although their demand function has a parametric form. In our setting, Assumption 1 rules out the possibility of discontinuity, and our algorithm is unlikely to work for discontinuous functions. This is because discontinuous functions cannot be concave, and they break the primal-dual formulation and thus the foundation of the algorithm. We also believe that in the nonparametric setting, no algorithm can achieve sublinear regret when the demand function can be discontinuous. Consider the following example: there is one type of consumer ($M = 1$); we

[6]The case we study is different from Wang et al. (2014), in which the market-clearing price $d^{-1}(c/T)$ can be learned with precision $n^{-1/2}$. In our case, there are $m$ types of consumers and there are still $m - 1$ degrees of freedom when a vector of prices $(p_1, \ldots, p_m)$ are market clearing.

normalize the time horizon ($T = 1$) and there is sufficient capacity ($c = 100$). Let the price range be $p \in [0, 1]$. Consider the demand (revenue) function satisfying

$$pd(p) = \mu_k, \ p \in ((k-1)/K, k/K].$$

Because the capacity is not constrained, the problem is conceptually equivalent to the multi-armed bandit problem with $K$ arms. (The difference of continuous/discrete time is non-consequential in this case.) It is well-known that the minimax lower bound for the regret of such a learning problem is $O(\sqrt{Kn})$. Since $K$ can be arbitrarily large, the regret cannot possibly be controlled.

## References


Agrawal, S. and N. R. Devanur (2014). Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 989–1006. ACM.

Araman, V. F. and R. Caldentey (2009). Dynamic pricing for nonperishable products with demand learning. *Operations research 57*(5), 1169–1188.

Auer, P., R. Ortner, and C. Szepesvári (2007). *Improved Rates for the Stochastic Continuum-Armed Bandit Problem*, pp. 454–468. Berlin, Heidelberg: Springer Berlin Heidelberg.

Badanidiyuru, A., R. Kleinberg, and A. Slivkins (2013). Bandits with knapsacks. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pp. 207–216. IEEE.

Badanidiyuru, A., J. Langford, and A. Slivkins (2014). Resourceful contextual bandits. In *Conference on Learning Theory*, pp. 1109–1134.

Ban, G. and N. B. Keskin (2017). Personalized dynamic pricing with machine learning. *Working paper*.

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

Besbes, O. and A. Zeevi (2012). Blind network revenue management. *Operations research 60*(6), 1537–1550.

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

Bubeck, S. and N. Cesa-Bianchi (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning 5*(1), 1–122.

Bubeck, S., R. Munos, G. Stoltz, and C. Szepesvári (2011). X-armed bandits. *Journal of Machine Learning Research 12*(May), 1655–1695.

Canonne, C. (2017). A short note on poisson tail bounds.

Cesa-Bianchi, N. and G. Lugosi (2006). *Prediction, learning, and games*. Cambridge university press.