Markdown Pricing Under Unknown Parametric Demand Models

Su Jia, Andrew A. Li, R. Ravi
Tepper School of Business, Carnegie Mellon University

Dynamic Pricing under unknown demand has been extensively studied but rarely deployed by real-world sellers, largely due to overlooking some practical constraints. As one example, the prices may oscillate, which is highly undesirable in practice. In many applications, few retailers would pursue extra revenues by increasing the price, which may potentially lead to stirring customer service complaints, social media outrage, or simply loss of customers. Rather, they may start with a high initial price and then offer price *markdowns*, usually in a conservative manner.

Thus motivated, we consider a single-product revenue management problem where, given infinite inventory, the objective is find the markdown policy, i.e. a pricing policy that monotonically reduces the price, over a finite sales horizon to maximize expected revenues. Recently, Jia et al. (2021) considered the problem where the underlying demand function is unknown, and showed a tight $T^{3/4}$ regret bound under *minimal* assumptions, that is, unimodality and Lipschitzness in the revenue function.

However, in practice the demand functions are usually assumed to have certain functional forms. We investigate two fundamental questions, assuming the underlying demand curve comes from a given parametric family: (1) Can we improve the regret bounds under this extra assumption? (2) Is markdown pricing still *harder* than unconstrained pricing? We introduce a concept called *dimension* that measures the complexity of performing markdown pricing on a family, and present tight regret bounds for each regime under this framework, thereby completely settling the problem.

Key words: dynamic pricing, markdown pricing, multi-armed bandits, revenue management, conservative decision-making

1. Introduction

Dynamic pricing under unknown demand has been extensively studied. Such problem arise naturally for new products, or for old products in new markets. Such problems are usually formulated as a Multi-Armed Bandit model. While bandit problems have been well-understood *theoretically*, in practice however, we rarely see retailers deploy such policies. This is a largely because some practical constraint are often overlooked by those policies. For example, the prices may oscillate, which is highly undesirable. As observed in Bitran and Mondschein (1997),

"Customers will hardly be willing to buy a product whose price oscillates...Most retail stores do not increase the price of a seasonal or perishable product despite the fact that the product is being sold successfully".

It is also pointed out in Harvard Business Review (Dholakia (2021)) that

"Communicating a price increase to customers is never a pleasant task. It has the potential to stir customer service complaints, social media outrage, or simply lose customers altogether."

Thus, in many applications retailers implicitly faces a natural monotonicity constraint that the prices can not go up. A pricing policy that satisfies such a constraint is usually referred to as *markdown pricing* policies.

Markdown pricing is ubiquitous in retailing. In practice, a retailer may use price markdowns to boost demands and hence increase revenues. For example, for fashion clothing, a retailer may start with a high retail price in the regular selling season, and then offer discounts in the clearance season, possibly for multiple times, to sell the remaining inventory.

A successful markdown pricing strategy can have a considerable impact on the gross margins. According to the retail think tank Coresight, "In 2018, price markdowns cost retailers $300 billion in the US alone, which accounts for 12% of total retail sales." A recent survey also (Google (2021)) suggests that up to $39 billion in value is being left on the table due to sub-optimal markdown pricing, and this number is just for one of many sectors of retail ("specialty" retail).

Thus motivated, in this work we consider the markdown pricing problem with unknown demand, under various assumptions. While *unconstrained* dynamic pricing under unknown demand has been extensively studied, little is known about *markdown* pricing under unknown demand. Recently, Jia et al. (2021) first considered this problem, under the most general case with only *minimal* assumptions for achieving meaningful performance guarantees. More precisely, they showed that unimodality and Lipschitzness in the *revenue* function (defined to be the price times mean demand) are necessary for attaining sublinear regret, and presented a tight $T^{3/4}$ regret bound under those assumptions. Noticeably, this bound is asymptotically higher than $T^{2/3}$, the known regret bound for *unconstrained* pricing, highlighting the extra complexity caused by the monotonicity constraint.

Nonetheless, in practice, demand functions are usually assumed to have certain parametric forms, such as linear, exponential or logit function. This motivates our first question:

"Q1) Can we strengthen the $T^{3/4}$ regret bound for markdown pricing in this setting?"

To see why such improvement is possible, observe that the proof of the $T^{3/4}$ lower bound considers pairs of "roof-shaped" revenue functions that are *completely* identical when the price is higher than some p, and diverging for prices lower than p. Thus, any reasonable policy has to carefully reduce the price, halting only when there is sufficient evidence for *overshooting* the optimal price.

However, this is not true in the parametric case. Take linear demand functions as an example. A policy may simply learn the slope and intercept of the underlying demand function at high prices, and then select the optimal price of the estimated demand function in all future rounds. This enables us to design a more powerful class of learn-then-earn type of policies.

Now that we believe the $T^{3/4}$ regret can be improved under certain parametric assumption, we naturally arrive at our second question:

"Q2) Is markdown pricing still harder than unconstrained pricing under these assumptions?"

Or more precisely, can we still show a *separation* between markdown and unconstrained pricing, under various parametric assumptions?

While one may answer these two questions for particular families such as linear family, as the real challenge, can we find a general framework to unify the regret bounds for different *categories* of families, rather than specific results for specific families? In this work, we do propose such a framework, by introducing a complexity index called *dimension*, that captures the hardness (or easiness) of performing markdown pricing on a family. Under this framework, we provide simple, efficient markdown policies for each dimension, which we also show to be *best* possible, thereby completely settling the problem of markdown pricing under unknown demand.

1.1. Our Contributions.

In this work, we make the following contributions. We introduced a new concept called *dimension*, that captures the complexity of performing markdown pricing on a family. Within this frame, we provide a *complete* settlement of the markdown pricing problem with unknown demand, as specified below.

1. **Markdown Policies with Theoretical Guarantees:** For each finite $d \geq 0$, we present a efficient markdown pricing policy. Our policies that proceeds in *phases*, wherein the seller learns the demand by selecting prices at suitable spacing, and estimates the true parameter and then makes conservative decisions. We show that for $d = 0$ and $d \geq 1$ our policies achieve regret $O(\log^2 T)$ and $\tilde{O}(T^{\frac{d}{d+1}})$ respectively, settling our first research question.

2. **Tight Minimax Lower Bound:** We complement our upper bounds with a matching lower bound for each of the three regimes $d = 0$, $1 \leq d < \infty$ and $d = \infty$. All of our lower bounds are established using the same tool, the generalized Wald-Wolfowitz Theorem (Theorem 10). Furthermore, our tight bounds are asymptoticly higher than that of unconstrained pricing in all regimes but $d = 1$, settling our second research question.

3. **Impact of Smoothness:** Further, we investigate the impact of smoothness of the revenue function around the optimal price, by extending our upper bounds for general sensitivity parameters $s \geq 2$. For both finite and infinite d, we obtained decreasing upper bounds as s increases from 2. Moreover for $d = \infty$, our tight $T^{\frac{2s+1}{3s+1}}$ regret bound is asymptoticly higher than that of unconstrained pricing, whose optimal regret is known to be $T^{\frac{s+1}{2s+1}}$.

The remainder of this paper is organized as follows: we conclude this section with a summary of the related literature. We then formally describe our model and assumptions in Section 2, and then state our policies and main results in Section 3. Next we discuss the three regimes $d = 0$, $1 \leq d < \infty$ and $d = \infty$ respectively in Sections 4,5 and 6 and prove the tight regret bounds.

1.2. Previous Work

The present work falls into two primary streams of work: dynamic pricing and multi-armed bandits. As mentioned above, the distinguishing feature of our work is the combination of a markdown constraint with a bandit-style (i.e. minimizing regret) analysis. Other important dimensions along which to contrast this work with the extant literature include: whether the underlying demand function is assumed to come from a parametric family (this work is non-parametric), whether infinite inventory is assumed (this work allows for a particular regime of finite inventory), and whether it is assumed that a prior distribution for the demand functions is given (this work does not).

Dynamic Pricing: Gallego and Van Ryzin (1994) characterized the optimal pricing policy when the demand function is known. Kleinberg and Leighton (2003) studied a revenue maximization problem for a seller with an unlimited supply of identical goods, and obtained tight regret bounds under different models of buyers. Besbes and Zeevi (2009) studied the dynamic pricing problem under finite inventory in a finite selling period. Their benchmark regret function is the optimal pricing algorithm which is non-adaptive and whose expected sales is at most the inventory level. They presented an algorithm which achieves nearly optimal regret bounds. Subsequently, Wang et al. (2014) improved their results by showing matching lower bound. Later, Babaioff et al. (2015)

and Badanidiyuru et al. (2013) considered a more practical scenario where the inventory is finite. Other works that formulate dynamic pricing as MAB include Bastani et al. (2019), Hu et al. (2016), Chen and Farias (2018), Lei et al. (2014), Keskin and Zeevi (2014), den Boer and Zwart (2013), Liu and Cooper (2015), Farias and Van Roy (2010), Lobel (2020), Qiang and Bayati (2016), Papanastasiou and Savva (2017), den Boer and Zwart (2015).

In practice, costs of implementing frequent price changes in a traditional retail setting can amount to a considerable portion of the seller's net margins. Thus motivated, Broder (2011) first formulated the demand learning problem with limited price changes and presented an $O(\sqrt{T})$ regret policy for parametric models using $O(\log T)$ price changes. Later, Perakis and Singhvi (2019) showed under stronger assumptions that the same regret may be achieved using $O(\log\log T)$ price-changes. Cheung et al. (2017) considered a given discrete demand functions and presented a regret bound that decreases in the number of allowed price-changes. Chen et al. (2020) considered the joint pricing and inventory management problem under limited price changes.

Orthogonal to the number of price changes, previous literature has also considered the *direction* of price changes. In practice, buyers usually have a *reference price* in mind, at which a higher (lower) price is considered a loss (gain), and customers are more sensitive to losses than to gains. Dynamic pricing with reference-price effects has been studied extensively in recent years, for example Nasiry and Popescu (2011), Heidhues and Kőszegi (2014), Wu et al. (2015), Hu et al. (2016), Wang (2016), Recently, den Boer and Keskin (2020) considered the setting where the demand function is unknown.

As an important variant of the dynamic pricing problem, the *Markdown Pricing* problem has been extensively studied. Book chapter by Ramakrishnan (2012) and surveys by Elmaghraby and Keskinocak (2003) and den Boer and Zwart (2015) provide a through overview. Most previous work on markdown pricing assume known demand function and focused on either empirical results (e.g. Smith and Achabal (1998), Heching et al. (2002)) or strategic customer behaviors (e.g. Yin et al. (2009), Boyacı and Özer (2010), Aviv and Vulcano (2012)). Recently, Jia et al. (2021) first considered the markdown pricing under unknown demand function, and showed a $\tilde{\Theta}(T^{3/4})$ regret bound assuming the unknown revenue functions are Lipschitz and Unimodal.

Multi-armed Bandits (MAB): There exist several MAB variants that are similar to our problem, but without the markdown constraint. In the *Discrete Multi-armed Bandit* problem, the player is offered a finite set of arms, with each arm providing a random revenue from an unknown

probability distribution specific to that arm. The objective of the player is to maximize the total revenue earned by pulling a sequence of arms (e.g. Lai and Robbins (1985)). Our pricing problem generalizes this framework by using an infinite action space $[0,1]$ with each price p corresponding to an action whose revenue is drawn from an unknown distribution with mean $R(p)$.

In the *Lipschitz Bandit* problem (e.g. Agrawal (1995)), it is assumed that each $x \in [0,1]$ corresponds to an arm with mean reward $\mu(x)$, and μ satisfies the Lipschitz condition, i.e. $|\mu(x)-\mu(y)| \leq L|x-y|$ for some constant $L>0$. Kleinberg (2005) proved a tight $\tilde{\Theta}(T^{2/3})$ regret bound for one-dimensional Lipschitz Bandits. The lower bound was proved by considering a family of "bump curves": each curve is $\frac{1}{2}$ at all arms except in a small neighborhood of the "peak", where the mean reward is slightly higher elevated. Since these bump curves are unimodal, this lower bound carries over to the family we study.

Another closely-related variant of MAB is the *Unimodal Bandits* problem (Cope (2009), Yu and Mannor (2011), Combes and Proutiere (2014)). In addition to the Lipschitzness assumption, the reward function $\mu : [0,1] \to [0,1]$ is assumed to be unimodal. It is also assumed that there is a constant $L'>0$ s.t. $|\mu(x)-\mu(y)| \geq L'|x-y|$ for all $x,y \in [0,1]$. Yu and Mannor (2011) proposed a binary-search type algorithm with regret $\tilde{O}(\sqrt{T})$.

Recently there is an emerging line of work on online learning with monotonicity constraint. Jia et al. (2021) considered the markdown pricing problem under unknown constraints, but only for nonparametric case. More precisely, they assumed the revenue function is unimodal and Lipschitz, and showed a tight $\tilde{\Theta}(T^{3/4})$ regret bound for finite and infinite setting. Chen (2021) independently considered a special case where the inventory is infinite under the name *Monotone Bandits*, and obtained the same results using different lower bound techniques. Gupta and Kamble (2019),Salem et al. (2021) considered a more general online convex optimization problem where the actions sequence is required to be monotone.

2. Model and Assumptions

We begin by formally stating our model. In this work we assume an unlimited supply of a single product. Given a discrete time horizon of T rounds, in each round t, the policy (representing the "seller"), selects a price x_t (the particular interval $[0,1]$ is without loss of generality, by scaling). This round's demand is then independently drawn from a fixed distribution with unknown mean $D(x_t)$, and the policy receives reward x_t for each unit sold. The only constraint the policy must satisfy is the *markdown* constraint: $x_1 \geq \cdots \geq x_T$ almost surely.

The function $D(x)$ which maps each price x to the mean demand at that price is known as the *demand function*, which is. naturally associated with a *revenue function*, $R(x) = x \cdot D(x)$. Sometimes we will deal directly with revenue functions, which we term more generally as *reward* functions.[1] For any policy π,[2] reward function $R(\cdot)$, we use $r(\pi, D)$ to denote the expected total reward of π under D.

Rather than evaluating policies directly in terms of $r(\pi, D)$, it is more informative (and ubiquitous in the literature on multi-armed bandits) to measure performance using the notion of *regret* with respect to a certain idealized benchmark. Here, we will define regret with respect to the best possible *fixed* price policy. Specifically, when the true reward function is known, then seller simply always selects a revenue-maximizing price $x_D^* = \arg\max_{x\in[0,1]} x \cdot D(x)$ at each round, and we denote $r^* = \max_{x\in[0,1]} x \cdot D(x)$. The regret of a policy is then defined with respect to this quantity, and we seek to bound the *worst-case* value over a given family of reward functions.

Definition 1 (Regret). For any policy π and demand function D, define the *regret* of policy π under D to be $\mathrm{Reg}(\pi, D) := r_D^* T - r(\pi, D)$. For any given family $\mathcal{F}$ of demand functions, the *worst-case regret* (or simply *regret*) of policy π for family $\mathcal{F}$ is $\mathrm{Reg}(\pi, \mathcal{F}) := \sup_{D\in\mathcal{F}} \mathrm{Reg}(\pi, D)$.

Sometimes (e.g. in Section 4) it will be convenient to work directly with the revenue functions. In such cases, by abuse of notations, we may write $r(\pi, R)$ as the regret of under reward function is R and $r(\pi, \mathcal{F})$ the worst case regret under a family $\mathcal{F}$ of revenue functions.

2.1. Assumptions

Now we state some assumptions that all of our results rely on. Throughout, all demand functions is assumed to be continuous. For any (parametric) demand function $\mathcal{F} = \{D(x;\theta) : \theta \in \Theta\}$, we define its revenue (or reward) function to be the $R(x;\theta) = x \cdot D(x;\theta)$. We make the following assumptions for the input family $\{D(x;\theta) : \theta \in \Theta\}$ of demand functions. We start with a standard assumption, e.g. Broder and Rusmevichientong (2012).

Assumption 1: Compact Parameter Space. We assume $\Theta \subseteq \mathbb{R}^m$ is compact.

This compactness assumption leads to many favorable properties. For instance, the maximum (resp. minimum) value of any demand functions in $\mathcal{F}$ exists, and hence we can without loss of

[1] The corresponding demand function can naturally be backed out from a reward function: $D(x) = R(x)/x$ for $x > 0$.

[2] For the sake of completeness, a *policy* is, formally, a time-indexed sequence of functions $\pi = \{\pi_t : ([0,1]\times[0,1])^{t-1} \to [0,1], t = 1, \ldots, T\}$, where each function π_t maps the prices selected and demands observed over the previous $t-1$ rounds to a price for round t.

generality re-scale both the price and the target space to be $[0,1]$. Moreover, by definition, the target space of any revenue function is contained in $[0,1]$.

Assumption 2: Smooth Parametrization. The mapping $D:[0,1]\times\Theta\to[0,1]$ is twice differentiable and admits continuous second partial derivatives. Consequently, by compactness of $[0,1]\times\Theta$, there exists a universal constant $C_2>0$ such that $|D^{(j)}(x,\theta)|\le C_2$ for any $j=0,1,2$ and any $(x,\theta)\in[0,1]\times\Theta$.

Let $x^*:\Theta\to[0,1]$ be the *Optimal Price Mapping* that maps a parameter θ to the *maximum*[3] optimal price of $D(x;\theta)$.

Assumption 3: Regular Optimal Price Mapping. We assume that the mapping x^* is differentiable and admits continuous first derivatives. Consequently, by compactness of Θ, there exists a constant $C_3>0$ s.t. $|\frac{d}{d\theta}x^*(\theta)|\le C_3$ for any $x\in[0,1]$.

Assumption 4: Inclusive Price Space. To utilize the smoothness around the optimal price, it is usually assumed (e.g. Broder and Rusmevichientong (2012)) that the derivative at any optimal price vanishes. To be more precise, in this work we assume that for any $\theta\in\Theta$, it holds $\frac{\partial}{\partial x}R(x^*(\theta),\theta)=0$.

Under this assumption, by Taylor expansion and Assumption 2, for any sufficiently small h it holds that

$$|R(x^*(\theta),\theta)-R(x^*(\theta)+h,\theta)|\le C_2h^2+o(h^2)=O(h^2).$$

Finally, we impose the only *distributional* assumption that facilitates the concentration bounds. To this aim we first introduce a standard (e.g. see Vershynin (2018)) definition.

Definition 2 (Subgaussian random variables). The *subgaussian norm* of a random variable X is

$$\|X\|_{\psi_2}:=\inf\{c>0:e^{X^2/c^2}\le 2\},$$

and X is said to be *Subgaussian* if $\|X\|_{\psi_2}<\infty$.

We will use the following concentration bound for subgaussian random variables.

Theorem 1 **(Hoeffding)**. *Suppose $X_1,..,X_n$ are independent subgaussian random variables, then for any $\delta>0$,*

$$\mathbb{P}\left[\sum_{i=1}^n(X_i-\mathbb{E}X_i)\ge\delta\right]\le\exp\left(-\frac{\delta^2}{2\sum_{i=1}^n\|X_i\|_{\psi_2}^2}\right).$$

Assumption 5: Subgaussian Noise. There exists a constant $C_5>0$ such that the random demand X at any price $x\in[0,1]$ and underlying parameter $\theta\in\Theta$ satisfies $\|X\|\le C_5$.

[3] Note that there may be multiple optima in the price space.

2.2. Measuring the Complexity of a Family

Intuitively, the exploration-exploitation trade-off for markdown pricing becomes harder to manage as the given family becomes more complex. Consider, for example, linear demand functions defined on $[p_{min}, p_{max}]$. If each function in the given family takes the form $D(x;b) = 1 - bx$ where only b is unknown and x, then the decision maker may simply estimate the (negative) slope b by sampling sufficiently many times at p_{max}. In contrast, if each function takes the form $D(x;a,b) = a - bx$ where *both* parameters a, b are unknown, then sampling at one price would *not* suffice, since for any price p and demand value d, there are infinitely many linear demand functions passing through the point (p, d). Rather, one needs to collect samples at two or more distinct (*sample prices*), thereby facing the following dilemma: far-apart sample prices may result in a high regret since the true optimal price may happen to lie between the sample prices, and squeezed sample prices lead to slow learning rate.

Indeed, later in this paper we will see that the single parameter families does admit lower regret bounds than the two-parameter family. However, instead of proving specific regret bounds for specific families, or under specific assumptions, can we find a more general theory that unifies the results? For instance, can we introduce a complexity index to measure the difficulty of performing markdown pricing for a given family, and provide tight regret bounds under this framework?

A natural idea is to use the number of parameters to define the complexity. However, this is not even well-defined, since there may be multiple ways to parametrize the same family, with different numbers of parameters. It turns out that finding such a complexity measure *per se* is the very first challenge we need to address.

2.3. Dimension of A Family

In this work, we do propose such a complexity index, called *dimension*, and provide a complete settlement of the problem in this framework. To present its formal definition, we first introduce the notion of profile mapping. Loosely, for a fixed subset of prices, this mappings assigns an "ID" to each demand function, by encoding them, using their values at those prices.

Definition 3 (Profile Mapping). Let $\mathcal{F} = \{D(x,\theta) : \theta \in \Theta\}$ be a set of functions defined on some interval $[p_{min}, p_{max}]$. For any fixed $x = (x_1, ..., x_d) \in [p_{min}, p_{max}]^d$, the *profile-mapping* is defined as

$$\begin{aligned} \Phi_{x_1,\ldots,x_d} : \Theta &\to \mathbb{R}^d, \\ \theta &\mapsto (D(x_1;\theta), ..., D(x_d;\theta)) . \end{aligned}$$

Now we are ready to define dimension. Roughly, a family of demand functions has dimension d if it can be uniquely identifiable at any $(d+1)$ prices, and moreover, the parametrization is *robust*, in the sense that if the observed mean demand is corrupted slightly by noise, then the uniquely determined demand function is not too far away from the true demand function.

DEFINITION 4 (DIMENSION). A family $\mathcal{F}$ of functions has dimension d if there exists constants $C, \delta_0 > 0$ (possibly dependent on d) such that for any $(d+1)$ *distinct* prices $x_0, x_1 ..., x_d \in [0,1]$, the profile mapping Φ_x has the following properties:

- **Unique Profile:** the mapping Φ_x is injective, i.e. distinct θ, θ' correspond to distinct *profiles*,
- **Robust Parametrization:** let $h = \min_{i \neq j} |x_i - x_j| > 0$, then for any $\theta \in \Theta, \delta \in (0, \delta_0)$ and $\hat{y} \in \mathbb{R}^{d+1}$ with $\|\hat{y} - \Phi(\theta)\|_2 \leq \delta$, it holds $\|\Phi_x^{-1}(\hat{y}) - \theta\|_2 \leq C\delta h^{-d}$.

Geometrically, the robustness condition states that the pre-image of any sufficiently small ball centered at the profile vector of θ (in the target space) must be contained in a small ball centered at θ (in the original parameter space Θ). Note that if $\mathcal{F}$ has dimension d then it is also dimension d' for any $d' \geq d$. Subsequently for any family $\mathcal{F}$ of mappings, we define $d(\mathcal{F})$ to be the minimal integer d that satisfies the above two conditions.

We illustrate our definition by considering the dimensions of some commonly used demand families. Loosely speaking, as the simplest family, a 0-dimensional family consists of *disjoint* demand functions that can be parameterized using just one parameter. In fact, one may verify that our definition of 0-dimensional family (under our assumptions) is equivalent to the *separable* family as defined in Section 4 of Broder and Rusmevichientong (2012). We provide more concrete examples below.

PROPOSITION 1. *The following families are* 0*-dimensional:*

- *single-parameter linear demand functions:* $\mathcal{F}_0 = \{1 - ax : b \in [a_{min}, a_{max}]\}$,
- *exponential demand functions:* $\mathcal{F}_1 = \{e^{1-ax} : b \in [a_{min}, a_{max}]\}$,
- *logit demand functions:* $\mathcal{F}_2 = \{\frac{e^{1-ax}}{1+e^{1-ax}} : b \in [a_{min}, a_{max}]\}$,

For finite $d \geq 1$, one may verify that any family of degree d polynomials is d-dimensional, assuming the parameters are bounded.

PROPOSITION 2. *Suppose* $\mathcal{F} = \{\theta_d x^d + ... + \theta_1 x + \theta_0 : \theta_i \in [\ell_i, u_i]\}$ *be a family degree-d polynomials where* $\ell_i < u_i$*, then* $d(\mathcal{F}) = d$.

PROPOSITION 3. *If* $\mathcal{F}$ *be the set of all 1-Lipschitz functions on* $[0,1]$*, then* $d(\mathcal{F}) = \infty$.

In this work, for each finite $d = 0, 1, 2, ...$, we will propose an efficient markdown pricing policy, which we also prove to have best possible theoretical guarantees.

2.4. Sensitivity of a Family

Essentially, dimension measure the degrees of freedom for a family of demand functions. But intuitively, the smoothness should also affect the regret. Consider the Taylor expansion of a reward function $R(x)$ around an optimal price x^*:

$$R(x) = R(x^*) + 0 + \frac{1}{2!}R''(x^*)(x - x^*)^2 + \frac{1}{3!}R^{(3)}(x^*)(x - x^*)^3 + ...$$

Suppose the first nonzero derivative is $R^{(k)}(x^*)$. Then, the higher k, the less the revenue is *sensitive* to overshooting (i.e. $x < x^*$).

This motivates us to introduce the following concept, *sensitivity*, that measures how fast the revenue function changes around the optimal price.

Definition 5. A family $\mathcal{F}$ of demand functions is *s-sensitive* if there exists a constant $C_6 > 0$ such that for any $D \in \mathcal{F}$ and any price $x \in [0, 1]$ it holds

$$D(x^*) - D(x) \leq C'|x - x^*|^s.$$

In words, the regret per round at a price distance h from x^* is bounded by $O(h^s)$.

For unconstrained pricing (or continuum bandits), Kleinberg (2005) showed that a tight $T^{\frac{s+1}{2s+1}}$ regret for Lipschitz reward functions. In particular, for $s = 1$, the regret becomes $T^{2/3}$, which is strictly lower than the $T^{3/4}$ result. As a natural question, can we show such separation for $s \geq 2$?

Note that by the smoothness assumption of $\mathcal{F}$, the second derivatives are bounded by some constant, hence $s \geq 2$. For any $s \geq 2$, one can easily construct a family of s-sensitive by considering revenue functions $R_\theta(\theta + h) := 1 - |h|^s$.

3. Policies and Results

We give an overview of our policies and prove regret upper bounds for zero, finite (non-zero) and infinite dimensional families of demand functions. Moreover, we show that our policies achieves nearly-optimal regret by providing lower bounds for each of these regimes.

3.1. Zero-Dimensional Families

We start with the simplest case, 0-dimensional demand functions. We propose a policy called *Cautious Myopic* which proceeds by *phases* and makes *conservative* decisions. As opposed to the *optimism* in the face of uncertainty in UCB type policies, our policy adopts *conservatism* in the face of uncertainty. More precisely, in the j-th phase, the policy estimates the parameter θ^* using the observations from the last phase and builds a confidence interval I around it, then selects the *largest* optimal price of any parameter in I in the next $t_j := \lceil 2^j \log T \rceil$ rounds. We write $t^{(j-1)} := \sum_{k=0}^{j-1} t_k$ and formally state this policy in Algorithm 1.

Algorithm 1 Cautious Myopic Policy.

Input: a family $\mathcal{F}$ of demand functions and time horizon T.
$p_1 \leftarrow 1$ ▷ Initialization
for $j = 1, ..., \log T$ **do**
  for $t = t^{(j-1)} + 1, ..., t^{(j-1)} + t_j$ **do** ▷ Phase j starts
    $x_t \leftarrow \hat{p}_j$ ▷ Select the same price in every round
    observe realized demand D_t
  $\bar{d}_j = \frac{1}{t_j} \sum_{\tau=1}^{t_j} D_{t^{(j-1)}+\tau}$ ▷ Empirical mean demand in phase j
  $\hat{\theta} \leftarrow \Phi_{p_j}^{-1}(\bar{d})$ ▷ Estimate parameter
  $w_j \leftarrow 2\sqrt{\frac{C_5 \log T}{t_j}}$ ▷ Width of the confidence interval
  $p_{j+1} \leftarrow \max\{x^*(\theta) : |\theta - \hat{\theta}_{t-1}| \le w_j\}$ ▷ Conservative estimation of the optimal price

By bounding the expected regret in each round using concentration bounds, we obtain our first following upper bound.

THEOREM 2 **(Zero-dimensional Upper Bound)**. *Let $\mathcal{F}$ be any zero-dimensional s-sensitive family of demand functions. Then the Cautious Myopic (CM) Policy has regret*

$$\mathrm{Reg}(\mathrm{CM}, \mathcal{F}) = \begin{cases} O(\log^2 T), & \text{if } s = 2, \\ O(\log T), & \text{if } s > 2. \end{cases}$$

As we discussed earlier, many simple families of demand functions, such as single-parameter linear or exponential demand functions satisfy $s = 2$, thereby having regret $O(\log^2 T)$. It is worth noting that this bound is asymptotically higher than the $O(\log T)$ bound in the absence of the markdown constraint (Broder and Rusmevichientong (2012)). Intuitively, this is because the CM policy strikes a balance between the risk of overshooting (the optimal price) and getting close to the

optimal price, by purposely distancing from the estimated optimal price. Is this trade-off optimal? In other words, can we achieve $o(\log^2 T)$ regret by taking more risk or being more conservative?

Somewhat surprisingly, we confirm that the amount of caution in CM is indeed optimal, by showing an $\Omega(\log^2 T)$ lower bound. Further, this results provides the first *separation* between the $O(\log T)$ regret for unconstrainted pricing and markdown pricing for 0-dimensional demand families.

THEOREM 3 **(Zero-Dimensional Lower Bound)**. *For any $\theta \in \mathbb{R}$, define $D_\theta(x) = 1 - \theta x$ for $x \in [\frac{1}{2}, 1]$ and consider $\mathcal{F} = \{D(x;\theta) : \theta \in [\frac{1}{2}, 1]\}$. Then, $\mathcal{F}$ is 0-dimensional and for any policy π, $\mathrm{Reg}(\pi, \mathcal{F}) = \Omega(\log^2 T)$.*

3.2. Finite-Dimensional Families

Now we consider finite, nonzero dimensional families. Different from the zero-dimensional case, now the learner is no longer able to estimate the true parameter θ at a single price. Rather, for dimension d, the learner needs to collect demand samples at $d+1$ distinct *sample prices*. This, however, introduces extra regret, since the optimal price may lie *between* these sample prices.

Intuitively, a reasonable policy needs to trade off between the overshooting risk and the learning rate. If the gap is large, the policy may learn the parameter efficiently, but there is potentially a higher regret due to overshooting, in case the true optimal price lie between the sample prices. On the other side, if the gap is small then there is less risk of overshooting but a slower rate of learning.

We introduce our Iterative Cautious Myopic (ICM) Policy (Algorithm 2) that strikes such balance nearly optimally, as we will soon see from Theorem 4 and Theorem 5. The policy consists of m phases. In phase $j \in [m]$, the policy collects T_j samples at each of the d evenly spaced *sample prices* with distance h. Then, based on the observed demands, the policy estimates the optimal price and constructs a confidence interval $[L_j, U_j]$ centered at $\hat{p}_j$.

To determine the initial price p_{j+1} in the next phase, the policy considers three cases the following three cases. In the *good* event, the current sample price, $p_j - dh$, lies on the right of the confidence interval $[L_j, U_j]$. In this case, we may simply select the next price to be the right endpoint U_j of thus interval. In the *dangerous* event, the current price is within the confidence interval, and we may (or may not) already overshot. In this case, we can no longer select p_{j+1} to be U_j, due to the markdown constraint. Instead, we select the current price $p_{j+1} = p_j - dh$. Finally in the

overshooting event, as the name suggests, our current price is already lower than the left endpoint of the confidence interval, and hence with high probability we have overshot the optimal price, so we exit the exploration phase (i.e. the outer for-loop) immediately and enter the exploitation phase by selecting the current price in all future rounds.

Algorithm 2 Iterative Cautious Myopic Policy.

Input: A family $\mathcal{F}$ of demand functions and time horizon $T > 0$.
$p_1 \leftarrow 1, L_0 \leftarrow 0, U_0 \leftarrow 1$. ▷ Initialization
for $j = 1, 2, ...m$ **do** ▷ Phases
  for $k = 0, 1, ..., d$ **do** ▷ Sample at $(d+1)$ equi-distant prices
    Select price $p_j - kh$ for T_j times.
    $\bar{D}_k \leftarrow \frac{1}{T_j} \sum_{i=1}^{T_j} D_i$ ▷ Mean demand at $p_j - kh$
  $\hat{\theta} \leftarrow \Phi^{-1}_{p_j,...,p_j-dh}(\bar{D}_0, ..., \bar{D}_d)$ ▷ Estimate Parameter
  $w_j \leftarrow 2Ch^{-d}\sqrt{\frac{C_5 d \log T}{T_j}}$ ▷ Width of confidence interval
  $L_j \leftarrow \min\{x^*(\theta) : \|\theta - x^*(\hat{\theta})\|_2 \le w_j\}$ ▷ Lower confidence bound
  $U_j \leftarrow \max\{x^*(\theta) : \|\theta - x^*(\hat{\theta})\|_2 \le w_j\}$ ▷ Upper confidence bound
  if $U_j \le p_j - dh$ **then** $p_{j+1} \leftarrow U_j$ ▷ Good event
  if $U_j > p_j - dh \ge L_j$ **then** $p_{j+1} \leftarrow p_j - dh$ ▷ Dangerous event
  if $p_j - dh < L_j$ **then** Break ▷ Overshooting event
Select the current price in every future round ▷ Exploitation

THEOREM 4 **(Upper Bound for Finite $d \ge 1$)**. *For suitable choice of parameters, the ICM policy achieves regret* $\mathrm{Reg}(\mathrm{ICM}, \mathcal{F}) = \tilde{O}\left(T^{\rho(m,s,d)}\right)$ *where*

$$\rho(m, s, d) = \frac{1 + \left(1 + \frac{s}{2} + ... + (\frac{s}{2})^{m-1}\right) d}{1 + \left(1 + \frac{s}{2} + ... + (\frac{s}{2})^{m-1}\right) \cdot (d+1) + \left(\frac{s}{2}\right)^m}.$$

In particular for $s = 2$,

$$\mathrm{Reg}\left(\mathrm{ICM}\left(T_1, ..., T_m, h\right), \mathcal{F}\right) = \tilde{O}\left(T^{\frac{d}{d+1}}\right).$$

In contrast to the upper bound for zero-dimensional family where the regret is only logarithmic in T, for $d \ge 1$ the regret increases to polynomial in T. We complement the upper bound with an almost-matching lower bound. In our proof, for each $d \ge 1$ we consider a sub-family of $(d+1)$-degree decreasing polynomial demand functions – which is also d-dimensional – and show that there is a pair of such demand functions on which any policy suffers regret $\Omega(T^{\frac{d}{d+1}})$.

THEOREM 5 **(Lower Bound for Finite $d \geq 1$).** *For any $d \geq 2$, there exists a d-dimensional family $\mathcal{F}$ of demand functions on $[0,1]$ s.t. for any markdown policy π,*

$$\mathrm{Reg}(\pi, \mathcal{F}) = \Omega(T^{\frac{d}{d+1}}).$$

3.3. Infinite Dimensional Families

For the infinite dimensional functions, it is more convenient to work with the reward (or revenue, which we use interchangeably) function $R(x) := x \cdot D(x)$ corresponding to the demand function $D(x)$. It is straightforward to verify that a family of demand functions have dimension $0 \leq d \leq \infty$ if and only if its corresponding reward functions have dimension d.

Many previous work on dynamic pricing and multi-armed bandits focused on *infinite* dimensional families of demand functions. For example, it has been shown that for the family of Lipschitz demand functions an $\tilde{O}(T^{2/3})$ regret can be achieved (Kleinberg (2005), Broder and Rusmevichientong (2012)). Another well-studied setting is when the reward functions corresponding to the demand functions are unimodal (Yu and Mannor (2011), Combes and Proutiere (2014)), where binary search type policy achieves $\tilde{O}(T^{1/2})$ regret under an additional lower Lipschitz assumption.

In contrast to the unconstrainted version, for the markdown pricing problem, no markdown policy achieves $o(T)$ regret on the family of Lipschitz reward functions (see Jia et al. (2021)). In fact, consider reward functions with possibly multiple local optima. Suppose a policy detects a local optimum at some high price p_{high}, then it faces a dilemma: if it stops at p_{high}, then a high regret is incurred since it may potentially earn rewards at a faster rate at some lower price. On the other side, if it does further reduce the price, it may be the case that no lower prices have as high reward rate as at p_{high}, and due to the markdown constraint, the policy may not increase the price back to p_{high} hence a high future regret is incurred.

It is also worth noting that for finite dimensional families such dilemma is lifted, since by definition of dimension, the learner may infer whether or not a lower price has higher reward rates by simply collecting more samples at p_{high}. This is, however, not true for the Lipschitz family, since two Lipschitz reward functions that behave drastically differently at low prices may be completely identical at higher prices.

Nonetheless, Jia et al. (2021) showed that if the underlying demand functions are assumed to be Lipschitz and unimodal (which are both satisfied by many commonly used families), then a tight $\tilde{\Theta}(T^{3/4})$ regret is achievable. With this unimodal assumption, the markdown pricing problem

essentially becomes finding the unique local optimum of the true revenue function. Specifically, their lower bound is derived on a family of Lipschitz reward functions where the reward rate may change abruptly at the peak.

Can the regret bound be improved if the reward functions are assumed to change smoothly? We answered this question definitely by generalize their result to incorporate the sensitivity parameter. Let $\mathcal{F}_s^U$ be the family of unimodal, s-sensitive reward functions.

Algorithm 3 Uniform Elimination Policy ($\text{UE}_{s,\delta}$).

Input: $w, \Delta > 0$. ▷ Step size and width of target confidence intervals.
Initialize: $x \leftarrow p_{max}$, $\text{LCB}_{\max} \leftarrow 0$, $k \leftarrow \lceil 3\delta^{-2} \log T \rceil$.
while $x > 0$ **do** ▷ Exploration phase starts.
    Select price x for the next k rounds and observe rewards $X_1, ...X_k$.
    $\bar{\mu} \leftarrow \sum_{i=1}^{k} X_i$ ▷ Compute mean rewards.
    $[\text{LCB}, \text{UCB}] \leftarrow [\bar{\mu} - w, \bar{\mu} + w]$. ▷ Compute confidence interval for reward at current price.
    if $\text{LCB} > \text{LCB}_{\max}$ **then** ▷ Update best LCB so far
        $\text{LCB}_{\max} \leftarrow \text{LCB}$.
    if $\text{UCB} < \text{LCB}_{\max}$ **then** ▷ Exploration phase ends.
        $x_h \leftarrow x$. Break. ▷ Define *halting* price.
    else $x \leftarrow s$.
Select price x_h in all future rounds. ▷ Exploitation phase.

THEOREM 6 **(Upper Bound for Infinite-dimensional Family)**. *For any $s \geq 2$, the Uniform Elimination Policy satisfies* $\text{Reg}(\text{UE}_{w,\Delta}, \mathcal{F}_s^U) = O(T^{\frac{2s+1}{3s+1}})$.

We then complement Theorem 6 with a lower bound that matches it for every $s \geq 2$. The challenge of the peak finding game is, on the one hand if we reduce the prices too fast, we may have overshot by a lot when we halt; on the other hand if the speed is too slow we may spend too much time at suboptimal prices, incurring high regret. We formalize the above idea and show the following lower bound.

THEOREM 7 **(Lower Bound for s-sensitive Family)**. *For any $s \geq 2$, there is a set $\mathcal{F}$ of s-sensitive family of unimodal revenue curves satisfying Assumptions (1)-(4) such that any markdown policy π satisfies* $\text{Reg}(\pi, \mathcal{F}) \geq \Omega(T^{\frac{2s+1}{3s+1}})$.

This tight regret bound, $T^{(2s+1)/(3s+1)}$, highlights how sensitivity helps reducing the regret for markdown pricing. Interestingly, as s grows to infinity, the regret approaches $T^{2/3}$, matching the regret of the unconstrained pricing problem *without* any smoothness assumption.

We summarize our results for $s=2$ in Table 1. We highlight our results in red, and emphasize that each entry corresponding to two results, an upper bound and a matching lower bound. Notation $\tilde{\Theta}$ mean ignoring $\log T$ terms.

dimension	Markdown	Unc. Pricing
$d=0$	$\Theta(\log^2 T)$	$\Theta(\log T)$
$1 \le d < \infty$	$\tilde{\Theta}(T^{d/(d+1)})$	$\tilde{\Theta}(\sqrt{T})$
$d=\infty, 2 \le s < \infty$	$\tilde{\Theta}(T^{(2s+1)/(3s+1)})$	$\tilde{\Theta}(\sqrt{T})$

Table 1 **Regret bounds for markdown and unconstrainted pricing under unknown demand for $s=2$.**

3.4. A Key Tool For Lower Bounds

Due to the monotonicity constraint, standard lower bound techniques fail to work for the markdown pricing problem. Thus, we rely upon Theorem due to Wald and Wolfowitz (1948) - which is not as well-known as it deserves to be, at least in the bandits literature - along with a a generalized version due to Jia et al. (2021), for adaptive sequential hypothesis testing between two *demand curves* on an interval.

As opposed to a "traditional" hypothesis testing algorithm that collects a *fixed* number of samples and then returns one of these hypotheses, an adaptive hypothesis testing algorithm collects samples until some certain criterion is satisfied, say when the likelihood ratios reaches some prescribed threshold. We consider adaptive hypothesis testing algorithms that return one of two given demand curves after *adaptively* collecting samples in a monotonic fashion. The Generalized Wald-Wolfowitz Theorem asserts that expected number of samples this algorithm collects must be lower bounded by a function in terms of the type I, II error and also the maximum KL-divergence between these two demand distributions, over all prices in the given interval.

The basic idea is to construct a family of reward functions each having a unique optimal price, then use GWW to prove that any low-regret policy has to spend **in expectation** at least certain number of rounds to distinguish between each pair of reward functions, whose maxima occur nearby. As a result, if the optimal price is small, a high regret is incurred since the policy "wasted" too much time in suboptimal prices.

4. Zero-Dimensional Families

In this section, we prove the following tight regret bounds for the markdown version. To highlight the technical challenges, we first rephrase the known tight regret bound for non-markdown version.

THEOREM 8 **(Broder and Rusmevichientong (2012))**. *For any zero-dimensional demand family $\mathcal{F}$, there is an algorithm with regret $O(\log T)$. Moreover, there exists a zero-dimensional demand family $\mathcal{F}$ on which any algorithm has regret $\Omega(\log T)$.*

They considered a simple policy that estimates the true parameter using maximum likelihood estimator (MLE), and then selects the optimal price of the estimated demand function. To bound the expected regret in round t, we showed that the *mean squared error* (MSE) of the estimated price is at most $1/t$, and hence the expected total regret is $\sum_{t=1}^{T} \frac{1}{t} \sim \log T$.

4.1. Upper Bound

While Theorem 8 is established by bounding the Mean Square Error (MSE), due to the monotonicity constraint for markdown pricing, it no longer suffices to consider the *mean* error. Rather, we need an error bound which (i) holds with high probability, so that we can make *conservative* decision by selecting a price that is extremely unlikely to overshoot the optimal price, and (ii) is sufficiently low, so that the total regret is also low. The following lemma can be obtained as a direct consequence of Hoeffding's inequality (Theorem 1). We provide a formal proof in Appendix A for completeness.

LEMMA 1. *Let θ^* be the underlying parameter and $\bar{d} = \bar{d}_t$ be the empirical mean of t i.i.d. samples at some price x. Let $\mathcal{E}$ be the event that $\left|D(x;\theta^*) - \bar{d}\right| \le 2\sqrt{\frac{C_5 \log T}{t}}$, then $\mathbb{P}[\overline{\mathcal{E}}] \le T^{-2}$.*

Conditional on $\mathcal{E}$, the true parameter θ^* is contained in the confidence interval $I = [\bar{d} - w, \bar{d} + w]$ where $w = 2\sqrt{\frac{C_5 \log T}{t}}$, so the next selected price $\hat{p} = \max\{x^*(\theta) : \theta \in I\} \ge x^*(\theta^*)$, i.e. our policy does not overshoot the true optimal price.

On the other side, to see why the estimated price is close to $x^*(\theta^*)$, recall that by definition of dimension, the parameter $\hat{\theta} = \Phi_x^{-1}(\bar{d})$ determined by the observed demands is at most $2C \cdot \sqrt{\frac{C_5 \log T}{t}}$ distance away from the true parameter θ^*. By Assumption 3, $|\frac{d}{d\theta} x^*(\theta)| \le C_3$ for some constant $C_3 > 0$, so the price $\hat{p}_j$ selected in phase $j+1$ satisfies

$$|\hat{p}_j - x^*(\theta^*)| \le C_3 \|\hat{\theta}_j - \theta^*\|_2 \le 2C \cdot C_3 \cdot \sqrt{\frac{C_5 \log T}{t_j}}.$$

Since the length of phase $j+1$ is t_{j+1}, the regret incurred in this phase is at most $C'\left(2C\cdot C_3\cdot\sqrt{\frac{C_5\log T}{t_j}}\right)^s\cdot t_{j+1}$. Summing over all $K\le\log T-\log\log T$ phases, we can bound the cumulative regret as

$$\begin{aligned}\mathrm{Reg}(\mathrm{CM},\mathcal{F}) &\le \sum_{j=1}^{K} C'\left(2C\cdot C_3\sqrt{\frac{C_5\log T}{t_j}}\right)^s\cdot t_{j+1}\\ &= C'\left(2C\cdot C_3\sqrt{C_5\log T}\right)^s\cdot\sum_{j=0}^{K}\frac{t_{j+1}}{t_j^{s/2}} \qquad (1)\end{aligned}$$

We substitute t_j with $2^j\log T$ and simplify the above for $s=2$ and $s>2$ separatly. When $s=2$,

$$\begin{aligned}(1) &= C'\left(2C\cdot C_3\sqrt{C_5\log T}\right)^2\cdot\sum_{j=0}^{K}\frac{t_{j+1}}{t_j}\\ &= C'\left(2C\cdot C_3\cdot\right)^2 C_5\log T\cdot(\log T-\log\log T)=O(\log^2 T).\end{aligned}$$

Now suppose $s>2$. Then,

$$\begin{aligned}(1) &= C'\left(2C\cdot C_3\sqrt{C_5\log T}\right)^s\cdot\sum_{j=0}^{K}\frac{2^{j+1}\log T}{2^{j\cdot\frac{s}{2}}\log^{s/2}T}\\ &\le C'\left(2C\cdot C_3\cdot\sqrt{C}_5\right)^s\cdot\log T\cdot\sum_{j=0}^{K}2^{(1-\frac{s}{2})j+1}\\ &\le 2C'\cdot\left(2C\cdot C_3\cdot\sqrt{C}_5\right)^s\cdot\log T\cdot\int_0^K 2^{(1-\frac{s}{2})x}dx\\ &= 2C'\cdot\left(2C\cdot C_3\cdot\sqrt{C}_5\right)^s\cdot\log T\cdot\frac{2}{(s-2)\ln 2}=O(\log T).\end{aligned}$$

Theorem 2 follows by combining these the analysis for $s=2$ and $s>2$. □

4.2. Lower Bound

We sketch the high level idea of the proof, while deferring the formal proof to Appendix E. The above algorithm motivates the following idea for showing a lower bound. Consider a policy π with $O(\log^2 T)$ regret. Fix a linear demand curve R, called the red curve, whose optimal price we denote p_R^*. For each t, we bound the expected regret in round t as follows. Choose $\Delta_t\sim\sqrt{\frac{\log T}{t}}$ and construct a blue linear demand function $B=B(t)$ whose optimal price is Δ_t greater than p_R^*. Consider the following classifier induced by the price choice of π: define the output of this classifier to be R if the price selected in round t is closer to p_R^*, and B otherwise.

Recall that the WW theorem asserts that if both the type I, II errors of a classifier are "low", then the expected numbers of samples under both hypotheses are "high". Now, on the one hand,

since the policy has low regret, under B the probability for overshooting should be extremely low. On the other hand, we will consider consider small t, so that the number t of samples that the classifier collects is "low". Now, in order not to contradict WW Theorem, the error probability under R must be large! In other words, with considerable probability, the price selected at time t is greater than $p_R^* + \Delta_t/2$, hence a high regret is incurred under R in round t. Thus, by summing the expected regret from round $t = \log T$ to $\sqrt{T}$, we can lower bound the regret by

$$\sum_{t=1}^{\sqrt{T}} \Delta_t^2 = \sum_{t=1}^{\sqrt{T}} \frac{\log T}{t} \sim \log^2 T.$$

5. Finite-Dimensional Families

In this section we first analyze the regret of the ICM policy and prove Theorem 4, and then complement this upper bound with an almost matching lower bound, using the Wald-Wolfowitz Theorem as stated in Section 3.4.

5.1. Upper Bound

Recall that the ICM policy is specified by two types of parameters: the gap h between neighboring sampling prices in each phase, and the number T_j of rounds to stay at each sampling price in phase j. To prove Theorem 4, we first present the following upper bound on the regret of ICM for arbitrary choice of parameters h and T_j's, and then optimize the choice of parameters (up to polylogarithmic factors in T) by solving a linear program.

PROPOSITION 4. *Let $\mathcal{F}$ be a d-dimensional, s-sensitive ($s \geq 2$) family of demand functions. Suppose $0 < T_1 < ... < T_m$ with $m = o(1)$ and $h > 0$. Then, the regret of* $\mathrm{ICM} = \mathrm{ICM}(T_1, ..., T_m, h)$ *is*

$$\mathrm{Reg}(\mathrm{ICM}, \mathcal{F}) \leq T_1 + C' \left(2C_3 C h^{-d} \sqrt{C_5 d \log T}\right)^s \cdot \left(\sum_{j=1}^{m-1} T_{j-1}^{-s/2} \cdot T_j + T_m^{-s/2} \cdot T\right) + C' (mdh)^s T.$$

We briefly explain the intuitions behind the above result before proceeding with finding the optimal parameters. As the name suggests, the Iterative Cautious Myopic policy iteratively computes a confidence interval $[L_j, U_j]$ around the true optimal price, and conservatively moves to the right endpoint of this interval. As a *simplistic* view, in phase j (assuming it ever takes place) the estimation error is $\sim h^{-d} T_{j-1}^{-1/2}$, and by definition of s-sensitivity, the regret incurred in phase j is $\sim (h^{-d} T_{j-1}^{-1/2})^s T_j$.

To understand the final term, observe that when h is sufficiently small compared to $U_j - L_j$, there is little risk of *overshooting* at the right endpoint U_j. However, when one selects larger h (for faster learning rate), it may happen that the last sample price $p_j - dh$ in this phase overshoots the optimal price, thereby incurring a regret term, as captured by the last term in the above bound.

Nonetheless, the actual proof involves carefully analyzing each of the three events (good, dangerous and overshooting) that can possibly occur at the end of each phase, as formally defined in Algorithm 2. We defer the proof to Appendix B.

We now determine the hyper-parameters to minimize this upper bound by solving a linear program. For simplicity, we only consider $s = 2$. The generalization to $s > 2$ is straightforward.

Write $T_i = T^{z_i}$, $h = T^{-y}$. Then,

$$(h^{-d} T_j^{-1/2})^s T_{j+1} = T^{sdy - \frac{s}{2} z_j + z_{j+1}},$$

and the optimal parameter can be derived by solving

$$\begin{aligned}
LP(d): \quad \min_{x,y,z} \quad & T^x \\
\text{s.t.} \quad & T^{z_1} \le T^x, && \text{Regret in phase 1} \\
& T^{2sdy + z_2 - \frac{s}{2} z_1} \le T^x, && \text{Regret in phase 2} \\
& \cdots \\
& T^{sdy + 1 - \frac{s}{2} z_m} \le T^x, && \text{Regret in the exploitation phase} \\
& T^{1 - sy} \le T^x, && \text{Regret for overshooting} \\
& x, y, z \ge 0, z \le 1
\end{aligned}$$

Taking logarithm with base T on both sides, the above becomes

$$\begin{aligned}
\min_{x,y,z} \quad & x \\
s.t. \quad & \begin{bmatrix}
-1 & 0 & 1 & 0 & 0 & 0 & \dots & 0 \\
-1 & sd & -\frac{s}{2} & 1 & 0 & 0 & \dots & 0 \\
-1 & sd & 0 & -\frac{s}{2} & 1 & 0 & \dots & 0 \\
-1 & sd & 0 & 0 & -\frac{s}{2} & 1 & \dots & 0 \\
 & & & \cdots\cdots & & & & \\
-1 & sd & 0 & 0 & 0 & \dots & -\frac{s}{2} & 1 \\
-1 & sd & 0 & 0 & 0 & 0 & \dots & -\frac{s}{2} \\
-1 & -s & 0 & 0 & 0 & 0 & \dots & 0
\end{bmatrix}
\begin{bmatrix} x \\ y \\ z_1 \\ z_2 \\ \vdots \\ z_{m-1} \\ z_m \end{bmatrix}
\le
\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ -1 \\ -1 \end{bmatrix} \\
& x, y, z \ge 0, \\
& z \le 1
\end{aligned}$$

By solving the above linear program, choose $h = T^{-\frac{m}{2(d+1)m+2}}$ and $T_j = T^{\frac{md+j}{(d+1)m+1}}$ for $1 \le j \le m$. Then, by Proposition 4, we have

$$\begin{aligned}\mathrm{Reg}(\mathrm{ICM}, \mathcal{F}) &= \tilde{O}\left(T_1 + \sum_{j=1}^{m-1} h^{-2d} T_j^{-1} T_{j+1} + h^{-2d} T_m^{-1} T + (mdh)^2 T\right) \\ &= \tilde{O}\left(m^2 d^2 T^{\frac{md+1}{(d+1)m+1}}\right).\end{aligned}$$

It is straightforward to verify that when $m = d^{-2} \log T$, the above becomes $\tilde{O}(T^{\frac{d}{d+1}})$.

We may generalize this to $s \ge 2$. In the appendix we show the how to and obtain Theorem **??**.

5.2. Lower Bound

We describe our proof at a high level and defer the details to Appendix E. For each d we construct a pair of demand functions D_{blue}, D_{red} on price space $[1/2, 1]$ with $D_{blue}(1) = D_{red}(1)$. Moreover, price 1 is the unique optimal price of D_{blue} and suboptimal for D_{red}. Since the gap between these two demand functions is tiny near price 1, to distinguish between them we have to select reduce the price away from 1. Thus learner faces the following trade-off: if she reduces the price by too much, then a high regret is incurred under D_{blue} since its optimal price is at 1; otherwise, the difference between these two curves is too small and she has to explore for too many rounds near price 1, which is suboptimal for D_{red}, hence incurring a high regret.

As the crucial step, in the appendix we explicitly construct a pair of demand curves with the following properties.

LEMMA 2. *For any $d \ge 1$, there exists a pair D_{red}, D_{blue} of degree-d polynomial demand functions satisfying the following properties.*

1. **Monotonicity:** *Both are non-increasing on $[1/2, 1]$,*
2. **First Order Optimality:** *Denote $R_i(x) = x \cdot D_i(x)$ for $i \in \{red, blue\}$, then $\max_{x \in [1/2,1]} R_{blue}(x)$ is attained at $x = 1$. Moreover, $R'_{blue}(1) = 0$,*
3. **Interior Optimal Price:** *The function R_{red} is maximized at some price $x \in [0, \frac{1}{2}]$,*
4. **Hardness Of Testing:** *Let $Gap(h) = \max_{x \in [1-h,1]}\{|D_{red}(x) - D_{blue}(x)|\}$, then $Gap(h) \le O(h^d)$ as $h \to 0^+$. In particular, this implies that $R_{blue}(1) = R_{red}(1)$.*

Consider a policy with low regret. Choosing suitable neighborhood $[1-h, 1]$, we convert this policy into a classifier that returns R_{red} or R_{blue} based on whether the price in round $\frac{T}{4}$ is within this neighborhood. We first argue that if the policy has low regret, then it has to perform reasonably well

on this classification problem, otherwise an $\Omega(h^2T)$ regret is incurred. Then, we use the generalized Wald-Wolfowitz Theorem (Theorem 10) to show that in order to distinguish between these two curves, the policy has to spend $\Omega(h^{-2d})$ rounds inside the neighborhood in expectation, incurring $\Omega(h^{-2d}T)$ regret under R_{red}.

6. Infinite-Dimensional Families

6.1. Upper Bound

In this section we first present a general regret upper bound for $\mathrm{UE}_{\Delta,w}$. Theorem 6 immediately follows then by choosing parameters $\Delta = T^{-1/(3s+1)}, w = T^{-2/(3s+1)}$.

PROPOSITION 5 **(Upper Bound)**. *Let $\mathcal{F}$ be any s-sensitive family for some $s \geq 2$. Then for any sufficiently small $\Delta, w > 0$, it holds that*

$$\mathrm{Reg}\left(\mathrm{UE}_{\Delta,w}, \mathcal{F}\right) = O\left(\Delta^{-1} w^{-2} \log T + (w + \Delta^s)T\right).$$

Proof. Consider a reward function $R \in \mathcal{F}$. Let $\ell = \arg\min_j\{|x_j - x^*|\}$ be the closest sample price to x^*, and $i = \arg\max_j\{L_j\}$ be the sample price with the highest confidence interval.

To analyze the regret in the exploitation phase, we next show that $R(x_{\ell+k}) - R(x^*) = O(\Delta^s + w)$. Suppose $0 \leq k \leq 2$, then the claim holds true trivially. In the more non-trivial case, suppose $k \geq 3$. In this case, the challenge is in showing that the policy will not stop "too late" after overshooting the optimal price, by incorporating the sensitivity of R. To this aim, we *lower* bound the difference between the mean reward rate at x_i and the halting price, $x_{\ell+k}$, in terms of the number k of steps from x_ℓ. We formally state this inequality below, whose proof is deferred to Appendix D.

LEMMA 3. *For sufficiently small Δ, for any $k \geq 3$ it holds that*

$$|R(x_{\ell+k}) - R(x_i)| \geq \frac{C}{2^s s!}(k\Delta)^s - 2w.$$

We complete the proof assuming this lemma holds. Suppose the stopping rule is not satisfied at some price $x_{\ell+k}$, formally, $[L(x_i), U(x_i)] \cap [L(x_{\ell+k}), U(x_{\ell+k})] \neq \emptyset$. Then, $|R(x_i) - R(x_{\ell+k})| \leq 4w$, and hence by Lemma 3,

$$4w \geq |R(x_i) - R(x_{\ell+k})| \geq \frac{C}{2^s s!}(k\Delta)^s - 2w. \tag{2}$$

Thus,

$$R(x_{\ell+k}) - R(x^*) \leq C_s \cdot ((k+1)\Delta)^s \leq \frac{2^s s!}{C}(\Delta^s + 6w) = O(\Delta^s + w).$$

where the second step follows from (2). Therefore, the regret incurred in the exploitation phase is $O((\Delta^s + w)T)$. Since there are $O(\Delta^{-1})$ sample prices, and the policy selects each of them for $O(w^{-2}\log T)$ times, the regret incurred in the exploration phase is then bounded by $O(\Delta^{-1}w^{-2}\log T)$. Combining these two terms, we have

$$\mathrm{Reg}(\mathrm{UE}_{\Delta,w}, \mathcal{F}) \le O\left(\Delta^{-1}w^{-2}\log T + (\Delta^s + w)T\right),$$

and the proof is complete. □

6.2. Lower Bound

The proof of this lower bound uses similar idea as in the lower bound proof in Jia et al. (2021). However, for each $s \ge 2$ we need to construct a family of s-sensitive family. Here, we sketch the high level idea and defer the technical details to Appendix E.

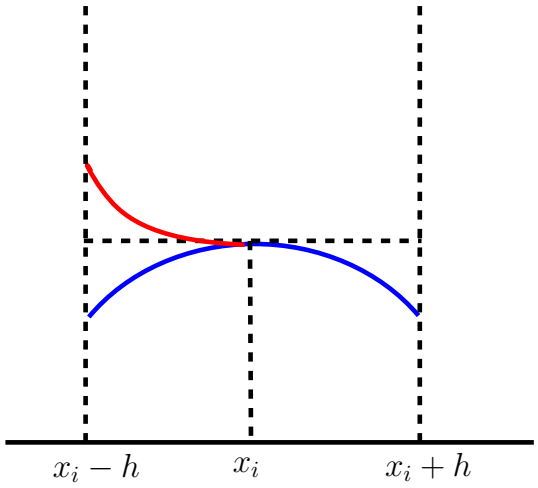

Figure 1 **Bow-shaped and S-shaped gadgets**

We consider the following s-sensitive family of unimodal reward curves. With some foresight, choose $h = T^{-\frac{1}{3s+1}}$. In the construction, we will use the following S-shaped curves (or *S curves*) and bow-shaped curves (or *B curves*), as shown in Figure 1.

We now formally describe those S and B curves. For simplicity we assume $m := \frac{1}{h}$ is an even integral. Define a decreasing sequence $x_i = 1 - (2i-1)h$ for each $i = 1, ..., m/2$ of prices. Each pair of curves B_i, S_i are defined in the interval $[x_i - h, x_i + h]$. These two curves are identical at higher prices than x_i and, scanning from right to left, start to diverge at a rate of h^s starting at x_i. Formally,

$$B_i(x_i + \xi) = y_i + |\xi|^s, \quad \forall \xi \in [-h, h],$$

and

$$S_i(x_i+\xi)=\begin{cases} y+|\xi|^s, & \text{if } \xi\in[-h,0],\\ y-|\xi|^s, & \text{if } \xi\in[0,h],\end{cases}$$

where $y_i=\frac{1}{2}+2ih^s$.

Now we are ready to construct the reward functions using these gadgets. For $i=1,...,m/2$, scanning from prices high to low, the reward function R_i is a concatenation of $(i-1)$ consecutive S-curves, followed by one B curve, and finally a curve extending downwards the left portion of B until reaching the x-axis. Formally for any $i=1,...,\frac{m}{2}$,

$$R_i(x)=\begin{cases} S_j(x), & \text{if } x\in[x_j-h,x_j+h] \text{ for } j\le i-1,\\ B_i(x), & \text{if } x\in[x_i-h,x_i+h],\\ sh^{s-1}x+\left(y_i-h^s-sh^{s-1}\left(x_i-h\right)\right) & \text{if } x\le x_i-h.\end{cases}$$

Finally, we need a special reward function R_0, that consists only of S-curves on $[\frac{1}{2},1]$, and extends upwards when the prices moves below $\frac{1}{2}$, as shown in Figure xxx. Formally,

$$R_0(x)=\begin{cases} R_m(x), & \text{if } x\ge 1/2,\\ y_m+(x_m-x)^s, & \text{if } x\in[0,1/2].\end{cases}$$

The lower bound is again showed using the Wald-Wolfowitz Theorem (Theorem 10). At a high level, any reasonable policy π needs to solve a hypothesis testing problem in each interval $[x_i-h,x_i+h]$, which aims at distinguishing between R_0 and R_i. Note that R_0 and R_i are completely identical on prices higher than x_i, and only starts to differ on prices lower than x_i, at a rate of h^s. Hence, the maximum KL divergence on this interval is $\sim h^{2s}$, and by the Wald-Wolfowitz Theorem (Theorem 10), in expectation $\Omega(h^{-2s})$ samples are necessary assuming the policy π is able to distinguish between these two reward functions.

To see why π *must* be able to distinguish between the two curves, for the sake of contradiction, suppose otherwise, say, under true reward curve R_i the policy π has a high probability of mistakenly return R_0 as the true curve, and hence reduces the price below x_i-h, incurring an $\Omega(h^s)$ regret per round. This leads to an $\Omega(h^sT)=\Omega(T^{\frac{2s+1}{3s+1}})$ regret in future rounds, contradicting the low-regret assumption (with suitably chosen constants). Since the number of intervals is $\Omega(h^{-1})$, we have

$$\mathrm{Reg}(\pi,R_0)\ge\Omega(h^{-2s})\cdot\Omega(h^{-1})=\Omega(h^{-1-2s})=T^{\frac{2s+1}{3s+1}},$$

and the proof follows. $\square$

References

Agrawal R (1995) The continuum-armed bandit problem. *SIAM journal on control and optimization* 33(6):1926–1951.

Aviv Y, Vulcano G (2012) Dynamic list pricing. *The Oxford handbook of pricing management*.

Babaioff M, Dughmi S, Kleinberg R, Slivkins A (2015) Dynamic pricing with limited supply. *ACM Transactions on Economics and Computation (TEAC)* 3(1):1–26.

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

Bastani H, Simchi-Levi D, Zhu R (2019) Meta dynamic pricing: Learning across experiments. *Available at SSRN 3334629* .

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

Bitran GR, Mondschein SV (1997) Periodic pricing of seasonal products in retailing. *Management science* 43(1):64–79.

Boyacı T, Özer Ö (2010) Information acquisition for capacity planning via pricing and advance selling: When to stop and act? *Operations Research* 58(5):1328–1349.

Broder J (2011) Online algorithms for revenue management .

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

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

Chen N (2021) Multi-armed bandit requiring monotone arm sequences. *arXiv preprint arXiv:2106.03790* .

Chen Y, Farias VF (2018) Robust dynamic pricing with strategic customers. *Mathematics of Operations Research* 43(4):1119–1142.

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

Combes R, Proutiere A (2014) Unimodal bandits: Regret lower bounds and optimal algorithms. *International Conference on Machine Learning*, 521–529.

Cope EW (2009) Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Transactions on Automatic Control* 54(6):1243–1253.

den Boer A, Keskin NB (2020) Dynamic pricing with demand learning and reference effects. *Available at SSRN 3092745* .

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

den Boer AV, Zwart B (2015) Dynamic pricing and learning with finite inventories. *Operations research* 63(4):965–978.

Dholakia UM (2021) If you're going to raise prices, tell customers why. *Harvard Business Review* .

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

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

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

Google (2021) Transforming specialty retail with ai. Technical report.

Gupta S, Kamble V (2019) Individual fairness in hindsight. *Proceedings of the 2019 ACM Conference on Economics and Computation*, 805–806.

Heching A, Gallego G, van Ryzin G (2002) Mark-down pricing: An empirical analysis of policies and revenue potential at one apparel retailer. *Journal of revenue and pricing management* 1(2):139–160.

Heidhues P, Kőszegi B (2014) Regular prices and sales. *Theoretical Economics* 9(1):217–251.

Hu Z, Chen X, Hu P (2016) Dynamic pricing with gain-seeking reference price effects. *Operations Research* 64(1):150–157.

Jia S, Li A, Ravi R (2021) Markdown pricing under unknown demand. *Available at SSRN 3861379* .

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

Kleinberg R, Leighton T (2003) The value of knowing a demand curve: Bounds on regret for online posted-price auctions. *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, 594–605 (IEEE).

Kleinberg RD (2005) Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems*, 697–704.

Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6(1):4–22.

Lei YM, Jasin S, Sinha A (2014) Near-optimal bisection search for nonparametric dynamic pricing with inventory constraint. *Ross School of Business Paper* (1252).

Liu Y, Cooper WL (2015) Optimal dynamic pricing with patient customers. *Operations research* 63(6):1307–1319.

Lobel I (2020) Dynamic pricing with heterogeneous patience levels. *Operations Research* .

Nasiry J, Popescu I (2011) Dynamic pricing with loss-averse consumers and peak-end anchoring. *Operations research* 59(6):1361–1368.

Papanastasiou Y, Savva N (2017) Dynamic pricing in the presence of social learning and strategic consumers. *Management Science* 63(4):919–939.

Perakis G, Singhvi D (2019) Dynamic pricing with unknown non-parametric demand and limited price changes. *Available at SSRN 3336949* .

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

Ramakrishnan R (2012) Markdown management. *The Oxford Handbook of Pricing Management*.

Salem J, Gupta S, Kamble V (2021) Taming wild price fluctuations: Monotone stochastic convex optimization with bandit feedback. *arXiv preprint arXiv:2103.09287* .

Smith SA, Achabal DD (1998) Clearance pricing and inventory policies for retail chains. *Management Science* 44(3):285–300.

Vershynin R (2018) *High-dimensional probability: An introduction with applications in data science*, volume 47 (Cambridge university press).

Wald A, Wolfowitz J (1948) Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics* 19(3):326–339.

Wang Z (2016) Intertemporal price discrimination via reference price effects. *Operations research* 64(2):290–296.

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

Wu S, Liu Q, Zhang RQ (2015) The reference effects on a retailer's dynamic pricing and inventory strategies with strategic consumers. *Operations Research* 63(6):1320–1335.

Yin R, Aviv Y, Pazgal A, Tang CS (2009) Optimal markdown pricing: Implications of inventory display formats in the presence of strategic customers. *Management Science* 55(8):1391–1408.

Yu JY, Mannor S (2011) Unimodal bandits. *ICML*, 41–48 (Citeseer).