Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and should not be used to distribute the papers in print or online or to submit the papers to another publication.

Markdown Pricing Under Unknown Demand

(Authors’ names blinded for peer review)

We consider the Unimodal Multi-Armed Bandit problem where the goal is to find the optimal price under an unknown unimodal reward function, with an additional “markdown” constraint that requires that the price exploration is non-increasing. This markdown optimization problem faithfully models a single-product revenue management problem where, given infinite inventory, the objective is to adaptively reduce the price over a finite sales horizon to maximize expected revenues.

We measure the performance of an adaptive exploration-exploitation policy in terms of the regret: the revenue loss relative to the maximum revenue that could have been attained when the demand curve is known in advance. For the case of $L$-Lipschitz-bounded unimodal revenue functions with infinite inventory, we present a natural policy that explores the price space at a uniform optimal speed in $T$ steps and has regret $O(T^{3/4}(L \log T)^{1/4})$. On the other side, we provide an almost-matching lower bound of $\Omega(L^{1/4}T^{3/4})$ on the regret of any policy. Further, under mild assumptions, we show that the above tight bounds also hold when the inventory is finite but is on the order of $\Omega(T)$. Our tight regret bounds highlight the additional complexity of the markdown constraint, and are asymptotically higher than the corresponding bounds without this markdown requirement of $\tilde{O}(T^{1/2})$ for unimodal bandits and $\tilde{O}(L^{1/3}T^{2/3})$ for $L$-Lipschitz bandits. We finally consider a generalization called Dynamic Pricing with Markup Penalty where the seller is allowed to increase the price by paying a markup penalty of magnitude $O(T^c)$ per markup where $c \in [0, 1]$ is a given constant. We extend our results to a tight $\tilde{O}(T^{\text{med} \{ \frac{2}{3}, \frac{3}{4}, c \}})$ regret bound for this variant*.

Key words: dynamic pricing; markdown pricing; multi-armed bandits; online learning

1. Introduction

Consider the problem of dynamic pricing under unknown demand. This problem is by now well-studied, and indeed “optimal” solutions exist under numerous variations on (a) the set of demand functions allowed, on (b) how inventory is treated, and on (c) the frequency at which prices are allowed to change, just to name a few. By and large, these problems are modeled as variants of

* $\text{med}\{a, b, c\}$ denotes the median of the numbers $a, b, c$. 
the classic *multi-armed bandit* problem, and optimality (with respect to a performance measure called *regret*) is achieved by striking a carefully-tuned balance between selecting prices to learn the unknown demand function (exploration), and prices to maximize revenue given what has previously been learned (exploitation).

Now a seemingly innocuous assumption made across all of this work, which appears to be critical in achieving meaningful results (i.e. sub-linear regret), is that the price is allowed to be both decreased (*marked down*) and increased (*marked up*). In reality, markdowns are quite common, but this treatment of markups as being common and harmless in fact stands in contrast to the *practice* of pricing, where it is well-understood that markups negatively impact customers’ perception of a product’s value. For example, as noted by Bitran and Mondschein (1997):

> “Customers will hardly be willing to buy a product whose price oscillates, from their point of view, randomly over the season...Most retail stores do not increase the price of a seasonal or perishable product despite the fact that the product is being sold successfully.”

For this reason, *markdown pricing* (i.e. where markups are not allowed) has long been ubiquitous in retail (Petro (2017)), and remains among the standard set of capabilities that retailers are still seeking to hone – a recent survey (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).

In short, despite the rich literature on dynamic pricing under unknown demand in recent years, a basic question remains open with respect to the salient challenge of markdown pricing: **Is it feasible to achieve any meaningful performance for markdown pricing under unknown demand, and if so, what is the “separation” from ordinary dynamic pricing?** Put another way, does a markdown constraint render dynamic pricing less “effective”, and if so, by how much? This work presents the first definitive answer to this basic question by providing an *optimal* policy for markdown pricing, which allows for a precise characterization of the separation between the regret bounds of markdown pricing and ordinary pricing.

### 1.1. Our Contributions.

We study a canonical pricing problem with an additional *markdown constraint*. Specifically, at each of $T$ discrete time periods, a price $x$ is chosen and a random demand is observed whose mean is given by an unknown demand function $D(x)$. The markdown constraint precisely means
that if price $x$ is selected at time period $t$, then the price at time period $t+1$ can be at most $x$. We place only minimal assumptions on the demand function: that the corresponding revenue function $R(x) = xD(x)$ be unimodal and Lipschitz (we will see later on that both are necessary), and inventory is assumed to be infinite (though we will quickly relax this assumption). The goal is to design a policy which minimizes regret (defined as the difference between the policy’s expected total revenue and the total revenue accrued by selecting the revenue-maximizing price at all $T$ periods).

Without the markdown constraint, this problem has previously been solved, and it has been shown that there exists a policy which achieves $O(T^{2/3})$ regret (Kleinberg (2005)). This policy selects a certain discrete subset of the prices and treats each price in this discretization as an “arm” in a classic multi-armed bandit problem. So in particular, many (approximately half) of the policy’s price changes are markups, and thus the introduction of the markdown constraint seems likely to (a) necessitate a different algorithmic approach, and (b) induce a “separation” in achievable performance as alluded to above.

Against this backdrop, we make the following contributions:

1. **A Markdown Policy and Performance Guarantee:** We introduce a policy which satisfies the markdown constraint, and show (via Theorem 1) that it achieves $\tilde{O}(T^{3/4})$ regret.\(^1\) This immediately answers the first part of our basic question affirmatively: we are able to achieve meaningful performance in the form of a sub-linear (in $T$) regret bound. Moreover, with small but non-trivial modifications to our policy and proof technique, we show that:

   (a) We can relax the assumption of infinite inventory and still achieve the same $\tilde{O}(T^{3/4})$ regret in the regime where the inventory scales as $\Omega(T)$; see Theorem 2.

   (b) Stronger regret guarantees can be obtained if more stringent restrictions are placed on the revenue function. For example, $\tilde{O}(T^{5/7})$ regret can be achieved under twice-differentiability of the revenue function; see Theorem 4.

2. **Optimality via a Minimax Lower Bound:** We prove that our policy is in fact order-optimal by showing (via Theorem 3) that the regret of any policy is at least $\Omega(T^{3/4})$. This answers the second part of our question: the separation between markdown and ordinary pricing is precisely that markdown pricing must incur at least $\Omega(T^{3/4})$ regret, whereas ordinary pricing can achieve $\tilde{O}(T^{2/3})$ regret.

\(^1\) We use $\tilde{O}$ to hide logarithmic terms.
Our proof uses a novel generalization of the classic Wald-Wolfowitz Theorem for hypothesis testing, which may be of independent interest for proving lower bounds for a broader class of online learning problems. As an example, we present an alternate (and perhaps simpler) proof of the $\Omega(\sqrt{KT})$ lower bound for the classic $K$-armed multi-armed bandit using our technique.

3. **Model Extension with Penalized Markups:** A natural generalization of our model would be one in which markups are allowed, but penalized. While a *complete* treatment of dynamic pricing with penalized markups would be substantial (indeed, we will see that even the choice of how to model these penalties is not obvious), we initiate this future direction of research by considering one (novel) version in which each markup incurs a fixed, known, additive cost that scales as $\Theta(T^c)$, for some $c \in [0, 1]$. We provide a complete solution for this model, showing that:

(a) A simple variant of the Successive Elimination Policy (applied on a suitable discretization of the price space) achieves $\tilde{O}(T^{\text{mod}(\frac{2}{3}, \frac{1}{4})})$ regret; see Theorem 6.

(b) This bound is optimal up to logarithmic factors; see Theorem 7.

These results completely characterize the manner in which our penalized markup model interpolates between ordinary pricing and markdown pricing. When the markup penalty is sufficiently low ($c \leq 2/3$), there is effectively no penalty for markups, since the achievable regret matches that for ordinary pricing. This is already quite surprising – for example, one corollary to this is that any sort of one-time or constant-sized penalty is an insignificant detractor to marking up (using carefully-constructed policies). When the penalty is sufficiently high ($c \geq 3/4$), this effectively imposes the hard markdown constraint, as it is optimal to never markup, and the resulting regret matches that for markdown pricing. Finally, the optimal regret interpolates smoothly between these two regimes for $c \in \left[\frac{2}{3}, \frac{3}{4}\right]$.

4. **Experimental Evaluation:** We test our policy on two of the most commonly-used families of demand functions, comparing against natural benchmarks designed specifically for these families. These experiments establish:

(a) Fast convergence rate of regret: compared to an explore-then-commit (ETC) type policy which knows the specific functional form of the demand function, the regret of our policy vanishes at a decent speed.

(b) Robustness to model misspecification: our policy has vanishing regret on various families of demand functions, whereas an ETC-type policy may incur non-vanishing regret when it assumes an incorrect demand model.

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, assumptions, policies, and core
results in Section 2. The proofs of our upper and lower regret bounds are given in Sections 3 and 4, respectively. Section 5 introduces our model and results for penalized markups. Experiments are described in Section 6, and conclusions are drawn in Section 7.

1.2. Previous Work

The present work falls into two primary streams of work: dynamic pricing in operations management, 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). Table 1 summarizes the most related works along these dimensions.

**Dynamic Pricing:** 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 (in the sense that the prices only depends on time, but not on the realized demands) and whose expected sales is at most the inventory level. They presented an algorithm which achieves nearly optimal regret bounds. Wang et al. (2014) improved their results.

Cheung et al. (2017) studied a dynamic pricing problem over a finite selling season $[T]$ with infinite inventory, and the unknown demand function arises from a discrete set of demand functions.

<table>
<thead>
<tr>
<th></th>
<th>Regret</th>
<th>Parametric</th>
<th>Markdown</th>
<th>$\infty$-Inv.</th>
<th>Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith and Achabal (1998)</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Kleinberg and Leighton (2003)</td>
<td>✓</td>
<td>No</td>
<td></td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Besbes and Zeevi (2009)</td>
<td>✓</td>
<td>Both</td>
<td></td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Wang et al. (2014)</td>
<td>✓</td>
<td>Both</td>
<td></td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Broder and Rusmevichientong (2012)</td>
<td>✓</td>
<td>Yes</td>
<td></td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Harrison et al. (2012)</td>
<td>No</td>
<td></td>
<td>✓</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Yin et al. (2009)</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Cheung et al. (2017)</td>
<td>✓</td>
<td>No</td>
<td></td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Ferreira et al. (2018)</td>
<td>✓</td>
<td>Yes</td>
<td></td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Combes and Proutiere (2014)</td>
<td>✓</td>
<td>No</td>
<td></td>
<td></td>
<td>No</td>
</tr>
<tr>
<td><strong>This work</strong></td>
<td>✓</td>
<td>No</td>
<td>✓</td>
<td></td>
<td>No</td>
</tr>
</tbody>
</table>

Table 1 Comparison of our work with prior related work along important model dimensions: whether or not (1) the metric used is regret; (2) the given family of demand/revenue curve is parametric; (3) the markdown constraint is considered, (4) infinite inventory is assumed and (5) a prior over the demand family, over which the Bayesian regret is considered, is given.
For any $m$, they showed a policy that is allowed to change price at most $m$ times with regret $O(\log^m(T))$ where $\log^m$ is the $m$-th iteration of the logarithm function. Ferreira et al. (2018) considered a price-based network version. Babaioff et al. (2015) and Badanidiyuru et al. (2013) consider the case with finite inventory.

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), Auer et al. (2002)). 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 $R(p)$.

Kleinberg and Leighton (2003) studied a revenue maximization problem for a seller with an unlimited supply of identical goods, interacting sequentially with a population of buyers through an online posted-price auction mechanism. Each customer has an unknown valuation of the good: if it is higher than the posted price, they buy; otherwise they leave. They assumed the buyers’ valuations are i.i.d. drawn from a fixed distribution on $[0,1]$. They obtained an upper bound on the regret of $\tilde{O}(pT)$ by discretizing the price space and reducing to the classic multi-armed bandits problem. An $\Omega(\sqrt{t})$ lower bound was showed under mild smoothness conditions.

In the Lipschitz Bandit problem (e.g. Agrawal (1995), Kleinberg (2005), Kleinberg et al. (2008)), it is assumed that each $x \in [0,1]$ corresponds to an arm with mean reward $\mu(x)$, and $\mu$ satisfies the Lipschitz condition, i.e. $|\mu(x) - \mu(y)| \leq L|x - y|$ for some constant $L > 0$. Kleinberg (2005) proved a tight $\tilde{O}(T^{2/3})$ regret bound for Lipschitz Bandits. The lower bound was proved by considering a family of “bump curves”: each curve is $1/2$ at all arms except in a small neighborhood of the “peak”, where the mean reward is slightly higher than $1/2$. Since these bump curves are unimodal, this lower bound carries over to the family we study.

The most closely-related variant of MAB is the Unimodal Bandits problem (Cope (2009), Yu and Mannor (2011), Combes and Proutiere (2014)). In addition to the assumptions for the Lipschitz bandits problem, the reward function $\mu : [0,1] \rightarrow [0,1]$ is assumed to be unimodal. It is also assumed

\[\text{Notation } \tilde{O} \text{ hides logarithmic terms in } T.\]
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 $\tilde{O}(\sqrt{T})$.

Another related variant is MAB with movement costs (Guha et al. (2010), Dekel et al. (2014), Koren et al. (2017)). In this variant, in addition to the settings in Lipschitz bandits, there is a given cost for moving between two arms (numbers). Finally, 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), Liu and Cooper (2015), Farias and Van Roy (2010), Lobel (2020), Qiang and Bayati (2016), Papanastasiou and Savva (2017), den Boer and Zwart (2015).

**Other Related Work:** Surveys by Aviv and Vulcano (2012), den Boer and Zwart (2015) and den Boer and Zwart (2013) provide a comprehensive overview of dynamic pricing. Gallego and Van Ryzin (1994) characterized the optimal pricing policy when the demand function is known. Boyacı and Özer (2010) considered a profit-maximization model in which a manufacturer collects advance sales information periodically prior to the regular sales season for a capacity decision. Smith and Achabal (1998) developed optimal clearance prices and inventory management policies that take into account the impact of reduced assortment and seasonal changes on sales rates. They tested these policies at three major retail chains and these applications are summarized and discussed. Heching et al. (2002) analyzed sales and price data from a speciality retailer of women’s apparel in spring 1993. They fitted a demand model to the data, and then obtain estimates of rewards under various pricing policies. The results also indicate that model-based pricing schemes can potentially increase revenue by approximately 4 percent.

### 2. Model

We begin by formally stating our model. Given inventory $I > 0$ and a discrete time horizon of $T$ rounds, in each round $t$, the policy (representing the “seller”), selects a price $x_t \in [0, 1]$ (the particular interval $[0, 1]$ is without loss of generality, by scaling). This round’s demand $d_t$ 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 (up to the smaller of the demand and remaining inventory):

$$\min\{d_t, I - (d_1 + \cdots + d_{t-1})\}x_t.$$

For simplicity, we will assume that the random demand $d_t$ is almost surely bounded, specifically in $[0, 1]$ (again, w.l.o.g.), though our results can be easily extended to sub-Gaussian distributions. 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. A demand function \( D(x) \) is naturally associated with a revenue function \( R(x) = xD(x) \). For most of this paper, we will deal directly with revenue functions, which we term more generally as reward functions.\(^3\) For any policy \( \mathcal{A} \),\(^4\) reward function \( R(\cdot) \), and inventory \( I \), we use \( r(\mathcal{A}, R, I) \) to denote the expected total reward of \( \mathcal{A} \) under \( R \) with initial inventory \( I \).

Rather than evaluating policies directly in terms of \( r(\mathcal{A}, R, I) \), 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, a Fixed Price Policy (FPP) selects the same price at each round, i.e. \( x_1 = \cdots = x_T = p \) for some \( p \). Let \( FPP(p) \) denote the FPP at price \( p \). We use \( \text{OPT}_R \) to denote the maximum achievable expected reward among all FPPs, i.e.

\[
\text{OPT}_R := \max_{p \in [0,1]} r(FPP(p), R, I).
\]

So for example, when \( I = \infty \), we have that \( \text{OPT}_R = r^*T \), where \( r^* = \max_{x \in [0,1]} R(x) \). 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).** Let \( \mathcal{F} \) be a family of reward functions, each a mapping from \([0, 1]\) to \([0,1]\). For any policy \( \mathcal{A} \) and \( R \in \mathcal{F} \), define the regret of policy \( \mathcal{A} \) under \( R \) to be

\[
\text{Reg}(\mathcal{A}, R, I) := \text{OPT}_R - r(\mathcal{A}, R, I).
\]

The worst-case regret of policy \( \mathcal{A} \) for family \( \mathcal{F} \) is

\[
\text{Reg}(\mathcal{A}, \mathcal{F}, I) := \sup_{R \in \mathcal{F}} \text{Reg}(\mathcal{A}, R, I).
\]

To summarize, the problem we seek to solve is: given a family \( \mathcal{F} \) of reward functions from \([0, 1]\) to \([0,1]\) and initial inventory \( I \), design a policy \( \mathcal{A} \) that satisfies the markdown constraint and that minimizes \( \text{Reg}(\mathcal{A}, \mathcal{F}, I) \). We will be particularly concerned with how a policy’s regret scales with the time horizon \( T \) – at the very least, we aim for sub-linear (i.e. \( o(T) \)) regret.

It is worth pausing here to note that our definition of regret has different implications when \( I \) is either infinite or finite. When \( I \) is infinite, the best offline policy (meaning one that knows the

---

\(^3\) The corresponding demand function can naturally be backed out from a reward function: \( D(x) = R(x)/x \) for \( x > 0 \).

\(^4\) For the sake of completeness, a policy is, formally, a time-indexed sequence of functions \( \mathcal{A} = \{\mathcal{A}_t : ([0,1] \times [0,1])^{t-1} \rightarrow [0,1], t = 1, \ldots, T\} \), where each function \( \mathcal{A}_t \) maps the prices selected and demands observed over the previous \( t - 1 \) rounds to a price for round \( t \).
reward function $R$) is precisely a fixed price policy, so regret here is really measured against the best offline policy (this is the “typical” definition of regret). However, when $I$ is finite, the best offline policy need not be a fixed price policy, and moreover even calculating the best offline policy for a given reward function can be non-trivial – in general, the policy can at best be characterized as the solution to a dynamic program (see Talluri and Van Ryzin (2004)). Thus, we measure regret only against fixed price policies. One reason this is fairly innocuous is that for the inventory regime we consider ($I = \Omega(T)$), the best fixed price policy is asymptotically optimal (as $T$ grows) in a manner that can be made formal.

2.1. Assumptions on the Reward Function

We have so far made just one assumption: that the random demands are bounded (and even this can be relaxed to sub-Gaussianity). We will, in addition, require two assumptions on the underlying reward function:

1. **Lipschitz**: The reward function is $L$-Lipschitz, i.e. $|R(x) - R(x')| \leq L|x - x'|$ for all $x, x' \in [0, 1]$. This assumption is standard for the version of our problem without markdown constraint (e.g. Kleinberg (2005)). Note, as an aside, that we are implicitly assuming here that $L$, or at least an upper bound on $L$ across the entire family of reward functions, is known.

2. **Unimodal**: The reward function is unimodal, i.e. there exists $x^* \in [0, 1]$ s.t. $R$ is non-increasing on $[x^*, 1]$ and non-decreasing on $[0, x^*]$. This assumption has also previously appeared for the non-markdown version of our problem (e.g. Yu and Mannor (2011)).

In addition to having appeared previously in the literature, both of these assumptions are in fact necessary for achieving sub-linear regret. Specifically, the Lipschitz assumption is necessary in the sense that there exists a family $\mathcal{F}$ of unimodal reward functions, whose Lipschitz constants are arbitrarily large, such that for any policy $A$, its regret is $\Omega(T)$ under some $R \in \mathcal{F}$.\footnote{One example family is $\mathcal{F} = \{R_c(x) = (-(x - 1)(x - 1 + 2c)/c^2)^+: c \in (0, 1/2)\}$.} The unimodal assumption is similarly necessary: there exists a family of $L$-Lipschitz reward functions such that any policy has regret $\Omega(T)$ under some function in the family.\footnote{Such a family can be constructed with just two reward functions: one with a single mode, and one with two modes.}

Finally, these two assumptions hold for the reward functions corresponding to some of the most commonly-used parametric families of demand functions:

1. **Linear Demand**: $\{D_{a,b}(x) = a - bx : 1 \geq a \geq b \geq 0\}$

2. **Exponential Demand**: $\{D_{a,b}(x) = e^{a-bx} : a \in \mathbb{R}, b \in \mathbb{R}_+\}$
These examples serve to illustrate that our assumptions are mild enough to allow for realistic models of demand, though we emphasize that our policies and results will not require that the reward functions be parameterizable.

2.2. Our Policies

We can now state our policies, beginning with the setting of infinite inventory (which captures the crux of the challenge), and then finite inventory.

Infinite Inventory. Our policy under infinite inventory operates under a simple idea: begin at the highest price, and decrease the price at a constant rate, stopping when the mode (or “peak”) of the reward function is detected, i.e. when the mean reward at the current price is significantly lower than some previous price. Intuitively, this idea should perform well as long as the rate of price decrease is neither too slow (or else the policy will spend too much time at sub-optimal prices before reaching the peak) nor too fast (or else the policy will not gather sufficient information to correctly identify when the peak is reached).

Our actual policy, dubbed the Uniform Elimination Policy (UEₙ,ₚ), implements a discretized version of this idea. As described in Algorithm 1, the policy is parameterized by two values: a step size s and a confidence interval width δ. Each price decrease is exactly of size s, and rather than

**Algorithm 1 Uniform Elimination Policy (UEₙ,ₚ).**

1: Input: \( s, \delta > 0 \). \( \triangleright \) Step size and width of target confidence intervals.

2: Initialize: \( x \leftarrow 1, \text{LCB}_{\text{max}} \leftarrow 0, k \leftarrow [3\delta^{-2}\log T] \).

3: while \( x > 0 \) do \( \triangleright \) Exploration phase starts.

4: Select price \( x \) for the next \( k \) rounds and observe rewards \( X_1,...X_k \).

5: \( \bar{\mu} \leftarrow \sum_{i=1}^{k} X_i \). \( \triangleright \) Compute mean rewards.

6: \( [\text{LCB}, \text{UCB}] \leftarrow [\bar{\mu} - \delta, \bar{\mu} + \delta] \). \( \triangleright \) Compute confidence interval for reward at current price.

7: if \( \text{LCB} > \text{LCB}_{\text{max}} \) then \( \triangleright \) Update best LCB so far

8: \( \text{LCB}_{\text{max}} \leftarrow \text{LCB} \).

9: if \( \text{UCB} < \text{LCB}_{\text{max}} \) then \( \triangleright \) Exploration phase ends.

10: \( x_h \leftarrow x \). Break. \( \triangleright \) Define halting price.

11: else \( x \leftarrow s \).

12: Select price \( x_h \) in all future rounds. \( \triangleright \) Exploitation phase.
decreasing each round, our policy remains at a price $x$ long enough that $R(x)$ can be estimated up to an additive error of at most $\delta$ with high confidence (via Hoeffding/Chernoff bound). The policy “halts” when the confidence interval at the current price lies completely below that of a previous price, indicating that we have likely “overshot” the optimal price.

As we will see in the next subsection, this policy is order-optimal (up to logarithmic factors) for certain values of $s$ and $\delta$. It is important to note that these “correct” choices of $s$ and $\delta$ depend on $L$ and $T$, meaning our policy itself requires knowledge of $L$ and $T$. The knowledge of $L$ is standard (see e.g. Kleinberg (2005)), and further, in practice, one may simply choose the maximum $L$ of fitted demand functions from past sales data as an upper bound for the Lipschitz constant of the unknown demand model. Knowledge of $T$ is more delicate. In the literature on MAB, one of the primary challenges (and successes) has been in designing so-called anytime policies, which achieve order-optimal regret without knowledge of $T$. One could ask if this is possible here – this is in fact impossible when the markdown constraint is present, a result we state and prove formally (see Proposition 3 in Appendix F).

**Finite Inventory.** Now assume that inventory is finite, and let $\rho := I/T$ be the inventory-to-time ratio. A simple observation allows us to modify the previous Uniform Elimination Policy for finite inventory. Fix a reward function $R$ (and corresponding demand function $D$), and let $p^*$ be the location of its peak: $p^* \in \arg \max_x R(x)$. Let $p_d$ be the depletion price, meaning the value satisfying $D(p_d) = \rho$ (we assume its existence just for convenience of discussion). This is the price at which, if the demands were not random, our inventory $I$ would be perfectly depleted after exactly $T$ rounds.

The simple observation is that $\max\{p^*, p_d\}$ is approximately the best fixed price (we show this formally in Section 3.2). Intuitively, if $p_d < p^*$, then there is effectively enough inventory to ignore the inventory constraint. If $p_d > p^*$, offering a price lower than $p_d$ is sub-optimal because the same number of units would get sold (i.e. all of them) at a lower price, and offering a price higher than $p_d$ is sub-optimal because $R(x)$ is decreasing for $x \geq p^*$.

Our Depletion-Aware Uniform Elimination Policy (DUE$_{s,\delta}$), described in Algorithm 2, adapts the Uniform Elimination Policy based on this observation. In particular, the Uniform Elimination Policy already seeks to decrease the price as quickly as possible until $p^*$ is reached. An extra subroutine which tracks the rate at which inventory is depleted ensures that price decreases are halted if $x_d$ is reached.
Algorithm 2 Depletion-Aware Uniform Elimination Policy (DUEs,δ).
1: Input: s, δ > 0. ▷ Step size and width of target confidence intervals.
2: while x > 0 do ▷ Exploration Phase.
3: Select price p in the next k = ⌈3δ−2 log T⌉ rounds (stop when inventory runs out), and observe demands X₁,...,Xₖ.
4: Set $\bar{d} = \frac{1}{k} \sum_{i=1}^{k} X_i$ ▷ Estimate D(x).
5: Set [LCB, UCB] ← [x$\bar{d}$ − δ, x$\bar{d}$ + δ] ▷ Compute confidence interval.
6: if LCB ≥ LCB max then ▷ Keep track of the highest LCB.
7: LCB max ← LCB.
8: if UCB(x) < LCB max or (x$\bar{d}$ + δ)T ≥ I then ▷ Termination condition.
9: $x_h \leftarrow x$. Break. ▷ Exploration halts.
10: else
11: $x \leftarrow x - s$. ▷ Reduce the price by s.
12: Use price $x_h$ for all future rounds ▷ Exploitation phase.

2.3. Our Results

The core results of this paper are a set of matching upper and lower regret bounds for markdown pricing, whose ideas also lay the cornerstone for our extensions. Throughout, we use $\hat{F}_L$ to denote the family of L-Lipschitz, unimodal functions from [0,1] to [0,1]. Our first result is an upper bound on the regret of our Uniform Elimination policy for the infinite inventory setting:

**Theorem 1 (Upper Bound, Infinite Inventory).** For any given $L > 0$ and $T \in \mathbb{N}$, the Uniform Elimination policy $\text{UE}_{s,\delta}$ satisfies

$$\text{Reg}(\text{UE}_{s,\delta}, \hat{F}_L, I = \infty) = O(T^{3/4}(L \log T)^{1/4}),$$

for $\delta = \sqrt{2}T^{-1/4}(L \log T)^{1/4}$ and $s = \delta/2L$.

The most immediate conclusion from Theorem 1 is that it establishes concretely that sub-linear regret is achievable for markdown pricing.

In practice, the initial inventory I is usually finite. Recall that $\rho := I/T$ is the inventory-time ratio. Since the range of demand functions are normalized to [0,1], the seller can sell at most T units in T rounds, so the problem reduces to the $I = \infty$ case if $\rho \geq 1$. On the other extreme, if $I = o(T)$, suppose the mean demand at $p = p_{\text{max}} = 1$ is non-zero, then for any $p \in [0,1]$ the FPP
is likely to sell out all units, so the optimal seller should select $p = 1$ in all rounds. Thus, the interesting scenario is $1 > \rho = \Omega(1)$.

**Theorem 2 (Upper Bound, Finite Inventory).** Given any $L, T, I > 0$ where $\rho = I/T = \Omega(1)$ (but not necessarily greater than 1), the Depletion-Aware Uniform-Elimination Policy $\text{DUE}_{s, \delta}$ with $s = \delta = T^{-1/4}$ satisfies

$$\text{Reg}(\text{DUE}_{s, \delta}, \mathcal{F}_L, I) = O(T^{3/4}(L \log T)^{1/4}).$$

To show a lower bound, we need to specify a family of problem instances on which any markdown policy suffers a such amount of regret. An $\Omega(L^{1/3}T^{2/3})$ lower bound for markdown pricing with $I = \infty$ is implied by the lower bound for Lipschitz bandits (Kleinberg (2005)), since the “bump curves” they used are unimodal and Lipschitz. Despite its low-adaptivity, the DUE policy surprisingly achieves the best possible regret among all deterministic policies, including adaptive ones:

**Theorem 3 (Lower Bound).** There is a family $\mathcal{M} \subset \mathcal{F}_L$ of reward functions s.t. any markdown policy $\mathcal{A}$ (that knows $L, T$) admits regret $\text{Reg}(\mathcal{A}, \mathcal{M}, I = \infty) = \Omega(L^{1/4}T^{3/4})$.

**Generalized Wald-Wolfowitz Theorem (GWW).** The existing lower bound techniques for MAB fail to address the extra complexity caused by the markdown constraint. We develop a novel technique to address this challenge and obtain Theorem 3. We generalize the classic Wald-Wolfowitz Theorem (WW) for sequential hypothesis testing from testing between point estimates to testing between appropriately defined intervals. 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.

<table>
<thead>
<tr>
<th></th>
<th>Upper Bound</th>
<th>Lower Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unimodal Bandits</td>
<td>$O(\sqrt{T})$</td>
<td>Unknown</td>
</tr>
<tr>
<td>Lipschitz Bandits</td>
<td>$O(T^{2/3}(L \log T)^{1/3})$</td>
<td>$\Omega(T^{2/3}L^{1/3})$</td>
</tr>
<tr>
<td>Markdown Pricing</td>
<td>$O(T^{3/4}(L \log T)^{1/4})$</td>
<td>$\Omega(T^{3/4}L^{1/4})$</td>
</tr>
</tbody>
</table>

Table 2 Distribution independent regret bounds for markdown pricing and related bandit problems. Our results (in red) are presented in Theorem 1 and 3. The lower bound for Lipschitz bandits is from Kleinberg (2005). Note that all but the last row hold for both known or unknown $T$. 
Table 2 summarizes the relevant previous results, along with our new results. The key observation is, with the markdown constraint, the regret bounds increase significantly, which matches our intuition. For unimodal bandits, one can apply a natural binary search type algorithm (see Yu and Mannor (2011)) to localize a mode with arbitrary precision. However, such a policy cannot be extended to our setting due to the markdown constraint.

3. Proof of Upper Bounds

3.1. Infinite Inventory: Proof of Theorem 1

Theorem 1 is immediately implied by the following lemma when \( \delta = \sqrt{2T^{-1/4}(L \log T)^{1/4}} \) and \( s = \delta / 2L \), which minimize the term inside big-O in (1). (In fact, let \( g(s, \delta) = (\delta + sL)T + s^{-1}\delta^{-2}\log T \). Then, \( \frac{\partial g}{\partial s} = T - 2\delta^{-3}s^{-1}\log T \) and \( \frac{\partial g}{\partial \delta} = LT - s^{-2}\delta^{-2}\log T \). Hence, \( \nabla g = 0 \iff s \delta^3 = \frac{2\log T}{T} \) and \( s^2 \delta^2 = \log T / LT \). One can then easily verify that our choices of \( s, \delta \) satisfies the above condition and is indeed a global minimum of \( g \).

**Lemma 1.** For any \( s, \delta > 0 \), it holds that

\[
\text{Reg}(\text{UE}_{s, \delta}, \hat{F}_L, I = \infty) = O((\delta + sL)T + s^{-1}\delta^{-2}\log T).
\]

**Proof** Let \( R \) be the true reward function with an optimal price \( p^* \). Define sample prices \( x_i = 1 - si \) for \( i \leq s^{-1} \). We use the following lemma that follows from standard tail bounds.

**Lemma 2.** Define \( C \) to be the event that \( R(x) \in [\mu - \delta, \mu + \delta] \) for any of the prices \( x \) selected by the policy \( \text{UE}_{s, \delta} \) with \( s \geq 1/T \). Then the probability of its complement \( P[\bar{C}] = O(T^{-1}) \).

Furthermore, when the policy halts at \( x_h \), the confidence interval for the reward at the price just before the halting price overlapped with that of the best confidence interval found so far centered around the reward at \( p^* \). Hence, the reward at the halting price is at most twice the length of a confidence interval, i.e. \( 4\delta \). This gives the following lemma (whose proof is in Section A).

**Lemma 3.** If \( x_h \) denotes the halting price of \( \text{UE}_{s, \delta} \) on an \( L \)-Lipschitz unimodal reward function \( R \in \hat{F}_L \), then conditional on event \( C \), \( R(x_h) \geq R(p^*) - 4\delta - 2sL \).

It follows that the regret in the exploitation phase is \( O((\delta + sL)T) \). On the other hand, in each case, there are \( O(\delta^{-2}\log T) \) explorations per price for up to \( s^{-1} \) different prices, giving a total of \( O(s^{-1}\delta^{-2}\log T) \) rounds for exploration. Lemma 1 is proved by combining the regret in the exploitation and exploration phases. \( \square \)

---

7 All asymptotic bounds with \( L \) are w.r.t. both \( L \) and \( T \). For example, our upper bound for markdown pricing policy \( A \) shows that there are constants \( C, T_0, L_0 > 0 \) s.t. for any \( L \geq L_0, T \geq T_0 \), we have \( \text{Reg}(A, \hat{F}) \leq CT^{3/4}(L \log T)^{1/4} \).
Smooth Reward Functions. In the above proof, we used the fact that if the exploitation price is \( \varepsilon \) distance away from the optimal price, then an \( O(\varepsilon) \) regret is incurred per round. If we assume the second derivatives of each reward functions exist, then by Taylor expansion,

\[
|R(x) - R(p^*)| = R'(p^*) + \frac{1}{2} R''(p^*)|x - p^*|^2 + o(|x - p^*|^2) \\
\leq 0 + \frac{C}{2} |x - p^*|^2 + o(|x - p^*|^2) = \left( \frac{C}{2} + o(1) \right) \cdot |x - p^*|^2, \quad \text{as } |x - p^*| \to 0.
\]

In other words, an \( \varepsilon \)-error in the estimation of optimal price only incurs regret \( O(\varepsilon^2) \) per round. This suggests that the \( \tilde{O}(T^{3/4}) \) upper bound may be improved for smooth reward functions. We first formally define the family under consideration.

**Definition 2 (Smooth Reward Functions).** Given \( L, C > 0 \), define \( \tilde{F}_{L,C} \) to be the family of all unimodal \( L \)-Lipschitz twice-differentiable reward functions whose second-derivatives are bounded by a common constant \( C \) in absolute value.

It turns out that such an improvement can be achieved by simply choosing a different combination of \( s, \delta \) in the UE policy. In fact, the proof is almost identical to the analysis above except that the term \( s \) in Lemma 3 can now be strengthened to \( O(s^2) \), as formally stated below.

**Lemma 4.** If \( x_h \) denotes the halting price of UE on a reward function \( R \in \tilde{F}_L \), then conditional on event \( C \), \( R(x_h) \geq R(p^*) - (2C + L)s^2 - 12\delta \).

Assuming this lemma (proof can be found in Appendix A), the regret in the exploitation phase becomes \( (s^2 + \delta)T \), so the total regret is now \( O(s^{-1}\delta^{-2}\log T + (s^2 + \delta)T) \). Choosing \( s \sim T^{-1/7} \) and \( \delta \sim T^{-2/7} \), we obtain an \( \tilde{O}(T^{5/7}) \) regret bound.

**Theorem 4 (Upper Bound for Smooth Reward Functions).** For any \( L, C > 0 \), with \( s = (L + C)^{-3/7}T^{-1/7} \log^{1/7} T \) and \( \delta = (L + C)^{1/7}T^{-2/7} \log^{1/7} T \), we have

\[
\text{Reg}(\text{UE}_s, \tilde{F}_{L,C}, I = \infty) = O((C + L)^{1/7}T^{5/7} \log^{2/7} T).
\]

### 3.2. Finite Inventory: Proof of Theorem 2

The first hurdle for showing Theorem 2 is that the optimal fixed price no longer enjoys a clean expression. Recall that when \( I = \infty \), for any reward function \( R \), the optimal FPP simply selects any \( p^* \in \arg\max_{x \in [0,1]} R(x) \). However, this is no longer true when \( I < \infty \). In fact, the optimal fixed price \( p_{OPT} \) has two equivalent characterizations.

**Characterization of the Optimal FPP.** It would be more convenient in this section to work with the demand functions (rather than the reward functions). By abuse of notation, write \( r(\mathcal{A}, D) \)
the expected reward of a policy $\mathcal{A}$ under a demand function $D$. Similarly define $r(p, D)$ as the expected reward of $\text{FPP}(p)$.

Consider the FPP at some price $p \in [0, 1]$. Let $\{X_t\}_{t \in [T]}$ be i.i.d. samples at price $p$ from demand function $D$, which by definition satisfy $D(p) = \mathbb{E}X_t$ for all $t \in [T]$.

- **Characterization 1:** Define the (random) depletion time $\tau_p$ to be the round in which inventory depletes, i.e. $\tau_p = \min\{t : \sum_{j=1}^t X_j \geq I\}$. Then by Wald’s identity (see e.g. Mitzenmacher and Upfal (2017)), the reward of $\text{FPP}(p)$ is

$$r(p, D) = \mathbb{E}[(\sum_{i=1}^{\tau_p} p \cdot X_i) = p \cdot \mathbb{E}[\tau_p] \cdot \mathbb{E}[X_i] = \mathbb{E}[\tau_p] \cdot R(p),$$

where $R(p) = D(p) \cdot p$ is the reward function for $D$. Thus, $p_{\text{OPT}} = \arg \max_{p \in [0,1]} \{\mathbb{E}[\tau_p] \cdot R(p)\}$.

- **Characterization 2:** Define the (random) sales to be $N_p = \min\{I, \sum_{t=1}^T X_t\}$. Then, $r(p, D) = \mathbb{E}[N_p] \cdot p$, and hence $p_{\text{OPT}} = \arg \max_{p \in [0,1]} \{\mathbb{E}[N_p] \cdot p\}$.

In neither characterizations, it is hard to derive a simple and precise expression for $p_{\text{OPT}}$. Fortunately, we can find a simple surrogate optimal price whose reward well approximates that of the optimal policy. To this aim, we first introduce the notion a depletion price, the price at which the inventory is perfectly depleted at the end of the time horizon, if the demands were deterministic.

**Definition 3 (Depletion Price).** The depletion price $p_d$ of a strictly decreasing demand function $D$ is the unique price $p$ such that $D(p) = \text{med}\{D(0), D(1), \rho\} \in (0, 1)$.

**Definition 4 (Surrogate Optimal Price).** The surrogate optimal price (SOP) for a demand function $D$ is

$$p_{\text{SOP}} = \max\{p_d, p^*\} = \begin{cases} p^*, & \text{if } p^* \geq p_d, \\ p_d, & \text{if } p^* < p_d. \end{cases}$$

**Lemma 5 (SOP is almost optimal).** Let $p_d$ be the depletion price of an $L$-Lipschitz demand function $D : [0, 1] \rightarrow [0, 1]$. Then for any $\Delta > 0$,

$$r(p_{\text{OPT}}, D) - r(p_{\text{SOP}}, D) \leq \begin{cases} (\Delta + e^{-\Omega(\Delta^2 I)}) \cdot T, & \text{if } p^* \geq p_d \\ (\Delta + e^{-\Omega(\Delta^2 I)}) \cdot I, & \text{if } p^* < p_d. \end{cases}$$

We defer the formal proof to Appendix B but outline the proof here. In the first case, $p^* \geq p_d$, $\text{FPP}(p^*)$ is unlikely to deplete the inventory, so the problem almost reduces to the infinite inventory version. Thus by Lipschitzness $p_{\text{OPT}} \approx p^*$ hence $p_{\text{SOP}} = p^*$ has low regret in this case.

Now suppose $p^* < p_d$. We argue that $p_d$ is almost optimal among prices lower than $p_d$ and higher than $p_d$ respectively:
Consider \( p \leq p_d \). Since FPP\((p)\) is likely to sell out all inventory, we have \( r(p) \sim pI \). Applying this observation on \( p_d \), we have \( r(p_d, D) \sim p_d I \leq pI \) for \( p \leq p_d \).

Consider \( p \geq p_d \). Since the inventory is unlikely to be depleted by FPP\((p)\), we have \( r(p, D) \sim pD(p)T = R(p)T \). By unimodality, \( R \) is non-increasing on \([p_d, 1]\), so \( R(p_d) \geq R(p) \) and hence \( r(p_d, D) \sim R(p_d)T > R(p)T \sim r(p, D) \). Thus \( p_d \) is almost optimal in \([p_d, 1]\).

We will first analyze the regret of DUE against the following surrogate regret, and then translate the bound back to the “real” regret.

**Definition 5 (Surrogate Regret).** The surrogate regret of a policy \( \hat{\pi} \) under demand function \( D \) is \( \text{SR}(\hat{\pi}, D) = r(p_{\text{SOP}}, D) - r(\hat{\pi}, D) \). Let \( \mathcal{D} \) be any family of demand functions, define \( \text{SR}(\hat{\pi}, D) := \max_{D \in \mathcal{D}} \text{SR}(\hat{\pi}, D) \).

Since we assumed \( I = \Omega(T) \), the regret of \( p_{\text{SOP}} \) in Lemma 5 can be simplified to \( O(\Delta T + e^{-\Omega(\Delta^2 T)}T) \), which becomes \( O(\sqrt{T} \log^{1/2} T) \) if we select \( \Delta = T^{-1/2} \log^{1/2} T \). By Lemma 5, any bound on the surrogate regret immediately implies a bound on regret for a policy.

**Corollary 1.** Let \( \hat{\pi} \) be a markdown policy, then for any L-Lipschitz demand function \( D \),

\[
\text{Reg}(\hat{\pi}, D) \leq \text{SR}(\hat{\pi}, D) + O(\sqrt{T} \log^{1/2} T).
\]

As in Section 3, we denote \( x_j = 1 - js \) the \( j \)-th sample price for \( j = 1, \ldots, s^{-1} \), and \( \bar{d}_j \) the mean demand at \( x_j \) as defined in Step 4 of Algorithm 2. The proof can be split into two parts, stated in Lemmas 6 and 7. Following the ideas of Lemma 1, we may bound the regret when \( p^* \geq p_d \).

**Lemma 6.** Suppose \( p^* \geq p_d \). Then for any \( \delta, s \in (0, 1) \) such that \( \delta^{-2}s^{-1} = O(T^{0.99}) \),

\[
\text{SR}(\text{DUE}_{s, \delta}, D) \leq O(s^{-1}\delta^{-2}\log T + (sL + \delta)T).
\]

The analysis becomes more involved when \( p^* \leq p_d \). In this case, we need to argue that our exploitation price \( x_h \) is not much worse than \( p_d \). It is easy to see that \( D(x_h) \) is not too far away from \( D(p_d) \), and hence \( x_h \) will not overshoot \( p_d \) by more than \( O(s) \) to the left. To bound the reward of \( x_h \) when it is to the right of \( p_d \), we need to lower bound the expected sales of DUE, but this is no longer straightforward since DUE is not an FPP. To circumvent this issue, the next proposition lower-bounds the expected sales of DUE in the exploitation phase, using ideas similar to those in proving Lemma 5 (see Appendix B).

**Lemma 7.** When \( p^* \leq p_d \), then for any \( s, \delta, \varepsilon > 0 \) with \( \varepsilon = \tilde{O}(s^{-1}\delta^{-2}I^{-1} + sL\rho^{-1}) \), it holds that

\[
\text{SR}(\text{DUE}_{s, \delta}, D) = O(\delta T + sI + e^{-\Omega(s^2 T)}I + \varepsilon I + s^{-1}\delta^{-2}\log T).
\]

Theorem 2 immediately follows by combining the Lemma 6, Lemma 7 and Corollary 1, with \( \delta = T^{-1/4}(L \log T)^{1/4}, s = \delta L \) and \( \varepsilon = (1 - \rho)\alpha + sL\rho^{-1} \).
4. Proof of Lower Bound (Theorem 3)

We now turn to proving our lower bound, which establishes minimax optimality of the policy described in the previous section in the setting of infinite inventory. Without loss of generality (by re-scaling the following arguments), we take $\rho = 1$, and thus abbreviate $\text{Reg}(A, M, I)$ as $\text{Reg}(A, M)$ for simplicity. Consider the following family of reward curves (see Fig 1): each curve has slope $L$ and $-1$ on the left and right of its unique optimal price $p$, and truncated from below to ensure non-negativity. Formally, for $p, x \in [0, 1]$, consider:

$$R_p(x) = \begin{cases} 
1 - x, & \text{if } x \geq p, \\
1 - (L + 1)p + Lx, & \text{if } \frac{(L+1)p-1}{L} \leq x \leq p, \\
0, & \text{otherwise}
\end{cases}$$

Let $\mathcal{M} = \{R_p : p \in [1/8, 7/8]\}$. We will consider Bernoulli reward at each price.

If $\text{Reg}(A, M) \geq \frac{1}{48} L^{1/4} T^{3/4}$, the theorem holds trivially. Therefore, suppose $\text{Reg}(A, M) < \frac{1}{48} L^{1/4} T^{3/4}$. Consider the case when the optimal price $p = \frac{1}{8}$. Partition $[3/4, 7/8]$ uniformly into subintervals of length $[T^{-1/4}]$ as follows. Let $k = \lfloor \frac{1}{8} L^{3/4} T^{1/4} \rfloor$ and $x_j = \frac{7}{8} - L^{-3/4} T^{-1/4} j$ for each $j \in [k]$. We show in the following key lemma that $A$ has to spend $\Omega(T^{1/2})$ rounds in expectation in each subinterval. For any $a, b \in [0, 1]$, denote $N(A; a, b)$ the (random) number of rounds that $A$ selects a price in $[a, b]$.

**Lemma 8 (Key Lemma).** Let $A$ be a markdown policy with $\text{Reg}(A, M) \leq \frac{1}{48} L^{1/4} T^{3/4}$ for all $T > 2^{12} L$. If $[a, b] \subset [3/4, 7/8]$ where $b = a + L^{-3/4} T^{-1/4}$ and $p \leq 1/8$, then $\mathbb{E}_{R_p}[N(A; a, b)] = \Omega(L^{-1/2} T^{1/2})$. 
By Lemma 8, since \([x_j, x_{j-1}] \subset [3/4, 7/8]\) for each \(j \in [k]\), we have

\[
\mathbb{E}_{R_p}[N(\mathcal{A}; \frac{3}{4}, \frac{7}{8})] = \mathbb{E}_{R_p}\left(\sum_{j=1}^{k} N(\mathcal{A}; x_j, x_{j-1})\right) = k \cdot \Omega(L^{-1/2}T^{1/2}) = \Omega(L^{1/4}T^{3/4}).
\]

By our choice of \(p\), the regret per round is \(\Omega(1)\) when the explored price remains in \([3/4, 7/8]\), thus

\[
\text{Reg}(\mathcal{A}, \mathcal{M}) \geq \text{Reg}(\mathcal{A}, R_p) \geq \Omega(\mathbb{E}_{R_p}[N(\mathcal{A}; \frac{3}{4}, \frac{7}{8})]) = \Omega(L^{1/4}T^{3/4}).
\]

\[
\square
\]

4.1. Proof of the Key Lemma 8

Our proof employs the following alternate view of a policy (see Fig 2).

**Definition 6 (Decision Tree).** A decision tree is a rooted binary tree \(T\) where each node \(v\) is labelled a node-price \(x_v \in [0, 1]\), and each edge corresponds to a realized binary demand. An adaptive markdown policy \(\mathcal{A}\) is essentially a decision tree where \(T\) is the complete binary tree of depth \(T\). Further, along each path, the node labels, which represent the prices, are non-increasing.

The effect of running a policy corresponding to a tree \(T\) on a specific reward function \(R\) is to induce a probability distribution of arriving at one of its leaves. We define this next.

**Definition 7 (Probability Measure on Leaves).** Let \(T\) be a binary decision tree (possibly not balanced) with node label \(x_v \in [0, 1]\) at each node \(v\), and \(L(T)\) be its leaves. Let \(R : [0, 1] \rightarrow [0, 1]\) be any reward function. Suppose a particle walks down \(T\) from the root to a leaf, where at each node \(v\), the particle goes to the left and right child w.p. \(R(x_v)\) and \(1 - R(x_v)\) respectively. Let \(P_R\) be the probability measure on \(L(T)\) induced by this random walk. By abuse of notation, for each \(\ell \in L(T)\), denote by \(P_R(\ell)\) the probability that the particle ends in \(\ell\) under reward function \(R\).

It will be convenient for the proof to only consider policies represented by trees where the node prices never change after \(T/2\). The following lemma argues that this is without loss of generality.

**Lemma 9.** For any markdown policy \(\mathcal{A}\), there is a policy \(\mathcal{B}\) which makes no price-change after \(T/2\) s.t. \(\text{Reg}(\mathcal{B}, R) \leq 2 \cdot \text{Reg}(\mathcal{A}, R)\) for all \(R \in \mathcal{F}_L\).

The area of interest in the decision tree for analyzing the statement of Lemma 8 is the region where the price enters the range \([a, b] \subset [3/4, 7/8]\), which we define next.

**Definition 8 (Entrance).** An entrance in a tree \(A\) for an interval \([a, b]\) is any node \(v\) with \(x_v \leq b\) and \(x_{\text{par}(v)} > b\), where \(\text{par}(v)\) is the parent of \(v\). If a path in \(A\) has no such node, then we define its node on level \(T/2\) as its entrance. Thus, for each path in \(A\), there is a unique entrance.

---

8 One may easily extend the definition to any finite-branch decision tree, but since proof only relies on Bernoulli rewards, we will proceed with binary trees for simplicity.

9 The level (or depth) of a node in a rooted tree is the hop-distance to the root.
Recall that \( b \in [3/4, 7/8] \). Since we are trying to determine whether the tree can differentiate between \( R = R_1 \) and \( B = R_b \), at each entrance \( v \) (for \([a, b]\)) we construct a truncated tree rooted at \( v \) along with its leaf-coloring \( f_v \).

Note that every entrance has depth at most \( \frac{T}{2} \) by definition. Our goal is to label the leaves of the subtree based on whether the policy predominantly uses prices in the range \([a, b]\) or below \( a \). Before we label the leaves however, we truncate the subtrees under each entrance at a further depth of \( \frac{T}{4} \) so that there is a further runway of at least an additional \( \frac{T}{2} - \frac{T}{4} = \frac{T}{4} \) steps for the labeled prices to contribute to the regret estimation. The number of steps \( N(A; a, b) \) that the policy selects prices in \([a, b]\) can be captured by the nodes in the corresponding decision tree \( T \) that are labeled in \([a, b]\).

We define these regions below each entrance by truncating the subtrees when they select prices below this range.

1. (Truncation) Remove all nodes that are at least \( \frac{T}{4} \) levels below \( v \). If a descendent \( u \) of \( v \) has node-price \( x_u \leq a = b - L^{-3/4}T^{-1/4} \), then remove all descendants of \( u \), hence \( u \) becomes a leaf. The subtree of \( T \) rooted at the entrance \( v \) after carrying out these descendant removals and the truncation in the previous step is denoted \( T_v \).

2. (Coloring) For every leaf \( \ell \in T_v \), if \( x_\ell \leq a \), set \( f_v(\ell) = R \), else \( f_v(\ell) = B \).

We formalize the notion of a tree differentiating between reward functions \( R \) or \( B \) using the induced probability measure on the leaves of the tree.

**Definition 9 (Confident Leaf-Coloring).** A leaf-coloring \( f : L(T) \rightarrow \{R, B\} \) is \((\alpha, \beta)\)-confident for some \( \alpha > \frac{1}{2} > \beta \), if

\[
P_D := \mathbb{P}_R(f(\ell) = R) \geq \alpha, \qquad \text{(Detection probability is high)}
\]

and

\[
P_{FA} := \mathbb{P}_B(f(\ell) = R) \leq \beta. \qquad \text{(False-Alarm probability is low)}
\]

This definition may remind readers of binary hypothesis testing, where \( R, B \) correspond to the null and alternate hypotheses, with type-I error \( 1 - \alpha \) and type-II error \( \beta \). The definitions of the leaf coloring and their confidence above are used to convert the quest for the lower bound on the regret of the policy to a lower bound of the depth of a decision tree for an appropriate hypothesis test.

An entrance \( v \) is said to be confident, if the leaf-coloring of its truncated subtree \( f_v \) is \((2/3, 1/3)\)-confident. Intuitively, if an entrance is confident, the policy represented by the tree is on track in distinguishing the right reward function between \( R \) and \( B \). In general, it is not true that all entrances in a low-regret tree are confident since a small fraction of non-confident entrances may not hurt the regret by too much. Hence we show the following weaker claim.
Claim 1. Let $V_c$ be the set of confident entrances in $A$ and $v_{\text{ent}}$ be the random entrance node, then $\mathbb{P}_\chi(v_{\text{ent}} \in V_c) \geq 1/2$ for any $\chi \in \{R,B\}$.

To finish the proof, we argue that confident entrances lead to sufficiently large depth subtrees. To lower-bound on $\mathbb{E}_R[N(A;a,b)]$, we need a lower bound on the expected depth of the truncated subtrees under the confident entrance nodes. Write $\mathbb{E}_\chi = \mathbb{E}_{\ell \sim \mathbb{P}_\chi}$ for $\chi = R, B$.

Claim 2. Suppose entrance $v$ is confident and $D(\ell)$ is the depth of $\ell \in L(T_v)$ in $T_v$. Then,

$$\mathbb{E}_R[D(\ell)] = \Omega(L^{-1/2}T^{1/2}), \quad \text{and} \quad \mathbb{E}_B[D(\ell)] = \Omega(L^{-1/2}T^{1/2}).$$

If every node in $T_v$ were labeled with the same price, this would use the same probabilities for the particle traveling left or right at each node under each of $R$ and $B$. In this case, the above claim can be reduced to a lower bound from the classical Wald-Wolfowitz hypothesis testing setting between these distributions defined by $R$ and $B$ at this single price. The proof of the above claim generalizes this lower bound to the setting of a continuum of distributions defined by the non-increasing prices selected in the truncated subtree.

Combining Claim 1 and Claim 2 (that are proved in the Appendix),

$$\mathbb{E}_R[N(A;a,b)] \geq \mathbb{E}_R[N(A;a,b)|v \in V_c] \cdot \mathbb{P}_R[v \in V_c] \geq \mathbb{E}_R[D(\ell)|v \in V_c] \cdot \mathbb{P}_R[v \in V_c] \geq \Omega(L^{-1/2}T^{1/2}) \cdot \frac{1}{2} = \Omega(L^{-1/2}T^{1/2}).$$

This completes the proof of Lemma 8.

5. Dynamic Pricing with Markup Penalty

The markdown pricing problem can be viewed as dynamic pricing problem where each mark-up incurs an infinite penalty, and we have just shown a tight $\tilde{O}(T^{3/4})$ regret bound. On the other hand, if prices can oscillate for free, the problem becomes ordinary Lipschitz bandits when $I = \infty$, which is known to admit an $\tilde{O}(T^{2/3})$ regret. Hence we arrive at a natural question:

**If there is a finite penalty for each mark-up, can we improve upon the $\tilde{O}(T^{3/4})$ bound for markdown pricing?**

More precisely, given finite markup penalty, can we interpolate its regret bound between $\tilde{O}(T^{3/4})$, the bound in the presence of penalty, and $\tilde{O}(T^{2/3})$, the bound for zero markup cost?

We provide an affirmative answer to this question. We organize this section as follows. We first show that the regret bound can be improved to $\tilde{O}(T^{2/3})$, if $O(\log T)$ number of markups is allowed. Then we proceed to introduce the problem of dynamic pricing with markup penalty (DPMP), and derive immediately a tight regret bound for DPMP using the results established so far.
5.1. Dynamic Pricing with Few Markups

Recall that for non-markdown pricing (i.e. Lipschitz bandits), a tight $\tilde{O}(T^{2/3})$ regret bound can be achieved by applying the Successive Elimination (SE) policy on a suitably discretized price space. However, such a policy may mark up for $\Omega(T^{2/3})$ times. As the cornerstone of this section, we first present a simple adaptation of the SE policy that achieves the same regret, $\tilde{O}(T^{2/3})$, but only marks up $O(\log T)$ times.

Algorithm 3 Geometric Successive Elimination Policy $GSE_{s,\varepsilon}$.

1: Input: $s, \varepsilon > 0$.
2: Initialize: $A \leftarrow \{i \cdot s | 0 \leq i \leq s^{-1}\}, \ell \leftarrow 2\log \varepsilon^{-1}$. $\triangleright$ Discretize price space into arms.
3: for $j = 0, 1, \ldots, \ell$ do $\triangleright$ Exploration phase consists of $\ell$ cycles.
4: for each $a \in A$ in decreasing order do
5: Select $a$ for $2^j$ times in a row and observe rewards $X_{2^j}(a), \ldots, X_{2^{j+1}-1}(a)$.
6: $\tilde{\mu}(a) \leftarrow (2^{j+1} - 1)^{-1} \sum_{i=1}^{2^{j+1} - 1} X_i(a)$. $\triangleright$ Empirical mean.
7: $[LCB(a), UCB(a)] \leftarrow [\tilde{\mu}(a) - \sqrt{\frac{\log T}{2^{j+1} - 1}}, \tilde{\mu}(a) + \sqrt{\frac{\log T}{2^{j+1} - 1}}]$. $\triangleright$ Confidence interval.
8: Eliminate $a$ from $A$ if there exists $a' \in A$ s.t. $UCB(a) < LCB(a')$. $\triangleright$ Elimination.
9: Select any $a \in A$ henceforth. $\triangleright$ Exploitation phase.

As in the SE policy, our Geometric Successive Elimination (GSE) policy (see Algorithm 3) discretizes the price space into arms and splits the time horizon into cycles. Different than the classical version where each cycle has the same length, in GSE the cycle lengths are given by a geometric sequence. In the $j$-th cycle, GSE sequentially selects each alive arm (i.e. not eliminated) for $2^j$ times consecutively, and at the end of each cycle eliminates the arms whose confidence intervals are dominated by some other arm. Specifically, the number of cycles that suffices to achieve the desired regret bound is only $O(\log T)$, in other words, GSE only marks up for $O(\log T)$ times. We formally state this result below.

**Theorem 5 (Pricing Policy with Few Markups).** Denote $GSE_{s,\varepsilon}$ the Geometric Successive Elimination Policy with parameters $s, \varepsilon > 0$. Then for $s = L^{-2/3}T^{-1/3}$ and $\varepsilon = L^{1/3}T^{-1/3}$,

$$\text{Reg}(GSE_{s,\varepsilon}, \mathcal{F}_L) = O(L^{1/3}T^{2/3}\sqrt{\log T}).$$

We conclude this subsection with a couple of observations. First, in this result we no longer require the reward functions to be unimodal. Further, when $T$ is unknown, we may apply the doubling trick (folklore, see e.g. Auer et al. (1995)) to achieve the same bound. Finally, compared to the $O(T^{2/3}(L \log T)^{1/3})$ bound for Lipschitz Bandits, this bound is only weaker by $O(\log^{1/6} T)$. 
5.2. Dynamic Pricing with Markup Penalty

In practice, an increase in price usually results in a decrease in demand (Homburg et al. (2005), Malc et al. (2016), Rotemberg (2002)). In this section, we model this effect by introducing a new concept, the **Markup Penalty Index** (MPI). As opposed to imposing a penalty on each markup, previous work has considered how promotions may boost the demand (Ramakrishnan (2012)). Building upon our lower bound techniques in the pure markdown setting, we show a tight regret bound in terms of the MPI for unimodal Lipschitz families.

**Random Utility Model.** We introduce a stylized model that captures the penalty incurred by mark-ups. Each buyer has two states: active and inactive. At each round, each active buyer’s valuation $v$ is i.i.d. drawn from some fixed unknown distribution $\mathcal{D}$, and she decides to buy if the utility $u(p, v) := v - p > 0$ where $p$ is the current price. Thus, the demand rate (i.e. fraction of buyers who buy) at price $p$ is $\mathbb{P}[u > 0] = \mathbb{P}_v \sim \mathcal{D}[v > p]$. To model the negative effect of markups, we assume that each time the seller marks up, a constant $\gamma$ fraction of active buyers become inactive and set their valuations to 0 permanently. Thus, after $k$ mark-ups, the amount of active buyers drops to $n(1 - \gamma)^k$ hence the demand rate becomes $(1 - \gamma)^k \mathbb{P}[v > p]$. Since $(1 - \gamma)^k = 1 - k\gamma + o(\gamma)$, compared to having no markup, the seller “loses” $(1 - (1 - \gamma)^k) = k\gamma + o(\gamma)$ fraction of demand at price $p$. This amounts to charging a constant, additive cost penalty for each markup.

**Extreme Cases.** If $\gamma = O(T^{-1/3})$, then each markup incurs at most $\gamma T = O(T^{2/3})$ loss over the entire time horizon. In this case, by Theorem 5, an $\tilde{O}(T^{2/3}) + \gamma T \log T = \tilde{O}(T^{2/3})$ regret can be achieved. If $\gamma = \Omega(1)$, then each markup incurs a $\Omega(T)$ loss, hence the markup penalty effectively becomes a hard constraint, and the problem reduces to the “pure” markdown problem discussed in the previous sections. Thus, the interesting case is when the order of $\gamma T$ is between $T^{2/3}$ and $T$.

**Definition 10 (MPI).** The **Markup Penalty Index (MPI)** is a number in $[0, 1]$ defined as $c = 1 + \log_T \gamma$, i.e. the unique $c$ satisfying $\gamma T = T^c$.

As in the previous sections, we compare the performance of a policy against the optimal FPP. Since the optimal FPP always selects $r^* = \max_{p \in [0, 1]} R(p)$ where $R$ is the underlying reward function, it pays no markup cost. Thus, the regret of a policy can be decomposed into the markup penalty and the cost of selecting suboptimal prices. Our goal is to find a policy that minimizes the regret, formalized below.

**Definition 11 (Regret with Markup Penalties).** Suppose the MPI is $c \in [0, 1]$. For any policy $A$, let $r(A, R)$ be its expected total reward and $\nu(A)$ be its total number of markups, i.e.

$$\nu = \sum_{t=1}^{T} 1[A(t) < A(t+1)]$$
where $A(t)$ is the price $A$ selects at $t$. Define the regret to be

$$\text{Reg}_c(A, R) = r^* T - r(A, R) + E_R[v(A)] \cdot T^c.$$  

The regret on a family $\mathcal{F}$ of reward functions is simply $\text{Reg}_c(A, \mathcal{F}) = \max_{R \in \mathcal{F}} \text{Reg}_c(A, R)$.

Note that the unimodality assumption is not needed here, so Theorem 5 immediately implies the following upper bound on $\mathcal{F}_L$.

**Corollary 2 (Upper Bound for Lipschitz Reward Functions).** Let $\text{GSE}_{s, \varepsilon}$ be the Geometric Successive Elimination Policy with parameters $s, \varepsilon > 0$. Then for $s = L^{-2/3} T^{-1/3}$ and $\varepsilon = L^{1/3} T^{-1/3}$, we have $\text{Reg}_c(\text{GSE}_{s, \varepsilon}, \mathcal{F}_L) = O(L^{1/3} T^{2/3} \sqrt{\log T} + T^c \log T)$.

However, when $c$ is large, the above regret bound becomes poor. Observe that if, in addition, all reward functions are unimodal (i.e. replace $\mathcal{F}_L$ with $\hat{\mathcal{F}}_L$), the Uniform-Elimination policy (Algorithm 1) achieves regret $\tilde{O}(T^{3/4})$. Thus if $c$ is known, by choosing the better policy between GSE and UE we achieve the following regret for $\hat{\mathcal{F}}_L$.

**Theorem 6.** Let $A$ be the policy that chooses $\text{GSE}_{s, \varepsilon}$ with $s = L^{-2/3} T^{-1/3}, \varepsilon = L^{1/3} T^{-1/3}$ when $c \leq 3/4$ and chooses $\text{UE}_{s, \delta}$ with $\delta = \sqrt{2} T^{-1/4} (L \log T)^{1/4}, s = \delta/2L$ otherwise. Then,

$$\text{Reg}_c(A, \hat{\mathcal{F}}_L) = \tilde{O}(T^{\text{med}(\frac{2}{3}, \frac{4}{3})}) = \begin{cases} O(T^{2/3} (L \log T)^{1/3}), & \text{if } c \leq 2/3, \\ O(T^c \log T), & \text{if } 2/3 < c \leq 3/4, \\ O(T^{3/4} (L \log T)^{1/4}), & \text{else}. \end{cases}$$

Surprisingly, this bound for $\hat{\mathcal{F}}_L$ turns out to be almost optimal as stated in the following theorem.

**Theorem 7 (Lower Bounds for General MPI).** For any policy $A$ that knows the time horizon $T$ and the true MPI $c$,

$$\text{Reg}_c(A, \hat{\mathcal{F}}_L) = \begin{cases} \Omega(T^{2/3}), & \text{if } c < 2/3 \\ \Omega(T^c), & \text{if } c \in [2/3, 3/4] \\ \Omega(T^{3/4}), & \text{if } c > 3/4. \end{cases}$$

Our results show that the markup penalty index (MPI) plays a critical role in the achievable regret, and that three regimes emerge (Table 3). First, when the MPI is low ($c < 2/3$), the penalty for marking up is dominated by the cost incurred for searching for the optimal price. The regret (both upper and lower) is thus $\tilde{\Theta}(T^{2/3})$ no matter how small the MPI is. Second, when the MPI is moderate ($c \in [2/3, 3/4]$), the markup penalty now dominates the search cost. In this regime, careful restriction of the use of markups allows the regret to be limited to $\tilde{\Theta}(T^c)$. Third and finally, when the MPI is high ($c > 3/4$), the story remains the same, and unfortunately the regret is even linear.
if \( c = 1 \) (corresponding to the pure markdown setting). However, if in addition to the Lipschitz assumption, we assume that (a) the reward function is unimodal and that (b) the time horizon is known, then a separate policy which never marks up is able to achieve regret \( \tilde{O}(T^{3/4}) \). These two assumptions together form a necessary and sufficient set of conditions for achieving sub-linear regret in the pure markdown setting.

All results until now have assumed that the time horizon \( T \) is known in advance. This assumption is, in a sense, necessary – we show in the appendix that if \( T \) is unknown, no policy achieves non-trivial regret (see Appendix F). We conclude this section by summarizing our results from a different lens: our results show that the MPI plays a critical role in the achievable regret, and that three regimes emerge (Table 3).

<table>
<thead>
<tr>
<th>Lipschitz</th>
<th>Low MPI ((c &lt; 2/3))</th>
<th>Med. MPI ((2/3 \leq c \leq 3/4))</th>
<th>High MPI ((c &gt; 3/4))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipschitz</td>
<td>( \tilde{O}(T^{2/3}) )</td>
<td>( \tilde{O}(T^c) )</td>
<td>( \tilde{O}(T^c) )</td>
</tr>
<tr>
<td>Lipschitz, unimodal, known ( T )</td>
<td>( \tilde{O}(T^{2/3}) )</td>
<td>( \tilde{O}(T^c) )</td>
<td>( \tilde{O}(T^{3/4}) )</td>
</tr>
</tbody>
</table>

Table 3 Upper and lower Regret bounds in three different regimes based on \( c \), the markup penalty index (MPI).

5.3. Proof of Theorem 5

We first describe the proof at a high level. The regret can be decomposed into the discretization error \( \epsilon T \) and the regret on the discretized instance. To bound the latter, we show that for any arm, the more suboptimal, the faster it gets eliminated (Lemma 11). Therefore, any alive arm at the end of the exploration phase is nearly optimal. We formalize this idea below.

We will use the following folklore regret-decomposition lemma (see e.g. Lemma 4.5 of Lattimore and Szepesvári (2020)).

**Lemma 10 (Decomposition of Regret).** Consider a Multi-armed Bandit instance \( I \) where each arm \( i \in [K] \) has mean reward \( \mu_i \in [0,1] \). Let \( \mu^* = \max_{i \in [K]} \mu_i \) and \( N_i \) be the number of times arm \( i \) is selected by a bandit policy \( \mathcal{A} \). Then,

\[
\text{Reg}(\mathcal{A}, I) = \sum_{i \in [K]} (\mu^* - \mu_i) \cdot \mathbb{E}N_i.
\]

We apply the above lemma to decompose the regret of our problem. Let \( r_{\max} = \max_{0 \leq k \leq s-1} R(a_k) \) be the optimal reward rate in the discretized instance and recall that \( r^* = \max_{x \in [0,1]} R(x) \) is the optimal reward rate of the (continuous) reward function. To relate to Lemma 10, denote \( \Delta_i = \)
Let \( r_{\text{max}} - R(a_i) \) and let \( N'_i, N''_i \) be the number of times \( i \) is selected in the exploration and exploitation phase respectively. Then,

\[
\text{Reg}(\text{GSE}_{s, \varepsilon}, R) = r^*T - r(A, R) = (r^*T - r_{\text{max}}T) + (r_{\text{max}}T - r(A, R)) 
\leq LsT + \sum_{0 \leq i \leq s-1} \mathbb{E}(N'_i + N''_i) \cdot \Delta_i 
= LsT + \sum_{i : \Delta_i \geq \varepsilon} \mathbb{E}N'_i \cdot \Delta_i + \sum_{i : \Delta_i < \varepsilon} \mathbb{E}N'_i \cdot \Delta_i + \sum_{0 \leq i \leq s-1} \mathbb{E}N''_i \cdot \Delta_i 
\leq LsT + \sum_{i : \Delta_i \geq \varepsilon} \mathbb{E}N'_i \cdot \Delta_i + \varepsilon T + \sum_{0 \leq i \leq s-1} \mathbb{E}N''_i \cdot \Delta_i. \quad (4)
\]

We need the following lemma (proof deferred to Appendix E) to bound the second and last term.

**Lemma 11.** Suppose \( R \in \mathcal{F}_L \) is the true reward function. For each \( i \leq s^{-1} \), let \( N'_i \) be the number of times arm \( a_i := s_i \) is selected in the exploration phase of \( \text{GSE}_{s, \varepsilon} \). Let \( \mathcal{E} \) be the event that \( N'_i \leq \min\{T, 16\Delta^{-2}_i \log T\} \) for all \( 0 \leq i \leq s^{-1} \), then \( \mathbb{P}[\mathcal{E}] \geq 1 - T^{-3} \).

Thus, the second term can be bounded as follows.

\[
\sum_{i : \Delta_i \geq \varepsilon} \mathbb{E}N'_i \cdot \Delta_i = \sum_{i : \Delta_i \geq \varepsilon} \Delta_i \cdot (\mathbb{P}(\mathcal{E}) \cdot \mathbb{E}[N'_i|\mathcal{E}] + \mathbb{P}(\overline{\mathcal{E}}) \cdot \mathbb{E}[N'_i|\overline{\mathcal{E}}]) 
\leq \sum_{i : \Delta_i \geq \varepsilon} \Delta_i \cdot (\mathbb{P}(\mathcal{E}) \cdot 16\Delta^{-2}_i \log T + \mathbb{P}(\overline{\mathcal{E}}) \cdot T) 
\leq \sum_{i : \Delta_i \geq \varepsilon} \Delta_i \cdot (16\Delta^{-2}_i \log T + T^{-2}) 
\leq 16s^{-1}\varepsilon^{-1} \log T + o(1) \quad \text{Since there are } s^{-1} \text{ arms.} \quad (5)
\]

We next bound the last term in (2), \( \sum_{0 \leq i \leq s-1} \mathbb{E}N''_i \cdot \Delta_i. \) Observe that each arm alive at the end of the exploration phase has been selected for at least \( 2^f = 2^{2\varepsilon^{-1}} = \varepsilon^{-2} \) times. On the other hand, Lemma 11 says conditional on \( \mathcal{E} \), each arm can be selected for at most \( 16\Delta^{-2}_i \log T \) times. Thus, conditional on \( \mathcal{E} \), for each arm \( i \) selected for exploitation, it holds

\[
\varepsilon^{-2} < 16\Delta^{-2}_i \log T,
\]

i.e. \( \Delta_i < 4\sqrt{\log T} \varepsilon. \) Therefore,

\[
\sum_{0 \leq i \leq s-1} \mathbb{E}N''_i \cdot \Delta_i = \sum_{0 \leq i \leq s-1} (\mathbb{E}[N''_i|\mathcal{E}] \cdot \mathbb{P}[\mathcal{E}] + \mathbb{E}[N''_i|\overline{\mathcal{E}}] \cdot \mathbb{P}[\overline{\mathcal{E}}]) \cdot \Delta_i 
\leq \sum_{0 \leq i \leq s-1} \mathbb{E}[N''_i|\mathcal{E}] \cdot \Delta_i + T^{-3} \cdot \sum_{0 \leq i \leq s-1} T 
\leq 4\sqrt{\log T} \varepsilon \cdot T + o(1). \quad (6)
\]
Substituting (5), (6) into (2), we have

$$\left(2\right) \leq 16\varepsilon^{-1}s^{-1}\log T + 4\sqrt{\log T}\varepsilon \cdot T + (sL + \varepsilon)T + o(1).$$

The proof follows by selecting \(\varepsilon = L^{1/3}T^{-1/3}\) and \(s = L^{-2/3}T^{-1/3}\sqrt{\log T}.\)

6. Experiments

In this section, we compare the empirical performance of the UE policy described in Section 2.2 with several alternative policies in the case of infinite inventory. Our results demonstrate that the UE policy is reasonably fast in convergence and robust to model misspecification that affect other parametric policies adversely.

6.1. Robustness Under Model Misspecification

We first compare the performance of our policy Uniform Elimination (UE) with two Explore-Then-Commit (ETC) type policies. A generic ETC policy assumes certain parametric form of the underlying demand function (which is possibly incorrect) and consists of an exploration phase and an exploitation phase. In the exploration phase, the policy randomly selects two sample prices \(p_1, p_2\) near the maximum price, each for sufficiently many times. At the end of the exploration phase, the policy estimates the true parameters from the observations, and commits to the optimal price of the estimated demand function throughout the exploitation phase. We provide more details below.

**ETC-Policies.** An ETC policy is specified by two parameters: \(h \in [0, 1]\) and \(k \in \mathbb{N}.\) It first uniformly draws two sample prices \(p_1, p_2\) from \([1-h, 1]\) and \([1-3h, 1-2h]\), and then selects each for \(k\) times. At the end of the exploration phase, the policy computes a demand function from the assumed demand family that best fits the empirical mean demands \(\bar{d}_i\) at each \(p_i.\) Formally,

- **ETC\(_{Lin}\)** fits a linear demand function \(\hat{D}(x) = \hat{a} - \hat{b}x\) given by
  
  \[
  \hat{b} = -\frac{\bar{d}_1 - \bar{d}_2}{p_1 - p_2}, \quad \hat{a} = \hat{b} \cdot p_1 + \bar{d}_1.
  \]

- **ETC\(_{Exp}\)** fits an exponential demand function \(\hat{D}(x) = \exp(\hat{a} - \hat{b}x)\) given by
  
  \[
  \hat{b} = -\frac{\log \bar{d}_1 - \log \bar{d}_2}{p_1 - p_2}, \quad \hat{a} = \hat{b} \cdot p_1 + \log \bar{d}_1.
  \]

Finally, note that the optimal price is \(\frac{a}{2b}\) for \(D(x) = a - bx,\) and \(1/b\) for \(D(x) = e^{a-bx},\) so in the exploitation phase the ETC policy selects the optimal price of \(\hat{D},\) given by \(\hat{p}_{lin} = \text{Clip}_{[0,1]}(\frac{\hat{a}}{2\hat{b}})\) and \(\hat{p}_{exp} = \text{Clip}_{[0,1]}(1/\hat{b})\) for each case\(^{10}.\)

\(^{10}\) For any \(a, b, x, \text{Clip}_{[a,b]}(x)\) is defined to be the median of \(a, b\) and \(x.\)
One of the merits of the UE policy is robustness: different than ETC policies, it does not assume a certain functional form on the underlying demand function. We compare the regret of the above three policies on linear demand functions $D(x) = \{a - bx\}$ and exponential demand functions $D(x) = \{e^{-dx}\}$ over 1000 independently randomly generated demand functions. In each epoch, we randomly generate demand functions by drawing

$$a \sim U(0,1), b \sim U(0,a), c \sim U(0,3), d \sim U(0,10).$$

Note that $b$ is capped at $a$ since otherwise the value of the linear function may be negative. Furthermore, we scale each demand function so that the maximum reward rate is 1.

For UE, we set the parameters to be order-optimal, $s = \delta = \frac{1}{4} T^{1/4}$, as proved in Theorem 1. For simplicity, we ignore the the Lipschitz constant $L$ in selecting the policy’s parameters. For both ETC policies, we set $h = 0.1$ and for fairness of comparison, we set $k$ to be the same as in UE, i.e. $k = \delta^{-2} \log T$. Our results in Fig 3 demonstrate the following properties of the UE policy:

1. Fast convergence rate of regret: the regret of UE vanishes at a decent speed for both linear and exponential demand functions.
2. Robustness to model misspecification: UE has vanishing regret on both demand functions. In comparison, the regret of each ETC policy converges to 0 at a speed much faster than UE, but does not converge under the incorrect model assumption.

![Figure 3](image-url)  
**Figure 3** Comparison between UE and ETC type policies on exponential and linear demand functions.

### 6.2. Impact of the Lipschitz Constant

Intuitively, as the Lipschitz constant $L$ increases, reward function may change faster around the optimum, rendering the problem trickier. In our theoretical analysis, the $\tilde{O}(L^{1/4} T^{3/4})$ regret bound, further confirms this intuition. In this section we numerically investigate the influence of $L$. 


Experiment Setup. We compare the performance of UE policy on several randomly generated families of exponential demand functions, each with a different range for Lipschitz constants.

We now describe how each curve in Figure 4 is obtained. Since we always scale the demand function so that the maximum reward rate is normalized to 1, the only parameter that matters in the demand function \( D(x; a, b) = e^{a-bx} \) is \( b \). For a fixed parameter range \([0, b_{\text{max}}]\), we first compute the (minimal) Lipschitz constant \( L = L(b_{\text{max}}) \) for the family \( \mathcal{F} = \{ D(x; 0, b) | b \in [0, b_{\text{max}}] \} \). Note that \( L(b_{\text{max}}) \) increases as \( b_{\text{max}} \) increases. Then, we compute the average regret of our UE policy with the (order-) optimal hyper-parameters \( s = L^{1/4}T^{3/4} \) and \( \delta = L^{-3/4}T^{3/4} \) over \( 10^3 \) randomly drawn demand models \( D(x; b) \) where \( b \sim U(0, b_{\text{max}}) \). Each curve in Figure 4 is obtained by repeating this process for integers \( b_{\text{max}} \) from 1 to 6.

![Figure 4](image-url)  
Figure 4  How the regret changes for different Lipschitz constants.

Finally, we let \( T \) to vary from \( 10^5 \) to \( 10^9 \) and obtain 5 curves. We observe that for each \( T \), the average regret is increasing in \( L \). This matches our intuition since as \( b_{\text{max}} \) increases, the reward function becomes steeper around the peak, hence it is trickier to decide when to halt.

7. Conclusion
In this paper we showed tight regret bounds for markdown pricing under unknown demand model. Our regret bounds reveal that the markdown constraint adds significantly more complexity to
dynamic pricing problems, since the corresponding regret bounds are asymptotically higher than without this constraint. Moreover, we introduced a new problem, dynamic pricing with markup penalty, that incorporates the negative effect of markups, and provided tight regret bounds. Finally, we showed through numerical experiments that our policy is robust to model misspecification and its regret vanishes rapidly.

This work opens up some new directions in dynamic pricing for future research:

1. First, this work made minimal assumptions (Lipschitz and unimodal). In practice, however, demand functions usually take some simple functional forms. We showed how to improve these bounds when the reward function is twice-differentiable. An open problem in this direction is whether we can improve the regret bounds for specific families such as linear, exponential or logit demand functions.

2. Second, we showed that our lower bounding technique has wider applicability in other bandit problems – for example, they can be applied to give a simple alternate lower bound proof for the ordinary MAB problem. An interesting direction is to extend this technique to prove lower bounds for other online learning problems.

3. Finally, our analysis compares against the best fixed price policy (FPP). For the infinite inventory version, the best FPP is optimal among all policies, but this is no longer true for the finite inventory version. An open question in this direction is, is it possible to analyze the regret against the best policy that knows the true demand function?

4. In this work we introduced a simple, elegant model that captures markup penalty and provided tight regret bounds. However, our stylized model fails to incorporate many practical aspects such as the heterogeneity in customer behaviors, and the impact of markup magnitude. We hope this work can open up a new direction in modeling the effect of markup.

References


