ONLINE ALGORITHMS FOR REVENUE MANAGEMENT

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Josef Meinrad Broder

August 2011

ONLINE ALGORITHMS FOR REVENUE MANAGEMENT

Josef Meinrad Broder, Ph.D.

Cornell University 2011

Traditional approaches to revenue management assume that a seller has a large degree of certainty about the market environment in which he operates. This dissertation focuses on the practice of revenue management in *uncertain* market environments, and describes strategies by which a seller can intelligently gather information about the demand for his products, to effectively price and distribute his goods. We describe theoretically sound policies for both pricing and distribution of goods under demand uncertainty, and establish insights into the structure of effective strategies, and the inherent challenges in such problems.

Biographical Sketch

Josef Broder was born on February 16^{th}, 1984 in Athens, Georgia. He received his B.S. in Mathematics from the University of Georgia in May 2006 and his M.S. in Applied Math from Cornell University in March 2010. He will receive his Ph.D. in Applied Math from Cornell in August 2011.

To my parents, Joe and Diane.

Acknowledgements

First, I would like to give my deepest thanks to my advisor, Paat Rusmevichientong, without whom this dissertation would not have been possible. Paat has been a limitless source of advice, encouragement, and support for me during my years in graduate school, and his genuine interest and enthusiasm for our work has kept me going, in good times and in bad. I thank my committee members, David Shmoys and Michael Nussbaum, for all of the helpful advice and feedback they have provided during my time at Cornell. Also, I am indebted to Bobby Kleinberg, for serving as a mentor to me during my early years in graduate school, and for generously sharing with me his insights and advice.

The Center for Applied Math has been a wonderful home to me during my time at Cornell. It has been a privilege to be a student in the department, and I am a better person for knowing the friends and colleagues I have met there. I would especially like to thank Steve Strogatz, Dolores Pendell, and Selene Cammer for their kindness, advice, and support, and for making CAM an amazing place to be.

Finally, I owe the greatest debt to my family, especially to my parents, Joe and Diane, and to my fiancee Miranda, who have provided me with constant love and support in all my endeavors.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This dissertation focuses on two fundamental problems faced by any revenue manager: how should I *set prices* for my products, and how should I *distribute* my products for sale? The answers to both of these questions depend intrinsically on the demand for the product, and accordingly, traditional studies in the revenue management literature assume that the seller has prior knowledge of the demand for his goods. Our concern here will be to understand how the seller can answer the above questions *without prior knowledge of the demand for his products*. Thus, our focus will be on a *learning* approach to revenue management, and we will seek to understand how the seller can effectively price and allocate his goods in the face of demand uncertainty.

The first two chapters are devoted to the question of pricing under demand uncertainty, and are based on the observation that to make intelligent pricing decisions, the seller must have some knowledge of the demand curve for his product. Thus, when the seller has imperfect information about the demand for the good, finding optimal pricing strategies becomes a challenging problem. We in-

vestigate strategies for price experimentation, and establish fundamental limits on the performance of pricing policies under demand uncertainty. We also consider the challenge of price experimentation with limited price adjustments, and derive insights into the structure and performance of price experimentation strategies in the presence of adjustment constraints.

In the final chapter, we address the question of how a seller should distribute his goods for sale, by studying a general resource allocation problem with uncertain demand. In this problem, a seller must decide how to allocate units of his products across multiple selling venues, with the goal of maximizing his total sales. We design policies that learn near-optimal allocations over a sequence of time periods, without prior knowledge of the demand at each venue. The policies developed here advance the state of the art for the online resource allocation problem, and our analysis highlights the ways in which effective allocation strategies should balance venue exploration with best-guess optimal allocation.

1.1. Pricing Under Demand Uncertainty

In Chapter 2, we consider the problem of a retailer choosing a price at which to sell a new product, with the objective of maximizing his total expected revenue. If the retailer had perfect information about the demand for the product $d(p)$ as a function of the price level p, then determining the revenue-maximizing price for the good would in principle be a straightforward optimization problem: the seller would simply compute $\arg\max_p\{pd(p)\}$. However, perfect information about the demand curve is typically not available in practice, because the relationship between price and customer purchase probability is generally not known to the seller in advance. To address this problem, we consider *dynamic pricing* strategies,

in which a seller adjusts the price of the good to gain information about the demand curve, and then exploits this information to offer a near-optimal selling price.

In this chapter, we consider two fundamental questions that apply to virtually any dynamic pricing formulation. First, what is the value of knowing the demand curve; in other words, what is the magnitude of the revenue lost due to uncertainty about the relationship between price and demand? Secondly, how should good pricing strategies balance price experimentation (exploration) and best-guess optimal pricing (exploitation)? We will see that the answers to both of these questions depend intrinsically on the nature of the demand uncertainty facing the seller.

To investigate these questions, we consider dynamic pricing under a general parametric model of demand uncertainty. We measure the performance of a pricing strategy in this model in terms of the *regret*: the difference between the expected revenue gained by the pricing strategy, and the revenue gained by an omniscient strategy that has full information about the demand curve in advance. We classify the order of regret of optimal pricing strategies, by describing policies that achieve provable worst-case performance guarantees, and by proving lower bounds on the minimum regret achievable by an *arbitrary* pricing strategy, which match our upper bounds to within a constant factor. In the course of this analysis, we illustrate important principles for the design of optimal pricing heuristics, and relate the best achievable performance for this problem to characteristics of the underlying demand model.

1.2. Pricing with Minimal Adjustments

In the previous chapter, we considered the problem of pricing under demand uncertainty, with the goal of maximizing total revenue. Implicit in our problem, and in nearly all dynamic pricing formulations, is the need for the seller to adjust the offer price as he gains information about the demand curve over time. Indeed, if the seller is constrained to offer a single fixed price throughout the entire selling season, then he will likely sacrifice a large amount of potential revenue, because without knowledge of the demand curve, the chosen price will almost certainly be sub-optimal. Accordingly, virtually all existing dynamic pricing policies place no constraints on the number of price adjustments, allowing the seller to change prices as many times as is necessary to perform demand learning and price exploitation.

This lack of restriction on price adjustments stands in contrast to a large body of evidence suggesting that frequent price changes are inherently undesirable. From the standpoint of the seller, there is a significant amount of evidence (see, for example, Levy et al. (1997), Levy et al. (1998), Zbaracki et al. (2004)) showing that the costs of implementing frequent price changes in a traditional retail setting can amount to a considerable portion of the seller's net margins. Even in an online retail setting, where price adjustments may be less costly (Brynjolfsson and Smith (2000)), there is evidence to suggest that frequent fluctuations in price may be undesirable to the customer. For example, Amazon.com was involved in a controversy after frequent price experiments lead to accusations of discrimination and negative press coverage regarding its pricing practices (Weiss and Mehrotra (2001)).

Motivated by these concerns, this chapter seeks to understand the fundamental

limit on the minimum number of price adjustments needed for optimal dynamic pricing. Perhaps the most straightforward means of understanding the relationship between adjustment constraints and regret would be to establish bounds on the regret of a policy in terms of the time horizon and a hard switching constraint. Such an *absolute* measure of performance, however, would make it difficult to distinguish between the loss in revenue due to demand uncertainty, and the revenue loss due to switching constraints. Thus, we consider a *relative* performance measure, by using the performance of the optimal unconstrained pricing policy as a natural benchmark. Specifically, we will consider the *minimal switching rate* of a problem instance, which we define to be the minimum number of price changes necessary for a switching-constrained policy to match the regret of the optimal unconstrained policy. We derive lower bounds on the minimum number of price adjustments needed for an online pricing policy to achieve the performance guarantees established in Chapter 2, and we describing online pricing strategies that achieve the optimal rate of performance, while adjusting pricing the minimal number of times. This analysis classifies the minimal switching rate for the online pricing problem, and gives a number of insights into the design of policies that jointly maximize revenue, while minimizing price adjustments.

1.3. Dynamic Resource Allocation

In Chapters 2 and 3, we considered a monopolist selling an unlimited supply of a product from a single venue, and considered optimization over the set of all online pricing strategies. In Chapter 4, we depart from this line of investigation to consider an alternative problem in the area of revenue management under demand uncertainty. In this chapter, we consider the problem of a retailer selling a fixed stock of inventory from *multiple* venues, with the goal of maximizing the total

number of sales across all venues. In contrast to the previous two chapters, the decision variable for the seller in this problem is no longer which price to charge, but rather which *allocation* of inventory to venues will result in the largest number of expected sales. As in the previous two chapters, we note that if the seller had perfect information about the demand at each venue, then determining the optimal allocation would be a straightforward optimization problem. Thus, we consider the natural extension in which the seller has no prior information about the demand at each venue, and must offer a sequence of allocations to maximize total overall sales.

In this chapter, we describe policies for the online resource allocation problem with stochastic demand. The primary difficulty for a retailer operating in this setting is the *censored* natured of the feedback that he receives. For example, if the retailer decides to allocate n units of his product to a given venue at the beginning of a selling period, and then observes at the end of the selling period that all n units of the product were sold, then it is impossible for him to determine whether the demand for the product was exactly n, or whether the demand for the product exceeded n by a large margin. This issue leads to a natural exchange between estimating the demand for products, and allocating goods to maximize consumption. On the one hand, since demand observations are censored, a policy must periodically over-allocate to each venue, to maintain an accurate estimate of the demand. On the other hand, excessive over-allocation comes at a cost in regret, because in so doing, a policy is most likely performing a suboptimal allocation.

We describe a natural class of policies for the online resource allocation problem, which carefully balance the exploration / exploitation tradeoff described above to achieve worst-case regret that is nearly optimal. These policies represent an ad-

vance in the state of the art for this problem; to our knowledge, these policies have the best-known performance guarantees for the online resource allocation problem with stochastic demand, and are the first policies whose regret guarantees match known lower bounds for this problem, up to sub-logarithmic factors in the number of time periods, and polynomial factors in the number of venues and units of resource. In addition to these theoretical results, we show that our policies perform well empirically, by evaluating them on both synthetic demand data, and demand data calibrated to a set of usage data from a local vehicle-sharing operation.

Chapter 2

Dynamic Pricing

2.1. Introduction

In this chapter, we consider the problem of a retailer choosing a price at which to sell a new product, with the objective of maximizing his expected revenue. If the retailer had full information about the demand at every price level, then he could determine the revenue-maximizing price for the good. However, perfect information about the demand curve is rarely available practice, and so we address the natural question of how a seller should set his prices to maximize his overall revenue in the face of demand uncertainty.

We will concern ourselves with two primary questions. First, how much revenue will be lost by the seller due to lack of information about the demand curve for the product, and secondly, how should we design pricing strategies to balance price experimentation for demand learning, and best-guess optimal pricing for revenue maximization? We investigate these questions under a general parametric

model of demand uncertainty. To measure the performance of a pricing strategy in this model, we compare the revenue generated by that strategy against the revenue generated by the optimal pricing strategy, which always offers the revenue-maximizing price. The difference in these two revenues is called the *regret*, and our goal will be to design pricing strategies which have small regret in a worst-case sense.

We classify the order of the regret of the optimal pricing policy under two scenarios: a scenario in which the demand curves satisfy a set of general assumptions, and a scenario in which the demand curves satisfy an additional "well-separated" condition, under which the demand curve for any one parameter value stochastically dominates the demand curve for a different value.

By analyzing the performance of pricing policies under these two scenarios, we derive a number of insights into the above questions. For the general case, we show that the worst-case T-period regret of an arbitrary pricing policy must be $\Omega(\sqrt{T})$.[1] To establish this result, we show that a parametric family of demand curves may include an "uninformative" price, and that the presence of uninformative prices makes demand learning intrinsically difficult for the seller. To complement this lower bound, we describe a pricing policy that achieves a regret guarantee of $\mathcal{O}(\sqrt{T})$ across all problem instances. To achieve the optimal order of regret, this policy hedges against the difficulties imposed by uninformative prices, by conducting a carefully chosen amount of dedicated exploratory pricing, during which it sacrifices immediate revenues to gain information about the demand curve. Thus, we address the above questions by showing that in a general parametric setting, the amount of revenue lost due to demand uncertainty is $\Theta(\sqrt{T})$, and by describing

[1] We use the notation $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ to represent upper and lower bounds, respectively, on the performance measure of interest (see Knuth (1997) for more details).

an explicit balance between price experimentation and best-guess optimal pricing that achieves the optimal rate of regret.

To contrast the general case discussed above, we consider a special scenario in which the demand curves satisfy a well-separated condition that precludes the possibility of an uninformative price. We show that in this case, dynamic pricing is intrinsically easier, in that the worst-case regret of the optimal pricing policy is $\mathcal{O}(\log T)$. We describe a "greedy" pricing policy that achieves this rate of regret, by simultaneously estimating the demand curve while offering the best-guess optimal price. Intuitively, in the absence of uninformative prices, a seller can learn from customer responses *at every price level*, making simultaneous exploration and exploitation possible, and leading to regret that is much smaller than in the general case. Additionally, we show that the stochastic nature of demand forces the worst-case regret of any pricing policy to be $\Omega(\log T)$, establishing that our greedy policy achieves the optimal order of regret. Thus, our study of this case exhibits a scenario in which the magnitude of the revenue lost due to demand uncertainty is significantly different, and in which the optimal rate of regret is achieved by a policy with entirely different structure.

In summary, our analysis provides a rich regret profile of dynamic pricing under a general parametric model of demand uncertainty. Moreover, our results demonstrate an intrinsic connection between the optimal order of regret and structural properties of the demand model, and provide insights which guide the design of provably effective pricing heuristics. We give a detailed outline of our results and their organization in Section 2.1.3; below, we describe the details of our dynamic pricing framework.

2.1.1 The Model

We assume that customers arrive in discrete time steps. For each $t \geq 1$, when the t^{th} customer arrives, he is quoted a price by the seller, and then decides whether to purchase the good at that price based on his willingness-to-pay V_t. We assume that $\{V_t : t \geq 1\}$ are independent and identically distributed random variables whose common distribution function belongs to some family parameterized by $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^n$. Let $d(\,\cdot\,;\mathbf{z}) : \mathbb{R}_+ \to \mathbb{R}_+$ denote the complementary cumulative distribution function of V_t, that is, for all $p \geq 0$,

$$d(p\,;\mathbf{z}) = \Pr_{\mathbf{z}}\{V_t \geq p\} \ . \tag{2.1}$$

We assume that each customer purchases the product if and only if his willingness-to-pay is at least as large as the product price. Thus, we will also refer to $d(\,\cdot\,;\mathbf{z})$ as the demand curve because it determines the probability that the customer will purchase a product at a given price. For any $p \geq 0$, the expected revenue $r(p\,;\mathbf{z})$ under the price p is given by

$$r(p\,;\mathbf{z}) = p\,d(p\,;\mathbf{z}) \ . \tag{2.2}$$

We will restrict our attention to families of demand curves for which the corresponding revenue function $r(\,\cdot\,;\mathbf{z})$ has a unique maximizer.

We will consider a *problem class* $\mathcal{C}$ to be a tuple $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$, where $\mathcal{Z} \subset \mathbb{R}^n$ is a compact and convex parameter set, $\mathcal{P} = [p_{min}, p_{max}]$ is a closed pricing interval with $p_{min} \geq 0$, and $d : \mathcal{P} \times \mathcal{Z} \to [0,1]$ is a smooth parametric family of demand curves such that $p \mapsto d(p;\mathbf{z})$ is non-increasing for each $\mathbf{z} \in \mathcal{Z}$. Finally, we assume that $p^*(\mathbf{z}) \in \mathcal{P}$ for all $\mathbf{z} \in \mathcal{Z}$.

For any $t \geq 1$, we denote by $\mathbf{y}_t = (y_1, \ldots, y_t) \in \{0,1\}^t$ a history of customer purchasing decisions, where $y_\ell = 1$ if the ℓ^{th} customer decided to purchase the

product, and $y_\ell = 0$ otherwise. A *pricing policy* $\psi = (\psi_1, \psi_2, \ldots)$ is a sequence of functions such that $\psi_t : \{0,1\}^{t-1} \to \mathcal{P}$ sets the price in period t based on the observed purchasing decisions in the preceding $t-1$ periods. To model the relationship between a pricing policy ψ and customer behavior, we consider the distribution $Q_t^{\psi,\mathbf{z}}$ on t-step customer response histories induced by the policy ψ, which we define as follows. For any policy ψ and $\mathbf{z} \in \mathcal{Z}$, let $Q_t^{\psi,\mathbf{z}} : \{0,1\}^t \to [0,1]$ denote the probability distribution of the customer responses $\mathbf{Y}_t = (Y_1, \ldots, Y_t)$ in the first t periods when the policy ψ is used and the underlying parameter is $\mathbf{z}$; that is, for all $\mathbf{y}_t = (y_1, \ldots, y_t) \in \{0,1\}^t$,

$$Q_t^{\psi,\mathbf{z}}(\mathbf{y}_t) = \prod_{\ell=1}^{t} d(p_\ell; \mathbf{z})^{y_\ell} (1 - d(p_\ell; \mathbf{z}))^{1-y_\ell} \ , \tag{2.3}$$

where $p_\ell = \psi_\ell(\mathbf{y}_{\ell-1})$ denotes the price in period ℓ under the policy ψ. It will also be convenient to consider the distribution on customer responses to a sequence of fixed prices $\mathbf{p} = (p_1, \ldots, p_k) \in \mathcal{P}^k$, rather than the prices set by a pricing policy. We represent these distributions by

$$Q^{\mathbf{p},\mathbf{z}}(\mathbf{y}) = \prod_{\ell=1}^{k} d(p_\ell; \mathbf{z})^{y_\ell} (1 - d(p_\ell; \mathbf{z}))^{1-y_\ell},$$

where $\mathbf{y} \in \{0,1\}^k$, and p_ℓ denotes the ℓ^{th} component of the price vector $\mathbf{p} \in \mathcal{P}^k$.

Finally, we formalize the performance measure used to evaluate pricing policies. For a problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$, a parameter $\mathbf{z} \in \mathcal{Z}$, a policy ψ setting prices in $\mathcal{P}$, and a time horizon $T \geq 1$, the T-period cumulative regret under ψ is defined to be

$$\text{Regret}(\mathbf{z}, \mathcal{C}, T, \psi) = \sum_{t=1}^{T} \mathbb{E}_{\mathbf{z}} \left[r\left(p^*(\mathbf{z}); \mathbf{z}\right) - r(P_t; \mathbf{z}) \right] \ ,$$

where $P_1, P_2, \ldots$ denotes the sequence of prices under the policy ψ, and $\mathbb{E}_{\mathbf{z}}[\,\cdot\,]$ denotes the expectation when the underlying parameter vector of the willingness-to-pay distribution is $\mathbf{z}$. We note that when the parameter $\mathbf{z}$ is known, minimizing

the T-period cumulative regret is equivalent to maximizing the total expected revenue over T periods.

As a convention, we will denote vectors in bold, and scalars in regular font. A random variable is denoted by an uppercase letter while its realized values are denoted in lowercase. We denote by $\mathbb{R}_+$ the set of non-negative real numbers, while $\mathbb{R}_{++}$ denotes the set of positive numbers. We use $\|\cdot\|$ to denote the Euclidean norm, and for any set $S \subset \mathbb{R}^n$ and any element $y \in \mathbb{R}^n$, we define $S - y = \{x - y : x \in S\}$. We use $\log(\cdot)$ to denote the natural logarithm. For any symmetric matrix $\mathbf{A}$, let $\lambda_{\min}(\mathbf{A})$ denote its smallest eigenvalue.

Before proceeding with a review of the relevant literature, we note several assumptions about the retail environment implicit in our model. We assume that the seller is a monopolist offering an unlimited supply of a nonperishable single product, with no marginal cost of production. We also assume that the seller has the ability to adjust prices and receive feedback in real time, at the level of individual customers. Although quite stylized, this model allow us to conduct a simple and tractable analysis of demand learning under parametric uncertainty, and clearly illustrate some of the difficulties facing a seller in such a scenario. Moreover, these assumptions have been adopted by previous works (e.g., Cope, 2006; Kleinberg and Leighton, 2003; Carvalho and Puterman, 2005; Besbes and Zeevi, 2009), and provide a convenient framework in which to study dynamic pricing. We now proceed to place this work in context with a review of the existing literature.

2.1.2 Literature Review

Many recent studies in the dynamic pricing literature consider heuristics for pricing under parametric notions of demand uncertainty. Carvalho and Puterman (2005) consider a dynamic pricing formulation in which the demand has a logistic distribution with two unknown parameters. The authors perform a numerical evaluation of several heuristic strategies, and demonstrate that a "one-step lookahead" policy, which sacrifices immediate revenue to compute a better estimate of the unknown demand parameters, outperforms a myopic policy. Lobo and Boyd (2003) consider a linear demand model with Gaussian noise, and investigate through numerical experiments a "price-dithering" policy, which adds a random perturbation to the myopically optimal price. Bertsimas and Perakis (2003) consider a similar demand model, and show through numerical experiments that approximate dynamic programming policies that balance immediate revenue rewards with long-term learning can outperform a myopic policy. The above works provide empirical evidence that, in a variety of settings, pricing policies that perform some sort of active exploration will outperform myopically greedy policies, indicating that there is some intrinsic value to price experimentation. However, none of these works establish provable performance guarantees for the heuristics described.

More aligned with this work, several other recent papers conduct more theoretical investigations of the value of price experimentation. In Besbes and Zeevi (2009), the authors consider demand learning under an uncapacitated Bernoulli demand model, in which the seller knows the initial demand curve. At some point in time unknown to the seller, the demand curve switches to a different (but known in advance) function of the price. The authors show that when the two demand curves satisfy a well-separated condition, a myopically greedy policy is optimal.

Additionally, they show that when the demand curves intersect, corresponding to the presence of an uninformative price, then the magnitude of the worst-case regret is larger, and exhibit an optimal policy that performs some forced exploration. Our work in this chapter is thematically related to Besbes and Zeevi (2009), in that we conduct a similar analysis of the worst-case regret under a well-separated versus intersecting demand model, and in that we consider myopic versus forced exploration policies. One may view our work as complementary to Besbes and Zeevi (2009), in that we consider demand learning in a stationary, parameter learning framework, while they consider a similar learning problem under a non-stationary, two-hypothesis setting.

A second related paper is Besbes and Zeevi (2008), in which the authors consider demand learning in a general parametric (as well as non-parametric) setting, and present policies based on maximum likelihood estimation. They suggest that the structure and performance of a rate-optimal pricing policy should be different in the general versus the well-separated case, but they provide the same lower bound on the performance measure for both cases. We complement the theme of their work by exhibiting a dynamic pricing formulation in which the regret profiles between the two cases are entirely different. Specifically, we prove in Theorem 2.3.1 that in the general case, the worst case regret under an arbitrary policy must be at least $\Omega(\sqrt{T})$, and that in the well-separated case, there is a policy whose regret is at most $\mathcal{O}(\log T)$ across all problem instances (Theorem 2.4.8). Aside from these thematic similarities, several crucial features differentiate this work from ours, including the presence of a known, finite time horizon, the presence of a known capacity constraint, and a performance measure that is parameterized by initial capacity and demand rate, rather than the time horizon. The aforementioned differences make direct comparisons difficult, and lead to a significantly different

analysis.

Pricing under parametric demand uncertainty has been considered under a variety of alternative models; see, for example, Aviv and Pazgal (2002), Araman and Caldenty (2005), and Farias and Van Roy (2007) for a recent line of investigation in a capacitated, Bayesian framework, and Harrison et al. (2010) for a formulation that investigates "uninformative prices" in a two-hypothesis, Bayesian model. Cope (2006) and Kleinberg and Leighton (2003) consider non-parametric approaches, and notably, Kleinberg and Leighton (2003) derive regret bounds for the non-parametric case that are comparable to our bounds for the general case. However, it is worth noting that the policies considered in that work have significantly different structure from the ones considered here: their policies operate by experimenting with fine mesh of prices across the entire pricing interval, whereas the policies considered here estimate the demand parameters from a relatively small number of test prices. Thus, by focusing on the design and performance of pricing strategies that utilize parametric information about the demand curve, we provide complementary insights into effective dynamic pricing under an alternative formulation of demand uncertainty. For further examples of dynamic pricing and a comprehensive review of the subject, we refer the reader to Talluri and van Ryzin (2004) and Bitran and Caldentey (2003).

Finally, we note that our pricing problem can be viewed as a special case of a general stochastic optimization problem, in which one wishes to iteratively approximate the minimizer of an unknown function, based only on noisy evaluation of the function at points inside a (usually uncountable) feasible set. A full review of the literature on this topic is beyond the scope of this dissertation; however, several notable references from the stochastic approximations literature include

Kiefer and Wolfowitz (1967), Fabian (1967), and more recently, Broadie et al. (2009) and Cope (2009), which examine the convergence properties of stochastic gradient-descent type schemes. Another standard approach is to apply the classical multi-armed bandit algorithm (Lai and Robbins (1985a) and Auer et al. (2002a)) to the general stochastic optimization setting via a discretization approach; see, for example, Agrawal (1995) and Auer et al. (2007), and Kleinberg and Leighton (2003) for an application of these techniques in the context of dynamic pricing. As a key distinction, we note that both of the aforementioned techniques are *non-parametric*, and thus the parametric, maximum-likelihood-based policies presented in this chapter are significantly different in both their structure and analysis.

We now proceed with a summary of our main contributions and organization.

2.1.3 Contributions and Organization

One of the main contributions of our work is a complete regret profile for the dynamic pricing problem under a general parametric choice model. In Section 2.3.1, we prove in Theorem 2.3.1 that in the general case, the regret of an arbitrary pricing policy is $\Omega(\sqrt{T})$, by exploiting the presence of "uninformative prices," which force a tradeoff between reducing uncertainty about the parameters of the demand curve and exploiting the best-guess optimal price. In Section 2.3.2, we present a pricing policy based on maximum-likelihood estimation whose regret is $\mathcal{O}(\sqrt{T})$ across all problem instances (Theorem 2.3.6).

In Section 2.4, we consider dynamic pricing when the family of demand curves satisfies a "well-separated" condition, which precludes the presence of uninformative prices. We show that in this scenario, the regret of the optimal policy is $\Theta(\log T)$. In Section 2.4.1, we establish a regret lower bound of $\Omega(\log T)$ for all poli-

cies (Theorem 2.4.1), based on a Cramér-Rao-type inequality. We also describe a pricing policy based on maximum-likelihood estimates (MLE) that achieves a matching $\mathcal{O}(\log T)$ upper bound (Theorem 2.4.8). The key observation is that in the well-separated case, demand learning is easier, in that a pricing policy can learn about the parameters of the demand curve from customer responses to any price.

As a by product of our analysis, we also provide a novel large deviation inequality and bound on mean squared errors for a maximum-likelihood estimator based on samples that are dependent and not identically distributed (Theorem 2.4.7). The proof techniques used here are of independent interest because they can be extended to other MLE-based online learning strategies.

2.2. Assumptions and Examples

Recall that a *problem class* $\mathcal{C}$ is a tuple $(\mathcal{P}, \mathcal{Z}, d)$, where $\mathcal{P} = [p_{min}, p_{max}] \subset \mathbb{R}_+$ is a feasible pricing interval, $\mathcal{Z} \subset \mathbb{R}^n$ is a compact and convex feasible parameter set, and $d : \mathcal{P} \times \mathcal{Z} \to [0, 1]$ is a parametric family of smooth demand functions such that $p \mapsto d(p; \mathbf{z})$ is non-increasing for each $\mathbf{z} \in \mathcal{Z}$. Throughout this work, we restrict our attention to problem classes $\mathcal{C}$ satisfying the following basic assumptions.

Assumption 1 (Basic Assumptions)**.** There exists positive constants d_{min}, d_{max}, L, and c_r such that

(a) $0 < d_{min} \leq d(p; \mathbf{z}) \leq d_{max} < 1$ for all $p \in \mathcal{P}$ and $\mathbf{z} \in \mathcal{Z}$.

(b) The revenue function $p \mapsto r(p; \mathbf{z})$ has a unique maximizer $p^*(\mathbf{z}) \in \mathcal{P}$.

(c) The function $\mathbf{z} \mapsto p^*(\mathbf{z})$ is L-Lipschitz, that is, $|p^*(\mathbf{z}) - p^*(\bar{\mathbf{z}})| \leq L \|\mathbf{z} - \bar{\mathbf{z}}\|$ for all $\mathbf{z}, \bar{\mathbf{z}} \in \mathcal{Z}$.

(d) The revenue function $p \mapsto r(p;\mathbf{z})$ is twice differentiable with $\sup_{p\in\mathcal{P},\mathbf{z}\in\mathcal{Z}} |r''(p;\mathbf{z})| \le c_r$.

Under Assumption 1(a), the demand is bounded away from zero and one on the pricing interval; that is, we will not offer prices at which customers will either purchase or decline to purchase with probability one. Assumption 1(b) is self-explanatory, and Assumption 1(c) says that if we vary the parameter $\mathbf{z}$ by a small amount, then the optimal price $p^*(\mathbf{z})$ will not vary too much. Assumption 1(d) imposes a smoothness condition on the demand curve $p \mapsto d(p;\mathbf{z})$.

In addition to these structural assumptions about the demand curve, we will also impose the following statistical assumption about the family of distributions $\{Q^{\mathbf{p},\mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$.

Assumption 2 (Statistical Assumption)**.** There exists a vector of exploration prices $\bar{\mathbf{p}} \in \mathcal{P}^k$ such that the family of distributions $\{Q^{\bar{\mathbf{p}},\mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$ is identifiable, that is, $Q^{\bar{\mathbf{p}},\mathbf{z}}(\,\cdot\,) \neq Q^{\bar{\mathbf{p}},\bar{\mathbf{z}}}(\,\cdot\,)$ whenever $\mathbf{z} \neq \bar{\mathbf{z}}$. Moreover, there exists a constant $c_f > 0$ depending only on the problem class $\mathcal{C}$ and $\bar{\mathbf{p}}$ such that $\lambda_{min}\{\mathbf{I}(\bar{\mathbf{p}},\mathbf{z})\} \ge c_f$ for all $\mathbf{z} \in \mathcal{Z}$, where $\mathbf{I}(\bar{\mathbf{p}},\mathbf{z})$ denotes the *Fisher information matrix* given by

$$\big[\mathbf{I}(\bar{\mathbf{p}},\mathbf{z})\big]_{i,j} = \mathbb{E}_{\mathbf{z}}\left[-\frac{\partial^2}{\partial z_i \partial z_j}\log Q^{\bar{\mathbf{p}},\mathbf{z}}(\mathbf{Y})\right] = \sum_{k=1}^{n} \frac{\left\{\frac{\partial}{\partial z_i} d(\bar{p}_k,\mathbf{z})\right\} \times \left\{\frac{\partial}{\partial z_j} d(\bar{p}_k,\mathbf{z})\right\}}{d(\bar{p}_k,\mathbf{z})(1-d(\bar{p}_k,\mathbf{z}))} \ .$$

Assumption 2 is a standard assumption, which guarantees that we can estimate the demand parameter based on the purchase observations at the exploration prices $\bar{\mathbf{p}}$ (see, for example, Besbes and Zeevi (2008)). As shown in the following examples, Assumptions 1 and 2 encompass many families of parametric demand curves (see Talluri and van Ryzin (2004) for additional examples).

Example 2.2.1 (Logit Demand)**.** Let $\mathcal{P} = [1/2, 2] \subset \mathbb{R}$, $\mathcal{Z} = [1, 2] \times [-1, 1] \subset \mathbb{R}^2$

and let

$$d(p, \mathbf{z}) = \frac{e^{-z_1 p - z_2}}{1 + e^{-z_1 p - z_2}}$$

be the family of logit demand curves. It is straightforward to check that $(\mathcal{P}, \mathcal{Z}, d)$ satisfies the conditions stated in Assumption 1 with $d_{\min} = e^{-5}/(1 + e^{-5})$, $d_{\max} = e^{1/2}/(1 + e^{1/2})$, $L = 2 + \log(2)$, and $c_r = 2e$. It is also straightforward to check that for any $\bar{\mathbf{p}} = (\bar{p}_1, \bar{p}_2) \in \mathcal{P}^2$ with $\bar{p}_1 \neq \bar{p}_2$, the associated family $\{Q^{\bar{\mathbf{p}}, \mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$ is identifiable. Moreover, for any $p \in \mathbb{R}_+$ and $\mathbf{z} \in \mathcal{Z}$, we have that

$$\frac{\partial}{\partial z_1} d(p, \mathbf{z}) = -p d(p, \mathbf{z})(1 - d(p, \mathbf{z})) \quad \text{and} \quad \frac{\partial}{\partial z_2} d(p, \mathbf{z}) = -d(p, \mathbf{z})(1 - d(p, \mathbf{z})) \ ,$$

which implies that the Fisher information matrix is given by

$$\mathbf{I}(\bar{\mathbf{p}}, \mathbf{z}) = d(\bar{p}_1, \mathbf{z})(1 - d(\bar{p}_1, \mathbf{z})) \begin{pmatrix} \bar{p}_1^2 & \bar{p}_1 \\ \bar{p}_1 & 1 \end{pmatrix} + d(\bar{p}_2, \mathbf{z})(1 - d(\bar{p}_2, \mathbf{z})) \begin{pmatrix} \bar{p}_2^2 & \bar{p}_2 \\ \bar{p}_2 & 1 \end{pmatrix}$$

By applying the trace-determinant formula, we can show that for all $\mathbf{z} \in \mathcal{Z}$,

$$\lambda_{min}\{\mathbf{I}(\bar{\mathbf{p}}, \mathbf{z})\} \geq \frac{(\bar{p}_1 - \bar{p}_2)^2}{\bar{p}_1^2 + \bar{p}_2^2 + 2} \cdot d_{min}^2 (1 - d_{max})^2 > 0$$

Example 2.2.2 (Linear Demand)**.** Let $\mathcal{P} = [1/3, 1/2]$, let $\mathcal{Z} = [2/3, 3/4] \times [3/4, 1]$, and let

$$d(p; \mathbf{z}) = z_1 - z_2 p$$

be a linear demand family. Then it is straightforward to check that this family satisfies Assumption 1 with $d_{\min} = 1/6$, $d_{\max} = 1/2$, $L = 2$, and $c_r = 2$. Moreover, for any $\bar{\mathbf{p}} = (\bar{p}_1, \bar{p}_2) \in \mathcal{P}^2$ with $\bar{p}_1 \neq \bar{p}_2$, the associated family $\{Q^{\bar{\mathbf{p}}, \mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$ is identifiable. A similar computation shows that the Fisher information matrix is given by

$$\mathbf{I}(\bar{\mathbf{p}}, \mathbf{z}) = \frac{1}{d(\bar{p}_1, \mathbf{z})(1 - d(\bar{p}_1, \mathbf{z}))} \begin{pmatrix} 1 & \bar{p}_1 \\ \bar{p}_1 & \bar{p}_1^2 \end{pmatrix} + \frac{1}{d(\bar{p}_2, \mathbf{z})(1 - d(\bar{p}_2, \mathbf{z}))} \begin{pmatrix} 1 & \bar{p}_2 \\ \bar{p}_2 & \bar{p}_2^2 \end{pmatrix} \ ,$$

and using the same argument as above, we can show that

$$\lambda_{min}\{\mathbf{I}(\bar{\mathbf{p}}, \mathbf{z})\} \geq \frac{(\bar{p}_1 - \bar{p}_2)^2}{\bar{p}_1^2 + \bar{p}_2^2 + 2} \cdot \frac{1}{d_{max}^2(1 - d_{min})^2} > 0 \ .$$

Example 2.2.3 (Exponential Demand)**.** Let $\mathcal{P} = [1/2, 1]$, let $\mathcal{Z} = [1, 2] \times [0, 1]$, and let

$$d(p; \mathbf{z}) = e^{-z_1 p - z_2}$$

be an exponential demand family. Then by the same techniques used in Examples 2.2.1 and 2.2.2, one may check that this problem class satisfies Assumptions 1 and 2, and that the associated family $\{Q^{\bar{\mathbf{p}},\mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$ is identifiable for any $\bar{\mathbf{p}} = (\bar{p}_1, \bar{p}_2) \in \mathcal{P}^2$ for which $\bar{p}_1 \neq \bar{p}_2$.

We now discuss an important observation that will motivate the design of pricing policies in our model. Suppose that the unknown model parameter vector is $\mathbf{z}$, and let $\hat{\mathbf{z}}$ denote some estimate of $\mathbf{z}$. We might consider pricing the product at $p^*(\hat{\mathbf{z}})$, which is optimal with respect to our estimate. When $\hat{\mathbf{z}}$ is close to the true parameter vector, we would expect that $p^*(\hat{\mathbf{z}})$ yields a near optimal revenue. We make this intuition precise in the following corollary, which establishes an upper bound on the loss in revenue from inaccurate estimation.

Corollary 2.2.4 (Revenue Loss from Inaccurate Estimation)**.** *For any problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ satisfying Assumption 1 and for any $\mathbf{z}, \hat{\mathbf{z}} \in \mathcal{Z}$,*

$$r(p^*(\mathbf{z}); \mathbf{z}) - r(p^*(\hat{\mathbf{z}}); \mathbf{z}) \ \leq \ c_r \, L^2 \, \|\mathbf{z} - \hat{\mathbf{z}}\|^2 \ .$$

Proof. First we will show that as a consequence of Assumption 1(b) and Assumption 1(d), we have that for $\mathbf{z} \in \mathcal{Z}$ and $p \in \mathcal{P}$,

$$0 \leq r(p^*(\mathbf{z}); \mathbf{z}) - r(p; \mathbf{z}) \leq c_r \, (p^*(\mathbf{z}) - p)^2 \ .$$

The result then follows from Assumption 1(c) (the Lipschitz continuity of the optimal price).

We will establish the quadratic inequality for $p > p^*(\mathbf{z})$. The same argument applies to the case where $p < p^*(\mathbf{z})$. For any $u \in \mathbb{R}_+$, let $r'(u;\mathbf{z})$ and $r''(u;\mathbf{z})$ denote the first and second derivatives of the revenue function at u, respectively. Since $r'(p^*(\mathbf{z});\mathbf{z}) = 0$, it follows that

$$\begin{aligned} |r(p^*(\mathbf{z});\mathbf{z}) - r(p;\mathbf{z})| &= \left| \int_{p^*(\mathbf{z})}^{p} \int_{p^*(\mathbf{z})}^{t} r''(u;\mathbf{z})\, du\, dt \right| \\ &\leq \sup_{u\in\mathcal{P}} |r''(u;\mathbf{z})| \int_{p^*(\mathbf{z})}^{p} \int_{p^*(\mathbf{z})}^{t} du\, dt \;=\; \frac{1}{2} \sup_{u\in\mathcal{P}} |r''(u;\mathbf{z})|\, (p^*(\mathbf{z}) - p)^2 \\ &\leq c_r (p^*(\mathbf{z}) - p)^2 \end{aligned}$$

□

Corollary 2.2.4 suggests a method for constructing a pricing policy with low regret. We construct an estimate of the underlying parameter based on the observed purchase history, then offer the greedy optimal price according to this estimate. If our estimate has a small mean square error, then we expect that the loss in revenue should also be small. However, the variability of our estimates depends on the past prices offered. As we will see, there is a nontrivial tradeoff between pricing to form a good estimate (exploration) and pricing near the greedy optimal (exploitation), and the optimal balance between these two will be quite different depending on the nature of the demand uncertainty facing the seller.

2.3. The General Case

In this section, we consider dynamic pricing under the general parametric model satisfying Assumptions 1 and 2. In Section 2.3.1, we show that the worst-case regret

of any pricing policy must be at least $\Omega(\sqrt{T})$, by constructing a problem class with an "uninformative price" that impedes demand learning. Then, in Section 2.3.2, we describe a pricing policy based on maximum likelihood estimation whose regret is $\mathcal{O}(\sqrt{T})$ across all problem instances, thus establishing that the order of regret for the optimal pricing policy in the general case is $\Theta(\sqrt{T})$.

2.3.1 A Lower Bound for the General Case

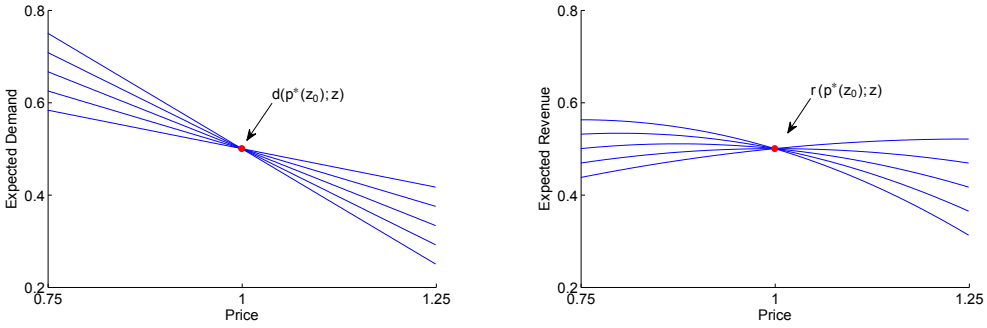
In this section, we establish a lower bound on the T-period cumulative regret for the general case. The main result is stated in the following theorem.

Theorem 2.3.1 (General Regret Lower Bound)**.** *Define a problem class* $\mathcal{C}_{\mathsf{GenLB}} = (\mathcal{P}, \mathcal{Z}, d)$ *by letting* $\mathcal{P} = [3/4, 5/4]$, $\mathcal{Z} = [1/3, 1]$, *and* $d(p; z) = 1/2 + z - zp$. *Then for any policy* ψ *setting prices in* $\mathcal{P}$, *and any* $T \geq 2$, *there exists a parameter* $z \in \mathcal{Z}$ *such that*

$$\mathrm{Regret}(z, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) \geq \frac{\sqrt{T}}{48^3} \ .$$

Using the same proof technique as in Example 2.2.2, one can show that the problem class $\mathcal{C}_{\mathsf{GenLB}}$ satisfies Assumptions 1 and 2, with $d_{min} = 1/4, d_{max} = 3/4, p^*(z) = (1 + 2z)/(4z), L = 3$, and $c_r = 2$. Before we proceed to the proof of Theorem 2.3.1, let us discuss the intuition underlying our arguments. Figure 2.1(a) shows examples of demand curves in the family given by $\mathcal{C}_{\mathsf{GenLB}}$. Note that for all $z \in \mathcal{Z}$, $d(1; z) = 1/2$, and thus all demand curves in this family intersect at common price $p = 1$. Note also that this price is the optimal price for some demand curve in this family, that is, $p^*(z_0) = 1$ for $z_0 = 1/2$ (see Figure 2.1(b) for examples of the revenue curves). Since the demand is the same at $p^*(z_0)$ regardless of the underlying parameter, the price $p^*(z_0)$ is "uninformative," in that no policy can gain information about the value of the parameter while pricing at

$p^*(z_0)$. To establish Theorem 2.3.1, we show that uninformative prices lead to a tension between demand learning (exploration) and best-guess optimal pricing (exploitation), which forces the worst-case regret of any policy to be $\Omega(\sqrt{T})$. This tension is made precise in two lemmas. We show in Lemma 2.3.3 that for a policy to reduce its uncertainty about the unknown demand parameter, it must necessarily set prices away from the uninformative price $p^*(z_0)$, and thus incur large regret when the underlying parameter is z_0. Then, in Lemma 2.3.4, we show that any policy that does not reduce its uncertainty about the demand parameter z must also incur a cost in regret.



(a) expected demand (b) expected revenue

Figure 2.1: Family of linear demand and revenue curves under $\mathcal{C}_{\mathsf{GenLB}}$ for $z \in \{1/3, 1/2, 2/3, 5/6, 1\}$. For $z = 1/2$, the optimal price is $p^*(1/2) = 1$, which is also the common intersection points for all demand curves in this family.

To give precise statements of Lemmas 2.3.3 and 2.3.4, we will need to quantify the heuristic notion of "uncertainty" about the unknown demand parameter. In our analysis, we will use a convenient quantitative measure of uncertainty, known as the *KL divergence*.

Definition 2.3.2 (Definition 2.26 in Cover and Thomas (1999))**.** For any probability measures Q_0 and Q_1 on a discrete sample space $\mathcal{Y}$, the *KL divergence* of Q_0

and Q_1 is

$$\mathcal{K}\left(Q_0; Q_1\right) = \sum_{y \in \mathcal{Y}} Q_0(y) \log\left(\frac{Q_0(y)}{Q_1(y)}\right).$$

Intuitively, the KL divergence is a measure of distinguishability between two distributions; if the KL divergence between Q_0 and Q_1 is large, then Q_0 and Q_1 are easily distinguishable, and if $\mathcal{K}(Q_0; Q_1)$ is small, then Q_0 and Q_1 are difficult to distinguish. Thus, we say that a pricing policy ψ has a large degree of certainty that the true underlying demand parameter is z_0, rather than some counterfactual parameter z, if the quantity $\mathcal{K}(Q_t^{\psi,z_0}; Q_t^{\psi,z})$ is large.

With this interpretation of the KL divergence, we now state Lemma 2.3.3. This lemma establishes that reducing uncertainty about the underlying parameter is costly, by establishing a lower bound on the regret incurred by an arbitrary pricing policy in terms of the KL divergence.

Lemma 2.3.3 (Learning is Costly)**.** *For any $z \in \mathcal{Z}$, $t \geq 1$, and any policy ψ setting prices in $\mathcal{P}$,*

$$\mathcal{K}\left(Q_t^{\psi,z_0}; Q_t^{\psi,z}\right) \leq \frac{9}{16}\left(z_0 - z\right)^2 \operatorname{Regret}\left(z_0, \mathcal{C}_{\mathsf{GenLB}}, t, \psi\right) \ ,$$

where $z_0 = 1/2$.

The proof of Lemma 2.3.4 is deferred to Appendix A.1.1, but here we give a high level description of the argument. Suppose the underlying demand parameter is z_0, and suppose a pricing policy ψ has the goal of reducing its uncertainty about whether the underlying demand parameter is in fact z_0, as opposed to some other value z. We may restate this goal of "reducing uncertainty" in terms of the KL divergence, by saying that the policy ψ wishes to offer a sequence of prices such that the KL divergence between the induced distributions Q_t^{ψ,z_0} and $Q_t^{\psi,z}$ of customer

responses is large. To accomplish this, ψ must offer prices at which the customer purchase probability will be significantly different under z_0 versus z; however, for all prices in a small neighborhood of the uninformative price $p^*(z_0)$, the probability of a customer purchase is virtually the same under z_0 and z. Thus, to distinguish the two cases (that is, increase the KL divergence $\mathcal{K}(Q_t^{\psi,z_0}; Q_t^{\psi,z})$), the policy ψ must offer prices away from $p^*(z_0)$, and thus incur large regret $\text{Regret}\left(z_0, \mathcal{C}_{\mathsf{GenLB}}, t, \psi\right)$ when the underlying parameter is in fact z_0.

We have now established in Lemma 2.3.3 that reducing uncertainty about the underlying demand curve is costly. However, this result alone is not enough to prove a lower bound on the regret. To establish the desired lower bound on regret, we need a complementary result, showing that any pricing policy that does not decrease its uncertainty about the demand curve must also incur a cost in regret. We establish this complement to Lemma 2.3.3 in the following lemma.

Lemma 2.3.4 (Uncertainty is Costly)**.** *Let ψ be any pricing policy setting prices in $\mathcal{P}$ Then, for any $T \geq 2$ and for demand parameters $z_0 = 1/2$ and $z_1 = z_0 + \frac{1}{4}T^{-1/4}$, we have*

$$\text{Regret}(z_0, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) + \text{Regret}(z_1, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) \geq \frac{\sqrt{T}}{12(48^2)} e^{-\mathcal{K}(Q_T^{\psi,z_0}; Q_T^{\psi,z_1})} \ .$$

The intuition for Lemma 2.3.4 is the following. Let us choose the special parameter $z_0 = 1/2$ such that the corresponding optimal price $p^*(z_0)$ is the uninformative price, and let us choose a second demand parameter $z_1 = z_0 + \frac{1}{4}T^{-1/4}$. The parameters z_0 and z_1 are chosen so that the optimal prices $p^*(z_0)$ and $p^*(z_1)$ are not too close to each other; in other words, z_0 and z_1 are far enough apart (with respect to the time horizon T) such that a near-optimal pricing decision when the demand parameter is z_0 will be sub-optimal when the demand parameter is z_1, and vice versa. Thus, for a pricing policy ψ to price well under both z_0 and z_1,

it must be able to distinguish which of the two is the true demand parameter, based on observed responses to the past prices offered. Consequently, if ψ cannot distinguish between the two cases z_0 and z_1 based on past prices offered (that is, the KL divergence $\mathcal{K}(Q_t^{\psi,z_0}; Q_t^{\psi,z_1})$ is small), then the worst-case regret of ψ must necessarily be large, as seen in the inequality of Lemma 2.3.4.

The proof of Lemma 2.3.4 follows from standard results on the minimum error probability of a two-hypothesis test, and we give a fully detailed proof in Appendix A.1.2. Equipped with Lemmas 2.3.3 and 2.3.4, we can immediately deduce the main result.

Proof of Theorem 2.3.1. Since $\text{Regret}(z_1, \mathcal{C}_{\mathsf{GenLB}}, T, \psi)$ is non-negative, and since $z_1 = z_0 + \frac{1}{4}T^{-1/4}$ by definition, it follows from Lemma 2.3.3 and the choice of z_1 that

$$\text{Regret}(z_0, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) + \text{Regret}(z_1, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) \geq \frac{\sqrt{T}}{9}\mathcal{K}\left(Q_T^{\psi,z_0}; Q_T^{\psi,z_1}\right) .$$

Adding this inequality to the result of Lemma 2.3.4, and using the fact that the KL divergence is non-negative, we have

$$\begin{aligned}
&2\left\{\text{Regret}(z_0, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) + \text{Regret}(z_1, \mathcal{C}_{\mathsf{GenLB}}, T, \psi)\right\} \\
&\geq \frac{\sqrt{T}}{9}\mathcal{K}\left(Q_T^{\psi,z_0}; Q_T^{\psi,z_1}\right) + \frac{\sqrt{T}}{12(48^2)}e^{-\mathcal{K}(Q_T^{\psi,z_0}; Q_T^{\psi,z_1})} \\
&\geq \frac{\sqrt{T}}{12(48^2)} \cdot \left\{\mathcal{K}\left(Q_T^{\psi,z_0}; Q_T^{\psi,z_1}\right) + e^{-\mathcal{K}(Q_T^{\psi,z_0}; Q_T^{\psi,z_1})}\right\} \;\geq\; \frac{\sqrt{T}}{12(48^2)}.
\end{aligned}$$

To see the last inequality, note that $\mathcal{K}\left(Q_T^{\psi,z_0}; Q_T^{\psi,z_1}\right) + e^{-\mathcal{K}(Q_T^{\psi,z_0}; Q_T^{\psi,z_1})} \geq 1$, since $x + e^{-x} \geq 1$ for all $x \in \mathbb{R}_+$. Thus, the tension between pricing optimally and learning the parameters of the demand curve is captured explicitly by the sum $\mathcal{K}\left(Q_T^{\psi,z_0}; Q_T^{\psi,z_1}\right) + e^{-\mathcal{K}(Q_T^{\psi,z_0}; Q_T^{\psi,z_1})}$. The first term in the sum captures the cost of learning the parameters of the demand curve, while the second term in the sum

captures the cost of uncertainty. The fact that this sum cannot be driven to zero, regardless of the choice of the pricing policy, captures the tradeoff between learning and exploiting in the presence of uninformative prices. The desired result follows from the fact that

$$\begin{aligned}\max_{z\in\{z_0,z_1\}} \mathrm{Regret}(z, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) &\geq \frac{\mathrm{Regret}(z_0, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) + \mathrm{Regret}(z_1, \mathcal{C}_{\mathsf{GenLB}}, T, \psi)}{2} \\ &\geq \frac{\sqrt{T}}{48^3}.\end{aligned}$$

□

Remark 2.3.5 (Statistical Identifiability)**.** The result of Theorem 2.3.1 leverages the presence of an "uninformative price" $p^*(z_0)$. Note that the family of distributions $\{Q^{p^*(z_0),z} : z \in \mathcal{Z}\}$ is not identifiable, that is, one cannot uniquely identify the true value of the underlying demand parameter z from observing customer responses *to the single price* $p^*(z_0)$. However, by the arguments of Example 2.2.2, the family $\{Q^{\bar{\mathbf{p}},z} : z \in \mathcal{Z}\}$ *is* identifiable for any $\bar{\mathbf{p}} = (p_1, p_2) \in \mathcal{P}^2$ with $p_1 \neq p_2$, that is, one *can* uniquely identify the value of the underlying parameter from observing customer responses to two distinct prices.

Before we proceed with Section 2.3.2, we briefly remark on the related literature. A very general version of the result of Theorem 2.3.1 was previously known in the computer science literature; Kleinberg and Leighton (2003) contains eight sufficient conditions under which a one-parameter family of demand curves yields regret that is not $o(\sqrt{T})$. It is worth noting that the family constructed in Theorem 2.3.1 does not satisfy the sufficient conditions provided by Kleinberg and Leighton (2003); in particular, the family presented in Theorem 2.3.1 contains an "uninformative price," while their lower bound proof exploits alternative properties.

The techniques used in the proof of Theorem 2.3.1 have appeared in several recent papers. A recent work in dynamic pricing is Besbes and Zeevi (2009), which contains a related lower bound result in a non-stationary demand learning framework. Examples of these techniques in the more general online learning literature can be found in Goldenshluger and Zeevi (2008) and Goldenshluger and Zeevi (2009), which concern optimal learning in a two-armed bandit setting.

2.3.2 A General Matching Upper Bound

In this section, we present a pricing policy called MLE-CYCLE whose regret is $\mathcal{O}(\sqrt{T})$ across all problem instances, matching the order of the lower bound of Section 2.3.1. We describe the policy MLE-CYCLE in detail below, but first we describe the general intuition behind the policy.

Suppose we had access to a good estimate of the underlying demand parameter. Then this would give us a good approximation of the true demand curve, and we would be able to price near-optimally (per the result of Corollary 2.2.4). However, any estimate of the demand parameter will depend on customer responses to the past prices offered, and as seen in Theorem 2.3.1, observing responses to prices near an "uninformative price" will do little to reduce uncertainty about the demand parameter. Thus, to learn the demand curve adequately, a pricing policy should be careful to offer prices at which a good estimate of the demand parameter can be computed.

Motivated by this discussion, we present a policy MLE-CYCLE based on maximum likelihood parameter estimation. The policy MLE-CYCLE operates in cycles, and each cycle consists of an exploration phase followed by an exploitation phase. These cycles are simply a scheduling device, designed to maintain the appropriate

balance between exploration and exploitation. During the exploration phase of a given cycle c, we offer the product to consecutive customers at a sequence of exploration prices $\mathbf{p} \in \mathcal{P}^k$, and then compute a maximum likelihood estimate of the underlying parameter based on the observed customer selections. The exploration prices $\mathbf{p}$ are fixed, and are chosen so that a good estimate of the demand parameter can be computed from the corresponding customer responses. Following the exploration phase of cycle c, there is an exploitation phase of c periods, during which we offer the best-guess optimal price corresponding to the current estimate of the demand parameter to c consecutive customers. Thus, the c^{th} cycle of MLE-CYCLE consists of $(k + c)$ periods: k periods in which we offer each of the k exploration prices, followed by c periods in which we offer the optimal price corresponding to our most recent estimate of the demand parameter. The cycle-based scheduling of MLE-CYCLE is carefully chosen to optimize the balance the amount of demand learning (exploration) with best-guess optimal pricing (exploitation). While we make this balance precise in the analysis of the policy, we note that the scheduling makes intuitive sense: the ratio of exploration steps to exploitation steps in MLE-CYCLE is high in the early time periods, when little is known about the demand curve, and is low in the later time periods, when the demand curve is known to a good approximation.

We now proceed with a formal description of the policy MLE-CYCLE.

We state the regret guarantee of MLE-CYCLE in the following theorem.

Theorem 2.3.6 (General Regret Upper Bound)**.** *For any problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ satisfying Assumptions 1 and 2 with corresponding exploration prices $\bar{\mathbf{p}} \in \mathcal{P}^k$, there exists a constant C_1 depending only on the exploration prices $\bar{\mathbf{p}}$ and the problem class $\mathcal{C}$ such that for all $\mathbf{z} \in \mathcal{Z}$ and $T \geq 2$, the policy* MLE-CYCLE

Policy MLE-CYCLE($\mathcal{C}, \mathbf{p}$)

Inputs: A problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ and exploration prices $\bar{\mathbf{p}} = (\bar{p}_1, \ldots, \bar{p}_k) \in \mathcal{P}^k$.

Description: For each cycle $c = 1, 2, \ldots,$

- Exploration Phase (k periods): Offer the product at exploration prices $\bar{\mathbf{p}} = (\bar{p}_1, \ldots, \bar{p}_k)$ and let $\mathbf{Y}(c) = (Y_1(c), \ldots, Y_k(c))$ denote the corresponding customer selections. Let $\widehat{\mathbf{Z}}(c)$ denote the maximum likelihood estimate (MLE) based on observed customer selections during the exploration phases in the past c cycles, that is,

$$\widehat{\mathbf{Z}}(c) = \arg\max_{\mathbf{z} \in \mathcal{Z}} \prod_{s=1}^{c} Q^{\bar{\mathbf{p}}, \mathbf{z}}(\mathbf{Y}(s)) \ ,$$

  where for each $1 \leq s \leq c$, $\mathbf{Y}(s) = (Y_1(s), \ldots, Y_k(s))$ denotes the observed customer responses to the exploration prices offered in the exploration phase of cycle s.
- Exploitation Phase (c periods): Offer the greedy price $p^*\left(\widehat{\mathbf{Z}}(c)\right)$ based on the estimate $\widehat{\mathbf{Z}}(c)$.

satisfies

$$\text{Regret}(\mathbf{z}, \mathcal{C}, T, \text{MLE-CYCLE}) \leq C_1 \sqrt{T} \ .$$

The main idea of the proof of Theorem 2.3.6 is the following. For a given time horizon T, it is straightforward to check that the number of cycles up to time T is $\mathcal{O}(\sqrt{T})$, and so to prove that the regret of MLE-CYCLE is $\mathcal{O}(\sqrt{T})$, it is enough to show that the regret in each cycle is $\mathcal{O}(1)$. Since each cycle consists of an exploration phase followed by an exploitation phase, it's enough to show that for an arbitrary cycle c, the regret incurred in the exploration phase is $\mathcal{O}(1)$, and the regret incurred during the exploitation phase is $\mathcal{O}(1)$.

First, to show that the regret during the exploration phase of an arbitrary cycle is $\mathcal{O}(1)$, note that during the exploration phase, MLE-CYCLE offers k exploration prices, and the regret incurred from offering each of these exploration prices is

$\mathcal{O}(1)$, by the smoothness of the revenue function, and the compactness of the pricing interval. Thus, the total regret incurred during the exploration phase is $\mathcal{O}(1)$. Secondly, to show that the regret incurred during the exploitation phase of an arbitrary cycle is $\mathcal{O}(1)$, recall that the price offered during the exploitation phase of cycle c is $p^*(\widehat{\mathbf{Z}}(c))$. This price is offered to c customers, and by Corollary 2.2.4, the instantaneous regret incurred for each customer is $O\left(\mathbb{E}_z\left[||\mathbf{z} - \widehat{\mathbf{Z}}(c)||^2\right]\right)$. But since $\widehat{\mathbf{Z}}(c)$ is a MLE computed from c samples, it follows from a standard result that $\mathbb{E}_z\left[||\mathbf{z} - \widehat{\mathbf{Z}}(c)||^2\right] = \mathcal{O}(1/c)$. Since this prices is offered to c customers, the total regret incurred during the exploitation phase is $c \cdot \mathcal{O}(1/c) = \mathcal{O}(1)$, as claimed.

We now proceed with a rigorous proof based on the above intuition. We begin by stating a bound on the mean squared error of the maximum likelihood estimator formed by MLE-CYCLE.

Lemma 2.3.7 (Mean Squared Errors for MLE based on IID Samples, Borovkov (1998))**.** *For any $c \geq 1$, let $\widehat{\mathbf{Z}}(c)$ denote the maximum likelihood estimate formed by the* MLE-GREEDY *policy after c exploration cycles. Then there exists a constant C_{mle} depending only on the exploration prices $\mathbf{p}$ and the problem class $\mathcal{C}$ such that*

$$\mathbb{E}_{\mathbf{z}}\left[\left\|\widehat{\mathbf{Z}}(c) - \mathbf{z}\right\|^2\right] \leq \frac{C_{mle}}{c} \ .$$

The proof of Lemma 2.3.7 follows from standard results on the mean-squared error of maximum-likelihood estimators, and is given in detail in Appendix A.2. We now give the proof of Theorem 2.3.6.

Proof. Fix a problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ with corresponding exploration prices $\bar{\mathbf{p}}$, and consider an arbitrary cycle c. First, we show that the regret incurred during the exploration phase of cycle c is $\mathcal{O}(1)$. Since the revenue function is smooth by assumption, and since the pricing interval $\mathcal{P}$ is compact, it follows that there exists

a constant $\bar{D}_1$ depending only on the problem class $\mathcal{C}$ such that

$$r(p^*(\mathbf{z}); \mathbf{z}) - r(p; \mathbf{z}) \leq \bar{D}_1$$

for all $\mathbf{z} \in \mathcal{Z}$ and all $p \in \mathcal{P}$. Consequently, the regret during the exploration phase of cycle c satisfies

$$\sum_{\ell=1}^{k} \mathbb{E}_{\mathbf{z}}[r(p^*(\mathbf{z}); \mathbf{z}) - r(\bar{p}_\ell; \mathbf{z})] \quad \leq \quad k\bar{D}_1.$$

Next, we show that the regret incurred during the exploitation phase of cycle c is also $\mathcal{O}(1)$. During the exploitation phase of cycle c, we use the greed price $p^*\left(\widehat{\mathbf{Z}}(c)\right)$, and we offer this price for c periods. It follows from Corollary 2.2.4 and Lemma 2.3.7 that the instantaneous regret during the exploitation phase satisfies

$$\mathbb{E}_{\mathbf{z}}\left[r(p^*(\mathbf{z}); \mathbf{z}) - r(p^*(\widehat{\mathbf{Z}}(c)); \mathbf{z})\right] \leq c_r \; L^2 \; \mathbb{E}_{\mathbf{z}}\left[\left\|\mathbf{z} - \widehat{\mathbf{Z}}(c)\right\|^2\right] \quad \leq \quad \frac{c_r L^2 C_{mle}}{c},$$

and since the price $p^*(\widehat{\mathbf{Z}}(c))$ is offered for c periods during the exploitation phase of cycle c, we have that the total regret incurred during the exploitation phase of cycle c is bounded above by $c_r L^2 C_{mle}$. Putting everything together, we have that the cumulative regret over K cycles (corresponding to $2K + \sum_{c=1}^{K} c$ periods) satisfies

$$\text{Regret}(\mathbf{z}, \mathcal{C}, 2K + \sum_{c=1}^{K} c, \textsf{MLE-CYCLE}) \leq \left(k\bar{D}_1 + c_r L^2 C_{mle}\right) K \; .$$

Now, consider an arbitrary time period $T \geq 2$ and let $K_0 = \lceil \sqrt{2T} \rceil$. Note that the total number of time periods after K_0 cycles is at least T because $2K_0 + \sum_{c=1}^{K_0} c \geq \sum_{c=1}^{K_0} c = K_0(K_0+1)/2 \geq T$. The desired result follows from the fact that

$$\text{Regret}(\mathbf{z}, \mathcal{C}, T, \textsf{MLE-CYCLE}) \leq \text{Regret}(\mathbf{z}, \mathcal{C}, 2K_0 + \sum_{c=1}^{K_0} c, \textsf{MLE-CYCLE}).$$

□

2.4. The Well-Separated Case

In the general case studied in Section 2.3.1, there are two major obstacles to pricing that force any policy to have $\Omega(\sqrt{T})$ worst-case regret. The first obstacle is the stochastic nature of the demand. A pricing policy never observes a noise-free value of the demand curve at a given price; it observes only a random variable whose expected value is the demand at that price. The second and more prominent obstacle is that of "uninformative prices," at which no pricing policy can reduce its uncertainty about demand.

Given this observation, a natural question is the following: how much does each of the two obstacles contribute to the difficulty of dynamic pricing? More specifically, are uninformative prices so difficult to deal with that they force a minimum regret of $\Omega(\sqrt{T})$, or is it simply the stochastic nature of the demand that forces this lower bound? In this section, we shed light on this issue by considering demand curves that satisfy a "well-separated" condition (Assumption 3), which precludes the possibility of uninformative prices. Under this assumption, we show in Section 2.4.1 a lower bound of $\Omega(\log T)$ on the T-period cumulative regret under an arbitrary policy. Then, in Section 2.4.2, we show that a greedy policy achieves regret matching the order of the lower bound.

We now state Assumption 3, which guarantees that it is possible to estimate demand from customer responses at *any* price in $\mathcal{P}$.

Assumption 3 (Well Separated Assumption)**.** The problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ has a parameter set $\mathcal{Z} \subset \mathbb{R}$, and for <u>*all*</u> prices $p \in \mathcal{P}$,

(a) The family of distributions $\{Q^{p,z} : z \in \mathcal{Z}\}$ is identifiable.

(b) There exists a constant $c_f > 0$ depending only on the problem class $\mathcal{C}$ such that the *Fisher information* $I(p, z)$, given by

$$I(p, z) = \mathbb{E}_z \left[-\frac{\partial^2}{\partial z^2} \log Q^{p,z}(Y) \right]$$

satisfies $I(p, z) \geq c_f$ for all $z \in \mathcal{Z}$.

Remark 2.4.1 (Geometric Interpretation of Assumption 3)**.** To make the notion of well separated more concrete, one may show that any problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ satisfying Assumption 3 also has the following property: there exists some constant $c_d > 0$ depending only on $\mathcal{C}$ such that

$$|d(p; z) - d(p; \hat{z})| \geq c_d \, |z - \hat{z}|$$

for any price $p \in \mathcal{P}$, and any $z, \hat{z} \in \mathcal{Z} \subset \mathbb{R}$. We defer the details of this derivation to Appendix A.5.1. Thus, for any fixed price $p \in \mathcal{P}$, if we vary the demand parameter z to some other value $\hat{z}$, then the demand at price p will vary by an amount proportional to $|z - \hat{z}|$. An obvious consequence of this property and the smoothness of the demand curves is that for any two demand parameters $z \neq \hat{z}$, it must be the case that either $d(p; z) > d(p; \hat{z})$ for all $p \in \mathcal{P}$, or $d(p; z) < d(p; \hat{z})$ for all $p \in \mathcal{P}$. Thus, we refer to this condition as a "well-separated" condition, since it implies that for any two demand parameters $z \neq \hat{z}$, the corresponding demand curves do not intersect with each other.

Since we will use the maximum likelihood estimator in our pricing model and this estimator is the minimizer of the function $z \mapsto -\log \, Q_t^{\mathbf{P},z}(\mathbf{Y}_t)$, we now state Assumption 4, which gives a convenient property of the likelihood function that allows for a simple analysis of the likelihood process. As shown in Examples 2.4.2, 2.4.3, and 2.4.4, Assumptions 3 and 4 are satisfied by many demand families of interest, including the linear, logistic, and exponential.

Assumption 4 (Likelihood Assumptions)**.** For any sequence of prices $\mathbf{p} = (p_1, \ldots, p_t) \in \mathcal{P}^t$, the function

$$z \mapsto -\log Q_t^{\mathbf{p},z}(\mathbf{Y}_t)$$

is convex on $\mathcal{Z} \subset \mathbb{R}$.

We now state some examples of problem classes satisfying Assumptions 3 and 4.

Example 2.4.2 (One-Parameter Logit Family)**.** Let $\mathcal{P} = [1/2, 2]$ and let $\mathcal{Z} = [1, 2]$. Define a family of logistic demand curves by

$$d(p, z) = \frac{e^{-zp}}{1 + e^{-zp}}.$$

Then by Example 2.2.1, we know that this problem instances satisfies the conditions of Assumption 1. It is also straightforward to check that for any $\bar{p} \in \mathcal{P}$, the associated family $\{Q^{\bar{p},z} : z \in \mathcal{Z}\}$ is identifiable. Moreover, for any $\bar{p} \in \mathcal{P}$ and $z \in \mathcal{Z}$, we have that

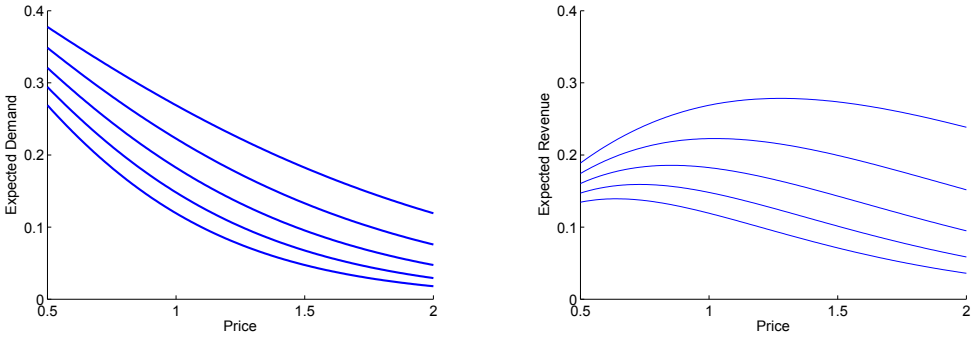
$$\frac{d}{dz} d(\bar{p}; z) = -\bar{p}\, d(\bar{p}; z)(1 - d(\bar{p}; z)) \ ,$$

and so by the formula given in Assumption 2, we have that the Fisher information is given by

$$I(\bar{p}, z) = \bar{p}^2\, d(\bar{p}; z)(1 - d(\bar{p}; z)) \geq p_{min}^2\, d_{min}(1 - d_{max}) \ .$$

Finally, it is a standard result (see, for example, Ben-Akiva and Lerman (1985)) that for the logit model, the negative log-likelihood function is globally convex, and so Assumption 4 is satisfied.

Example 2.4.3 (One-Parameter Linear Family)**.** Let $\mathcal{P} = [1/3, 1/2]$, let $\mathcal{Z} = [3/4, 1]$, and let $b = 2/3$ be a fixed constant. Define a linear family of demand



(a) expected demand (b) expected revenue

Figure 2.2: Family of well separated logit demand and revenue curves from Example 2.4.2 for $z \in \{1, 5/4, 6/4, 7/4, 2\}$.

curves by $d(p; z) = b - zp$. By Example 2.2.2, we know that this problem instances satisfies the conditions of Assumption 1. It is also straightforward to check that for any $\bar{p} \in \mathcal{P}$, the associated family $\{Q^{\bar{p},z} : z \in \mathcal{Z}\}$ is identifiable. Moreover, for any $\bar{p} \in \mathcal{P}$ and $z \in \mathcal{Z}$, we have that

$$\frac{d}{dz} d(\bar{p}; z) = -\bar{p} \ ,$$

and we have that the Fisher information is given by

$$I(\bar{p}, z) = \frac{\bar{p}^2}{d(\bar{p}; z)(1 - d(\bar{p}; z))} \geq \frac{p_{min}^2}{d_{max}(1 - d_{min})} \ .$$

Finally, to verify Assumption 4, we have that for any vector of prices $\mathbf{p} = (p_1, \ldots, p_t) \in \mathcal{P}^t$,

$$Q_t^{\mathbf{p},\mathbf{z}}(\mathbf{y}_t) = \prod_{\ell=1}^{t} (b - zp_\ell)^{y_\ell} (1 - b + zp_\ell)^{1-y_\ell} \ ,$$

so that

$$-\log Q_t^{\mathbf{p},\mathbf{z}}(\mathbf{y}_t) = -\sum_{\ell=1}^{t} \{y_\ell \log(b - zp_\ell) + (1 - y_\ell) \log(1 - b + zp_\ell)\} \ .$$

Taking derivatives twice, we have

$$\frac{d^2}{dz^2} \{-\log Q_t^{\mathbf{p},\mathbf{z}}(\mathbf{y}_t)\} \ = \ \sum_{\ell=1}^{t} \frac{y_\ell p_\ell^2}{(b - zp_\ell)^2} + \frac{(1 - y_\ell) p_\ell^2}{(1 - b + zp_\ell)^2} \ > \ 0 \ ,$$

which implies that the negative log-likelihood function is globally convex, as desired.

Example 2.4.4 (One-Parameter Exponential Family)**.** Let $\mathcal{P} = [1/2, 1]$ and let $\mathcal{Z} = [1, 2]$. Define an exponential family of demand curves by

$$d(p; z) = e^{-zp}.$$

By the same techniques used in Examples 2.4.2 and 2.4.3, one can check that this problem class satisfies all the conditions of Assumptions 1 and 3. Moreover, to verify Assumption 4, one can check that for any vector of prices $\mathbf{p} = (p_1, \ldots, p_t) \in \mathcal{P}^t$,

$$\frac{d^2}{dz^2} \left\{ -\log Q_t^{\mathbf{p},\mathbf{z}}(\mathbf{y}_t) \right\} \;=\; \sum_{\ell=1}^{t} \frac{(1 - y_\ell) p_\ell^2 e^{-z_\ell p}}{(1 - e^{-z_\ell p})^2} \;>\; 0 \;,$$

which implies that the negative log-likelihood function is globally convex, as desired.

2.4.1 A Lower Bound

In this section we establish a lower bound of $\Omega(\log T)$ for the well-separated case. The main result of this section is stated in the following theorem.

Theorem 2.4.5 (Well-Separated Lower Bound)**.** *Define a problem class* $\mathcal{C}_{\mathsf{WellSepLB}} = (\mathcal{P}, \mathcal{Z}, d)$ *by letting* $\mathcal{P} = [1/3, 1/2]$, $\mathcal{Z} = [2, 3]$, *and letting* $d(p; z) = 1 - (pz)/2$. *Then for any policy* ψ *setting prices in* $\mathcal{P}$ *and any* $T \geq 1$, *there exists a constant* $z \in \mathcal{Z}$ *such that*

$$\mathrm{Regret}(z, \mathcal{C}_{\mathsf{WellSepLB}}, T, \psi) \geq \frac{1}{405\pi^2} \log T \;.$$

There are two key observations that lead to the proof of Theorem 2.4.5. First, recall that in our model, the price offered by a pricing policy ψ to the t^{th} customer

is given by $P_t = \psi_t(\mathbf{Y}_{t-1})$, where $\psi_t : \{0,1\}^{t-1} \to \mathcal{P}$ is any function and $\mathbf{Y}_{t-1}$ is a vector of observed customer responses. Thus, we may think of P_t as an "estimator," since P_t is just a function ψ_t of the observed data $\mathbf{Y}_{t-1}$. Consequently, we may apply standard results on the minimum mean squared error of an estimator to show that $\mathbb{E}[(p^*(Z) - P_t)^2] = \Omega(1/t)$. We make this precise in Lemma 2.4.6 whose proof is given in Appendix A.3.

Secondly, as a converse to Corollary 2.2.4, we will see that it is easy to construct problem classes under which the instantaneous regret in time t is bounded *below* by the mean squared error of the price P_t with respect to the optimal price $p^*(z)$ (times some constant factors). Combining this result with the above estimate on the minimum mean squared error of P_t established the theorem.

Our proof technique follows that of Goldenshluger and Zeevi (2009), who have used van Trees' inequality to prove lower bounds on the performance of sequential decision policies.

Lemma 2.4.6 (Instantaneous Risk Lower Bound)**.** *Let $\mathcal{C}_{\mathsf{WellSepLB}} = (\mathcal{P}, \mathcal{Z}, d)$ be the problem class defined in Theorem 2.4.5, and let Z be a random variable taking values in $\mathcal{Z}$, with density $\lambda : \mathcal{Z} \to \mathbb{R}_+$ given by $\lambda(z) = 2\{\cos(\pi(z - 5/2))\}^2$. Then for any pricing policy ψ setting prices in $\mathcal{P}$, and for any $t \geq 1$,*

$$\mathbb{E}\left[(p^*(Z) - P_{t+1})^2\right] \geq \frac{1}{405\pi^2} \cdot \frac{1}{t},$$

where P_{t+1} is the price offered by ψ at time $t+1$, and $\mathbb{E}[\,\cdot\,]$ denotes the expectation with respect to the joint distribution of P_t and the prior density λ of the parameter $Z \in \mathcal{Z} = [2,3]$.

Here is the proof of Theorem 2.4.5.

Proof of Theorem 2.4.5. By checking first and second order optimality conditions, it is straightforward to check that $p^*(z) = 1/z$. By noting that $r'(p^*(z); z) = 0$ and $r''(p; z) = -z \leq -2$, it follows from a standard result that for any $z \in \mathcal{Z}$ and $p \in \mathcal{P}$,

$$r(p^*(z); z) - r(p; z) \geq (p^*(z) - p)^2 \ .$$

Applying this fact and Lemma 2.4.6, we have

$$\begin{aligned}
\sup_{z \in \mathcal{Z}} \text{Regret}(z, \mathcal{C}_{\mathsf{WellSepLB}}, T, \psi) &\geq \sup_{z \in \mathcal{Z}} \mathbb{E}_z \left[\sum_{t=1}^{T-1} [r(p^*(Z); Z) - r(P_{t+1}; Z)] \right] \\
&\geq \mathbb{E} \left[\sum_{t=1}^{T-1} r(p^*(Z); Z) - r(P_{t+1}; Z) \right] \\
&\geq \frac{1}{405\pi^2} \sum_{t=1}^{T-1} \frac{1}{t} \\
&\geq \frac{1}{405\pi^2} \log T
\end{aligned}$$

where the last line follows from the fact that $\sum_{t=1}^{T-1} \frac{1}{t} \geq \int_1^T \frac{dx}{x} = \log T$. □

2.4.2 A Matching Upper Bound for Well Separated Problem Class

In this section, we present a simple greedy pricing strategy called MLE-GREEDY whose regret is $\mathcal{O}(\log T)$ across all well separated problem instances, matching the order of the lower bound established in Section 2.4.1. We describe MLE-GREEDY in detail below, but here we sketch the intuition behind the policy.

Intuitively, we know that if we form a good estimate of the underlying demand parameter, then the optimal price corresponding to this estimate will be close to the true optimal price. More specifically, Corollary 2.2.4 establishes that if we compute an estimator whose mean squared error is $\mathcal{O}(1/t)$ in each time period t, then by offering the optimal prices corresponding to these estimates, we will incur instantaneous regret $\mathcal{O}(1/t)$ in each time period t, and thus incur regret that is

$\mathcal{O}(\log T)$ up to time T. Thus, a natural approach is to compute an estimate of the demand parameter based on the observed customer responses to past prices offered, and then offer the best-guess optimal price corresponding to this estimate.

Although this intuition is essentially correct, there is a wrinkle to the analysis. Suppose that in time periods $1, \ldots, t$, we could observe the *actual* willingness-to-pay of each customer; that is, if we could observe the realized values $(v_1, \ldots, v_t)$ of the i.i.d. willingness-to-pay random variables $(V_1, \ldots, V_t)$. Then by standard results on maximum likelihood estimation (e.g. Theorem A.2.1), we could compute an estimator whose mean squared error was $\mathcal{O}(1/t)$, and by Corollary 2.2.4, incur regret $\mathcal{O}(1/t)$ by offering the optimal price corresponding to our estimator. However, in our model, a pricing policy does not have access to the actual willingness-to-pay of each customer. Rather, the policy observes a Bernoulli random variable $Y_t = \mathbf{1}\{V_t \geq P_t\}$ specifying whether the willingness-to-pay V_t of customer t exceeded the price offered P_t. Consequently, the observations $Y_1, Y_2, \ldots, Y_t$ are dependent random variables, because for any ℓ, Y_ℓ is a function of the price P_ℓ in period ℓ, which depends on the customer responses $Y_1, \ldots, Y_{\ell-1}$ in the preceding $\ell - 1$ periods. Thus, a pricing policy must form an estimate based on samples that are *dependent and not identically distributed*, and the standard bound for MLE estimates (Theorem A.2.1) does not apply. Thus, to establish an upper bound on the regret of MLE-GREEDY using the approach described above, it is enough to establish that the mean squared error of the estimate formed by MLE-GREEDY from t samples is in fact $\mathcal{O}(1/t)$.

With this intuition, we proceed with our analysis of the greedy pricing policy. For brevity in the following analysis, we denote by $\mathcal{G} = (\mathcal{G}_1, \mathcal{G}_2, \ldots)$ the pricing policy MLE-GREEDY described below.

Policy MLE-GREEDY$(\mathcal{C}, p_1)$

Inputs: A problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$, and an initial price $p_1 \in \mathcal{P}$.
Initialization: At time $t = 1$, offer the initial price p_1, and observe the corresponding customer decision $Y_1 = \mathbf{1}\{V_1 \geq p_1\}$.
Description: For time $t = 2, 3, \ldots$,

- Compute the maximum likelihood estimate $\widehat{Z}(t-1)$ given by

$$\widehat{Z}(t-1) = \arg\max_{z \in \mathcal{Z}} Q_{t-1}^{\mathcal{G},z}(\mathbf{Y}_{t-1}) \ ,$$

  where $\mathbf{Y}_{t-1} = (Y_1, \ldots, Y_{t-1})$ denotes the observed customer responses in the first $t-1$ periods.
- Offer the greedy price $p^*\left(\widehat{Z}(t-1)\right)$ based on the estimate $\widehat{Z}(t-1)$.

We now state the main result on the mean squared error on the maximum-likelihood estimator computed by MLE-GREEDY, which we prove in Appendix A.4.

Theorem 2.4.7 (MLE Deviation Inequality for Dependent Samples)**.** *Let* $\widehat{Z}(t) = \arg\max_{z \in \mathcal{Z}} Q_t^{\mathcal{G},z}(\mathbf{Y}_t)$ *be the maximum-likelihood estimate formed by the* MLE-GREEDY *policy. Then for any* $t \geq 1$, $z \in \mathcal{Z}$, *and* $\epsilon \geq 0$,

$$\Pr_z\{|\widehat{Z}(t) - z| > \epsilon\} \leq 2\, e^{-t c_H \epsilon^2/2} \qquad \textit{and} \qquad \mathbb{E}_z[(\widehat{Z}(t) - z)^2] \leq \frac{4}{c_H} \cdot \frac{1}{t}$$

The above theorem immediately yields the upper bound the regret, which is the main result of this section.

Theorem 2.4.8 (Well-separated Regret Upper Bound)**.** *For any problem class* $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ *satisfying Assumptions 1, 3, and 4, and any initial price* $p_1 \in \mathcal{P}$, *there exists a constant* C_2 *depending only* $\mathcal{C}$ *and* p_1 *such that for all* $z \in \mathcal{Z}$ *and* $T \geq 2$, *the* MLE-GREEDY *policy satisfies*

$$\mathrm{Regret}(z, \mathcal{C}, T, \text{MLE-GREEDY}) \leq C_2 \cdot \log T \ .$$

Proof. To bound the regret incurred by MLE-GREEDY in the first period, note

that since the revenue function is a smooth on the compact set $\mathcal{P} \times \mathcal{Z}$, there exists a constant $\bar{D}_2$ depending only on $\mathcal{C}$ such that $r(p^*(z); z) - r(p_1; z) \leq \bar{D}_2$ for any choice of p_1 and any $z \in \mathcal{Z}$.

To bound the regret in the subsequent periods, we apply Corollary 2.2.4 and Theorem 2.4.7 to see that

$$\begin{aligned} \mathbb{E}_z \left[\sum_{t=1}^{T-1} r(p^*(z); z) - r(p^*(\widehat{Z}(t)); z) \right] &\leq c_r L^2 \sum_{t=1}^{T-1} \mathbb{E}_z \left[(\widehat{Z}(t) - z)^2 \right] \\ &\leq \frac{4 c_r L^2}{c_H} \sum_{t=1}^{T-1} \frac{1}{t}. \end{aligned}$$

Taking $C_2 = \bar{D}_2 + 4 c_r L^2 c_H / (4 \log 2)$ proves the claim. □

2.5. Numerical Experiments

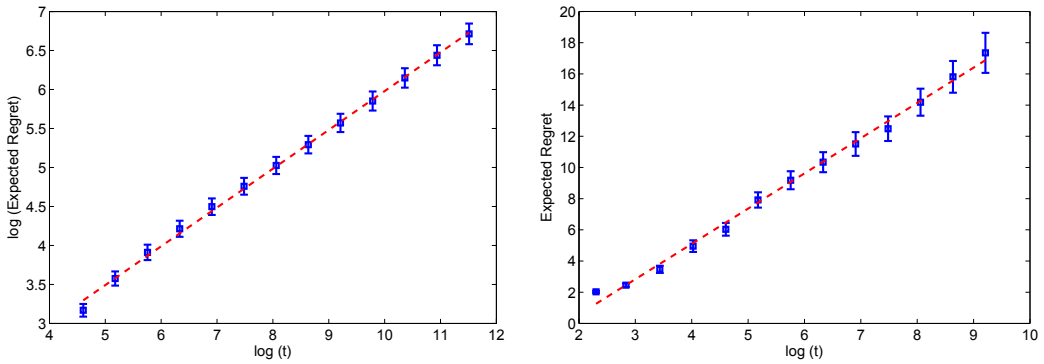
In this section, we evaluate the empirical performance of the MLE-CYCLE and MLE-GREEDY policies described in Sections 2.3.2 and 2.4.2. We investigate their rates of regret, and compare their performance to the performance of several alternative policies, over a variety of problem instances. For all of our simulations, we focus on a logistic demand problem class given by $\mathcal{P} = [1/2, 8]$, $\mathcal{Z} = [0.2, 2] \times [-1, 1]$ and

$$d(p; \mathbf{z}) = \frac{e^{-z_1 p - z_2}}{1 + e^{-z_1 p - z_2}}.$$

2.5.1 First Simulation: Rates of Regret

For our first simulation, we investigate the rates of regret of MLE-CYCLE and MLE-GREEDY on a specific problem instance from the problem class described above. We compute the average regret of both policies over 50 independent trials for parameter values $z_1 = 1$ and $z_2 = -1$, normalizing the regret by the maximum

possible per-period revenue for this instance. For MLE-CYCLE, we fix the exploration prices to be $\bar{p}_1 = 1/2$ and $\bar{p}_2 = 4.25$, corresponding to the left endpoint and midpoint of the pricing interval, and we fix the time horizon to be $T = 10^5$. For MLE-GREEDY, we fix the initial price to be $\bar{p}_1 = 4.25$, and we fix the time horizon to be $T = 10^4$.



(a) Average Regret of MLE-CYCLE (b) Average Regret of MLE-GREEDY

Figure 2.3: An illustration of the rates of regret of MLE-CYCLE and MLE-GREEDY. In Figure 2.3 (a), the line of best fit in the log-log plot of expected regret versus T has slope 0.49, indicating that the rate of regret of MLE-CYCLE is approximately $\Theta(\sqrt{T})$. In Figure 2.3 (b), the expected regret of MLE-GREEDY versus $\log(T)$ is approximately linear, indicating that the rate of regret is $\Theta(\log T)$.

In Figure 2.3 (a), we plot the logarithm of the average regret of MLE-CYCLE versus $\log(t)$. We note that the line of best fit to the mean regret values has a slope of 0.49, which is consistent with the $\Theta(\sqrt{T})$ rate of regret established in Section 2.3. In Figure 2.3 (b), we plot the average regret of MLE-GREEDY versus $\log(t)$. The linear trend of the mean regret values is consistent with the $\Theta(\log T)$ rate of regret established in Section 2.4. These results provide a simple empirical example of the rates of regret of the two policies.

2.5.2 Second Simulation: The General Case

For our second simulation, we compare the performance of MLE-CYCLE with several alternative heuristics. We describe these alternative heuristics below.

1. FP: As a baseline for comparison, we consider a fixed-price policy FP that chooses a price uniformly at random from the pricing interval, and offers this price for all time periods. Note that this policy will have regret that is linear in T.
2. MLE-CYCLE-S: The MLE-CYCLE-S policy is a variant of MLE-CYCLE that uses samples from both the exploration *and* exploitation phases to compute its estimates of the unknown parameters (recall that the MLE-CYCLE policy computes estimates only from its explorations periods).
3. MLE-CYCLE-SU: The MLE-CYCLE-SU policy is a further refinement of MLE-CYCLE, in which all samples are used for computing the estimates, and in addition, the exploration prices are updated at each step to be close to the estimated optimal price. Specifically, at the beginning of each cycle, we choose the first exploration price P_1 to be equal to the current estimated optimal price, and we set the second exploration price P_2 to be $P_1 + t^{-1/4}$, where t is the current time period. This scheme balances the competing objectives of having the exploration prices close to the optimal price, and having them far enough apart to provide good estimates of the demand parameters. We note that this scheme is closely related to the Controlled Variance Pricing idea introduced in den Boer and Zwart (2010a).
4. KW: To compare our policies with general stochastic optimization techniques, we will consider a Kiefer-Wolfowitz-type stochastic optimization pol-

icy. Given a current price P_t, the KW policy sets

$$P_{t+1} = P_t + c_n \qquad P_{t+2} = P_t - c_n \qquad P_{t+3} = P_t + a_t \frac{Y_{t+1}P_{t+1} - Y_{t+2}P_{t+2}}{2c_t},$$

where $Y_{t+1} = \mathbf{1}\{V_{t+1} \geq P_{t+1}\}$ and $Y_{t+2} = \mathbf{1}\{V_{t+2} \geq P_{t+2}\}$. This is a stochastic gradient-ascent optimization scheme, and we implement this scheme with $a_t = t^{-1}$ and $c_t = t^{-1/4}$.

Recall that in Section 2.3.2, we were concerned with describing a pricing policy whose regret matched the order of the $\Omega(\sqrt{T})$ lower bound established in Section 2.3.1. The MLE-CYCLE policy proposed in Section 2.3.2 was sufficient to achieve this goal, and its simple structure facilitated a straightforward analysis of its regret, which was desirable for the theoretical development of Section 2.3. However, although MLE-CYCLE achieves the optimal $\mathcal{O}(\sqrt{T})$ regret, there are a number of natural modifications of this policy that one might suspect would improve its performance – specifically, the use of *all* samples to compute estimates of the demand parameters, and the updating of exploration prices as information is gained. We empirically investigate both of these modifications is this section by studying the performance of MLE-CYCLE-S and MLE-CYCLE-SU.

We investigate the performance of all pricing policies on an ensemble of problem instances drawn from a Gaussian distribution over the parameter set. We generate 500 independent random samples $(\mathbf{z}^1, \ldots, \mathbf{z}^{500})$, by drawing independent random values z_1^i in the interval $[0.2, 2]$ according to a Gaussian distribution with mean $(2+0.2)/2$ and variance $(2-0.2)/4$, truncating so that all samples lie in the interval. We then generate 500 independent random samples z_2^i for the interval $[-1, 1]$ in a similar fashion, and set $\mathbf{z}^i = (z_1^i, z_2^i)$.

To evaluate the performance of each policy, we consider the **Percentage Revenue Loss**, which is defined to be the average over the random sample of problem instances of the cumulative regret divided by the total optimal revenue. Thus, if $\mathbf{z}^1, \ldots, \mathbf{z}^m \in \mathcal{Z}$ is the sample of problem parameters, we have

$$\textbf{Percentage Revenue Loss}\ (T) \ = \ \frac{1}{m} \sum_{i=1}^{m} \frac{\sum_{s=1}^{T} r(p^*(\mathbf{z}^i); \mathbf{z}^i) - r(P_s^i; \mathbf{z}^i)}{T \ \cdot \ r(p^*(\mathbf{z}^i); \mathbf{z}^i)} \times 100\% \ .$$

Equivalently, this quantity describes the total amount of revenue lost by each policy with respect to the optimal policy, as a percentage of the total optimal revenue.

In Table 2.1, we report the results of these experiments. For all simulations, the exploration prices for MLE-CYCLE and its variants, and the initial price for the KW policy, are chosen uniformly at random from the pricing interval. The standard error of the figures reported in the **Percentage Revenue Loss** columns is less than 0.2% for MLE-CYCLE, MLE-CYCLE-S, and MLE-CYCLE-SU, and is less than 1.8% for FP and KW, at all reported values of T.

Table 2.1: Comparison of the Percentage Revenue Loss of the heuristics on the Gaussian instance.

	Percentage Revenue Loss				
$T \times 10^3$	FP	KW	MLE-CYCLE	MLE-CYCLE-S	MLE-CYCLE-SU
1	61.9 %	58.7 %	20.4 %	14.3 %	6.0 %
2	61.9 %	58.0 %	16.1 %	10.7 %	5.0 %
3	61.9 %	57.6 %	13.9 %	9.0 %	4.5 %
4	61.9 %	57.3 %	12.5 %	7.8 %	4.2 %
5	61.9 %	57.1 %	11.5 %	7.1 %	4.0 %

First, we note that all policies lose a smaller percentage of the optimal revenues than the FP policy, and more importantly, all policies have a percentage revenue loss that is decreasing with the number of time steps. We note that all three variants of MLE-CYCLE lose a significantly smaller proportion of the optimal revenue than the FP and KW policies; moreover, we see that both the use of all samples to compute the estimates of the demand parameters, as well as the updating of the exploration prices, lead to a significant improvement in the percentage revenue lost.

2.5.3 Third Simulation: The Well-Separated Case

As a final simulation, we will investigate the percentage revenue loss of MLE-GREEDY when problem parameters are drawn from two different distributions. Recall that to prove the lower bound of Section 2.4.1 on the performance of MLE-GREEDY, we showed that expected regret was $\Omega(\log T)$, when the problem parameters were drawn from a specially chosen distribution. A natural question is whether this distribution is somehow pathological, or whether the expected regret of MLE-GREEDY would be similar when problem parameters are drawn from some other type of distribution. To address this question, we generate three sets of 100 independent random problem instances for the logistic demand problem class described at the beginning of this section. Each set is generated by fixing $z_2 = 0$, and drawing z_1 from one of two distributions over the interval $[0.2, 2]$ (note that the value of $z_2 = 0$ is known to MLE-GREEDY). The first is the distribution $\frac{10}{9}\left\{\cos\left(\frac{5\pi}{9}\left(x - \frac{11}{10}\right)\right)\right\}^2$, similar to the one used in the proof of the lower bound of Section 2.4.1, and the second is the uniform distribution on $[0.2, 2]$. For all simulations, the starting price of MLE-GREEDY is chosen uniformly at random from the pricing interval. In Table 2.2, we report the percentage revenue loss of

MLE-GREEDY for each of the three experiments.

Table 2.2: Comparison of the Percentage Revenue Loss of MLE-GREEDY on two distributions

	Percentage Revenue Loss			
	Lower Bound		Uniform	
$T \times 10^3$	FP	MLE-GREEDY	FP	MLE-GREEDY
1	65.2 %	1.20 %	62.3 %	1.10 %
2	65.2 %	0.67 %	62.3 %	0.61 %
3	65.2 %	0.48 %	62.3 %	0.43 %
4	65.2 %	0.37 %	62.3 %	0.34 %
5	65.2 %	0.30 %	62.3 %	0.28 %

The standard error for all percentage revenue loss figures reported in Table 2.2 is less that 0.07%. We note that the percentage revenue loss of MLE-GREEDY is much smaller than that of the fixed price policy, as well as all of the policies tested in the simulation for the general case. Moreover, we note that when averaged over 100 trials, the percentage revenue loss of MLE-GREEDY is practically identical for both problem distributions. This suggests that the lower bound distribution used in Section 2.4.1 is not pathological, and that we should expected similar average-case behavior for MLE-GREEDY when instances are drawn from other natural distributions.

2.6. Discussion

We studied a stylized dynamic pricing problem under a general parametric choice model. For the general case, we constructed a forced-exploration policy based on maximum likelihood estimation that achieved the optimal $\mathcal{O}(\sqrt{T})$ order of regret. We also considered the special case of a "well-separated" demand family, for which a myopic maximum likelihood policy achieved the optimal $\mathcal{O}(\log T)$ order of regret. Finally, we performed an empirical investigation of the rate of regret of our policies, and compared the performance of several variations thereof. There are many possible extensions of this work, including extensions to account for the sale of multiple products and for competition among sellers. Other interesting directions would involve a more complex model of customer behavior, accounting for strategic customer decision making, or a model in which the parameter values varied over time.

Chapter 3

Pricing with Minimal Adjustments

3.1. Introduction

In the previous chapter, we saw the existence of effective strategies for pricing under demand uncertainty, which allowed a seller to generate near-optimal revenues with limited prior knowledge of the demand curve. To facilitate revenue generation in this scenario, we allowed our pricing strategies to adjust prices freely over time. This lack of restriction on price adjustments is in keeping with virtually all of the existing literature on pricing under demand uncertainty; however, this assumption stands in contrast to a large body of evidence suggesting that frequent price changes are inherently undesirable. As noted in Chapter 1, there is a significant amount of evidence (see, for example, Levy et al. (1997), Levy et al. (1998), Zbaracki et al. (2004)) showing that the costs of implementing frequent price changes in a traditional retail setting can amount to a considerable portion of the seller's net margins. Even in an online retail setting, where price adjustments may be less costly (Brynjolfsson and Smith (2000)), there is evidence to suggest that frequent

fluctuations in price may be upsetting to customers (Weiss and Mehrotra (2001)).

In this chapter, we address these concerns by investigating the fundamental limit on the minimum number of price adjustments needed for optimal dynamic pricing. Specifically, we will consider the *minimal switching rate* of a problem instance, which we define to be the minimum number of price changes necessary for a switching-constrained policy to match the regret of the optimal unconstrained policy, as determined in Chapter 2.

As the main contribution of this chapter, we study dynamic pricing under a general parametric choice model, and show that among the class of pricing policies that adjust prices according to an arbitrary but deterministic schedule, the worst-case minimum number of price adjustments needed to match the performance of the rate-optimal unconstrained policy is $\Omega(\log T)$. To prove this result, we leverage the result proved in Chapter 2 that the worst-case rate of regret of the optimal unconstrained pricing policy is $\Theta(\sqrt{T})$, and establish an intrinsic connection between this rate of regret and the minimum number of price adjustments needed to achieve this performance. We also leverage the techniques developed in Chapter 2 to provide a simple policy achieving the optimal order of regret with only $\mathcal{O}(\log T)$ price adjustments, showing that the lower bound on the number of price adjustments needed is tight up to constant factors.

We also consider pricing with minimal price adjustments in the well-separated case studied in Chapter 2, in which the optimal rate of regret is $\Theta(\log T)$, as opposed to $\Theta(\sqrt{T})$. Despite the large difference in the optimal rate of regret, we show that in this case, at least $\Omega(\log T)$ price adjustments are still needed to achieve the optimal $\Theta(\log T)$ regret. To establish this bound, we actually prove a stronger result: we show that, for the well-separated case, any pricing policy that

switches prices at most $\tau - 1$ times in T time periods must incur regret that is $\Omega(\tau T^{1/\tau})$, for $\tau \in \{1, \ldots, \lceil \log T \rceil\}$. Thus, in this special case, we derive a more fine-grained result which describes an explicit relationship between the exact number of switches and the best achievable rate of regret; in particular, this result shows how the best achievable performance of a pricing policy improves as the number of allowed price adjustments increases. We also describe a simple heuristic achieving the optimal rate of regret with the minimal number of price adjustments, showing that our bounds for the well-separated case are tight up to constant factors.

To our knowledge, these bounds on the number of price adjustments are the first results of their kind, establishing fundamental limits on the number of price adjustments necessary for effective dynamic pricing under a general parametric demand model. Moreover, the proofs of these results establish intrinsic connections between the structure of the demand families, the optimal rate of regret, and the minimal rate of price adjustments.

The Explore / Exploit Tradeoff: Ubiquitous in dynamic pricing studies of this type is the *exploration / exploitation* tradeoff: How should a seller balance exploratory pricing, which may sacrifice immediate revenues for better information about the demand curve, with exploitative pricing, which acquires the maximum amount of short-term revenue without regard for reducing uncertainty about demand? This tradeoff usually plays a central role in studies of pricing under demand uncertainty, and so a discussion of this concept in the context of adjustment-constrained pricing is in order. The results of this work suggest that, in our formulation, price adjustment constraints do not impede exploration. Indeed, due to the parametric notion of demand uncertainty considered in our model, we can compute a good estimate of the entire demand curve by offering a relatively small

number of "test prices," and so frequent price adjustments are not necessary for demand learning. In contrast, our results suggest that adjustment constraints do impede a seller's ability to efficiently exploit the information gained about the demand curve, since the offered price may only be updated a limited number of times, regardless of the amount of knowledge that may be acquired.

In this chapter, we focus on the same pricing model described in Section 2.1.1, with the addition of the following performance measure and terminology. We will say that a pricing policy $\psi = (\psi_1, \psi_2, \ldots)$ makes a *price change* (or price switch, or price adjustment) in time period t if the pricing function ψ_t used by ψ to set the price in period t is not equal to the pricing function ψ_{t-1} used to set the price in the previous period. Accordingly, the cumulative number of price switches over T periods is defined to be

$$\text{Switch}(\mathcal{C}, T, \psi) = 1 + \left|\left\{\, 2 \le t \le T \,:\, \psi_t(\,\cdot\,) \neq \psi_{t-1}(\,\cdot\,) \,\right\}\right| \ .$$

3.1.1 Contributions and Organization

In Theorem 3.2.1 of Section 3.2, we establish a stronger version Theorem 2.3.1, giving an $\Omega(\sqrt{T})$ lower bound on the risk under an arbitrary policy, without any constraint on the number of price adjustments. This lower bound represents the smallest achievable worst-case regret for an arbitrary pricing policy, and pricing policies that achieve this lower bound (up to constant factors) are considered to be *regret-optimal*. As we will see, this stronger version of the regret lower bound will be needed to prove the desired lower bound on the minimal switching rate of an arbitrary pricing policy. In Section 3.3, we consider the minimum number of price adjustments needed for effective dynamic pricing, and show in Theorem 3.3.1 that without advance knowledge of the time horizon, any regret-optimal pricing policy

that switches price according to an arbitrary but pre-specified schedule must adjust is prices at least $\Omega(\log T)$ times. These results establish a fundamental lower limit on both the amount of revenue lost by a pricing policy due to demand uncertainty, and also on the minimum number of price adjustments needed by a policy to achieve the optimal rate of performance. Moreover, the proof of Theorem 3.3.1 establishes an intrinsic connection between the performance of a pricing policy and the minimum number of price adjustments necessary.

In Section 3.4, we establish that the lower bounds presented in Sections 3.2 and 3.3 are tight up to constant factors, by constructing a pricing policy that achieves the optimal order of regret, while performing the minimum number of price adjustments. We begin Section 3.4.1 by considering the special case of *horizon-dependent policies*, in which the time horizon is fixed and known to the seller in advance, and the seller is only concerned with the weaker benchmark of meeting a regret guarantee *at the single pre-specified time horizon*. In this case, we show that one can leverage advance knowledge of the time horizon to design a pricing policy that performs well under this weaker benchmark with only a *constant* number of price adjustments, independent of value of the time horizon. Such a policy, however, is designed to perform well only at a specific time horizon, and will not work well in settings where the time horizon is not known in advance. In Section 3.4.2, we extend the ideas developed in the horizon-dependent case to design policies that are truly regret-optimal, in that they operate without advance knowledge of the time horizon, and achieve the optimal regret *uniformly over time*. Specifically, we describe a regret-optimal policy that performs $\mathcal{O}(\log T)$ price adjustments while achieving the optimal $\mathcal{O}(\sqrt{T})$ regret, matching the lower bounds on the regret and switching rate established in Sections 3.2 and 3.3.

In Section 3.5, we consider the well-separated case of the dynamic pricing problem, in which the optimal rate of regret is $\Theta(\log T)$. We show that despite the large difference in the optimal rate of regret between this case and the general case, a pricing policy achieving the optimal rate of regret in this scenario must still switch prices $\Omega(\log T)$ times. Additionally, we prove bounds relating the exact number of price changes to the best achievable performance, establishing for this special case a stronger style of result than for the general case.

3.1.2 Literature Review

The problem of dynamic pricing without price adjustment constraints has been widely studied in the revenue management literature, under a variety of modeling assumptions. The related literature on this topic is discussed extensively in Section 2.1.2. To our knowledge, none of the works discussed there explicitly consider the number of price changes made by the seller. The number of dynamic pricing studies that do consider price adjustment constraints is quite limited, and we review a handful of notable examples here. Feng and Gallego (1995) consider the problem of selling a fixed stock of items over a finite horizon, and analyze the optimal timing of a single price change from a known initial price to a given second price. Bitran and Mondschein (1997) consider a capacitated pricing problem in which the price can only be adjusted at a set of pre-specified times. Netessine (2006) considers a deterministic demand model, and derives structural results on the optimal timing of price changes, and Çelik et al. (2009) take a dynamic programming approach to dynamic pricing under price-adjustment costs, and provide an analysis of several heuristic policies. All of the aforementioned studies assume that the distribution of demand at each price level is *known in advance*, and focus on the problem of pricing under inventory constraints. In contrast, we consider a

model without capacity constraints, in which the relationship between price and demand is *not* known to the seller. Thus, to our knowledge, this is the first work that considers the problem of learning the probabilistic relationship between price and demand with minimal price adjustments, and it represents a significant departure from the antecedent literature.

3.2. Minimum Regret Under Arbitrary Policies

The goal of this section is to establish an $\Omega(\sqrt{T})$ lower bound on the worst-case cumulative risk under an *arbitrary* policy (without any constraint on the price adjustments). This lower bound will serve as a benchmark for assessing the impact of price adjustment constraints, and will be leveraged in Section 3.3, in which we prove a lower bound on the number of price adjustments necessary for a policy to have the optimal $\mathcal{O}(\sqrt{T})$ regret. While a similar lower bound of $\Omega(\sqrt{T})$ was proved in Theorem 2.3.1, we now give a stronger version of this result, and in doing so, derive intermediate results that are crucial for analyzing the minimal switching rate of an arbitrary pricing policy. The connection between the result of this section and Theorem 2.3.1 is discussed in detail in Remark 3.2.5.

To prove the desired lower bound on the risk of an arbitrary policy, we will construct a problem class $\mathcal{C}_{\mathsf{GenLB}} = (\mathcal{P}, \mathcal{Z}, d)$ and a density λ over $\mathcal{Z}$ such that for any policy ψ,

$$\mathbb{E}_{\lambda}[\text{Regret}(Z, \mathcal{C}_{\mathsf{GenLB}}, T, \psi)] = \Omega(\sqrt{T}) \ .$$

Note that this immediately implies the existence of some $z \in \mathcal{Z}$ such that $\text{Regret}(z, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) = \Omega(\sqrt{T})$, which gives a worst-case lower bound on the regret. To establish the desired lower bound on the risk, we will focus on the problem class $\mathcal{C}_{\mathsf{GenLB}} = (\mathcal{P}, \mathcal{Z}, d)$ with $\mathcal{P} = [\sqrt{2}/2, \sqrt{3}/2]$, $\mathcal{Z} = [1/3, 2/3]$, and

$d(p;z) = \sqrt{z} - pz$. It is straightforward to check that $\mathcal{C}_{\mathsf{GenLB}}$ satisfies the conditions of Assumption 1 and 2. The main result of this section is stated in the following theorem.

Theorem 3.2.1 (Risk Lower Bound)**.** *For any policy ψ and $T \geq 2$,*

$$\mathbb{E}_\lambda[\text{Regret}(Z, \mathcal{C}_{\mathsf{GenLB}}, T, \psi)] \geq e^{-10} \cdot \sqrt{T} \ ,$$

where the random variable $Z \in [1/3, 2/3]$ has a density function $\lambda : [1/3, 2/3] \to \mathbb{R}_+$ given by $\lambda(x) = 6\{\cos(3\pi(x - 1/2))\}^2$ for all $x \in [1/3, 2/3]$.

To prove Theorem 3.2.1, we prove two lemmas that establish a fundamental tension between two competing objectives: reducing uncertainty about the demand curve to minimize future revenue losses, and pricing close to the optimal to minimize immediate losses. Specifically, Theorem 3.2.1 makes use of Lemma 3.2.2, which provides a lower bound on the expected *instantaneous* risk in period $t+1$ in terms of the *expected Fisher information* gained by the policy up to time t, which is defined to be $\mathbb{E}\left[\left(\frac{d}{dz}\log Q_t^{\psi,Z}(\mathbf{Y}_t)\right)^2\right]$. [1] The Fisher information is a well-studied quantity in the theory of parameter estimation, and can be thought of as quantifying the amount of "certainty" one has about the parameters of a distribution, based on a set of observed samples from that distribution. In our setting, the Fisher information of the unknown parameter z serves as a measure of the total amount of exploration that we have done thus far. Lemma 3.2.2 follows directly from van Trees' inequality (Theorem 2, Gill and Levit (1995)), and its proof is deferred to Appendix B.1.

Lemma 3.2.2 (Little Exploration Implies Large Instantaneous Risk)**.** *For any*

[1] For a scalar parameter z, this definition of Fisher information is equivalent to the definition provided in Assumption 2 (see, for example, Cover and Thomas (1999), Section 11). We find it convenient to use this alternative definition in the analysis.

policy ψ *and* $t \geq 1$*,*

$$\mathbb{E}[r(p^*(Z); Z) - r(P_{t+1}; Z)] \geq \frac{(1/27)}{\mathbb{E}\left[\left(\frac{d}{dz} \log Q_t^{\psi,Z}(\mathbf{Y}_t)\right)^2\right] + 36\pi^2} \ ,$$

where $\mathbb{E}[\,\cdot\,]$ *denotes expectation with respect to the joint distribution of* $\mathbf{Y}_t$ *and* Z*.*

The result of Lemma 3.2.2 shows that to have small instantaneous risk in the current period, we must have performed sufficient exploration of the demand curve in the past. However, we will show in the following lemma that there is an inherent cost to demand exploration, in that any policy that performs a sufficient amount of exploration must also incur large cumulative risk. This intuition is made precise in Lemma 3.2.3, which establishes a lower bound on the cumulative risk in terms of the Fisher information. (Note that while the result of Lemma 3.2.2 would hold for a generic problem class, the result of Lemma 3.2.3 leverages the specially chosen demand family of the problem class $\mathcal{C}_{\mathsf{GenLB}}$ to prove the desired relationship between the Fisher information and the cumulative risk). The proof of Lemma 3.2.3 is given in Appendix B.2.

Lemma 3.2.3 (Large Exploration Implies Large Cumulative Risk)**.** *For any policy* ψ*, any* $z_0 \in [1/3, 2/3]$*, and* $t \geq 1$*, we have*

$$\mathbb{E}_{z_0}\left[\left(\frac{d}{dz} \log Q_t^{\psi,z}(\mathbf{Y}_t)\,\Big|_{z=z_0}\right)^2\right] \leq 30\,\mathrm{Regret}(z_0, \mathcal{C}_{\mathsf{GenLB}}, t, \psi) \ ,$$

where $\mathbb{E}_{z_0}[\,\cdot\,]$ *denote the expectation when the underlying parameter is* z_0*.*

Lemmas 3.2.2 and 3.2.3 demonstrate the tradeoff between reducing the instantaneous risk in the current time period, and minimizing the total cumulative risk incurred by performing sufficient exploration. We will exploit the tension between these competing objectives to prove Theorem 3.2.1; however, we first need the

following technical lemma, which gives a lower bound on the instantaneous risk in the first period. The proof of this result is given in Appendix B.3.

Lemma 3.2.4. *For any policy ψ, $\mathbb{E}_\lambda[\mathrm{Regret}(Z, \mathcal{C}_{\mathsf{GenLB}}, 1, \psi)] \geq 1/26244$.*

Here is the proof of Theorem 3.2.1.

Proof. Combining Lemmas 3.2.2 and 3.2.3, we have that for any $t \geq 1$,

$$\mathbb{E}[r(p^*(Z); Z) - r(P_{t+1}; Z)] \geq \frac{(1/27)}{30\ \mathbb{E}_\lambda[\mathrm{Regret}(Z, \mathcal{C}_{\mathsf{GenLB}}, t, \psi)] + 36\pi^2}\ .$$

Since the cumulative risk is non-decreasing, it follows from Lemma 3.2.4 that

$$\mathbb{E}_\lambda[\mathrm{Regret}(Z, \mathcal{C}_{\mathsf{GenLB}}, t, \psi)] \geq \mathbb{E}_\lambda[\mathrm{Regret}(Z, \mathcal{C}_{\mathsf{GenLB}}, 1, \psi)] \geq \frac{1}{26244}\ .$$

Moreover, it is easy to verify that for any $(a, b, c) \in \mathbb{R}^3_{++}$, $\frac{a}{bx+c} \geq \frac{ax_0/(bx_0+c)}{x}$ for all $x \geq x_0$. Let $a = 1/27$, $b = 30$, $c = 36\pi^2$, and $x_0 = 1/26244$. It follows that

$$\mathbb{E}[r(p^*(Z); Z) - r(P_{t+1}; Z)] \geq \frac{c_1}{\mathbb{E}_\lambda[\mathrm{Regret}(Z, \mathcal{C}_{\mathsf{GenLB}}, t, \psi)]} \tag{3.1}$$

for $c_1 = \frac{(1/27)(1/26244)}{(30/26244)+36\pi^2} \geq 1/e^{20}$.

Letting $R_t = \mathbb{E}_\lambda[\mathrm{Regret}(Z, \mathcal{C}_{\mathsf{GenLB}}, t, \psi)]$, and noting that $\mathbb{E}[r(p^*(Z); Z) - r(P_{t+1}; Z)] = R_{t+1} - R_t$ by definition, the above argument shows that for all $t \geq 1$,

$$R_{t+1} \geq R_t + \frac{c_1}{R_t}\ .$$

We will now prove by induction on t that the above recursion implies that $R_t \geq \sqrt{c_1} \cdot \sqrt{t}$ for all $t \geq 2$, which gives the desired result because $\sqrt{c_1} \geq 1/e^{10}$. The case when $t = 2$ is trivial because $R_2 \geq R_1 + \frac{c_1}{R_1} \geq \sqrt{R_1^2 + 2c_1} \geq \sqrt{c_1}\sqrt{2}$. So suppose that the claim holds for $t \geq 2$, that is, $R_t \geq \sqrt{c_1} \cdot \sqrt{t}$. Then,

$$R_{t+1} \geq R_t + \frac{c_1}{R_t} \geq \sqrt{R_t^2 + 2c_1} \geq \sqrt{c_1 t + 2c_1} = \sqrt{c_1(t+1) + c_1} \geq \sqrt{c_1} \cdot \sqrt{t+1}\,,$$

which completes the induction. □

Remark 3.2.5 (Contrast Between Theorem 3.2.1 and Previous Results in the Literature)**.** Theorem 3.2.1 is a stronger result than the $\Omega(\sqrt{T})$ lower bound established in Theorem 2.3.1. Specifically, the result of Theorem 2.3.1 states that there exists some constant c such that, given a fixed time horizon T_0, it is possible to construct a pair of problem instances *depending on* T_0 such that the worst-case regret of any pricing policy on these two instances *at the pre-specified time* T_0 must be bounded below by $c\sqrt{T_0}$. Note that for a fixed time horizon T_0, this result does not give any information on the regret of a pricing policy at intermediate time steps $t \in \{1, \ldots, T_0 - 1\}$. In contrast, the result of Theorem 3.2.1 states that there exists some fixed density λ over the problem parameters, which is independent of T, such that the expected regret with respect to λ grows *uniformly* like $\sqrt{t}$. As we will see in Section 3.3, the new result of Theorem 3.2.1 will be crucial in our analysis of the minimal switching rate.

3.3. Minimum Number of Price Adjustments

In this section, we establish a lower bound on the minimum number of price changes needed for optimal dynamic pricing, by showing that among regret-optimal policies that switch prices according to an arbitrary but pre-specified schedule, the minimum number of price-adjustments is $\Omega(\log T)$. Let Ψ_F denote the class of pricing policies that change the price at fixed pre-specified time points, *independent* of the observed customer responses. In the next theorem, we show that any pricing policy in Ψ_F whose regret is $\mathcal{O}(\sqrt{T})$ across all problem instances must *necessarily* adjust its pricing function at least $\Omega(\log T)$ times in the worst case. Later, in Section 3.4.2, we will introduce a pricing policy in the class Ψ_F called the DOUBLING policy, which achieves the optimal order of regret while switching prices only $\mathcal{O}(\log T)$ times, demonstrating that $\Theta(\log T)$ is in fact the minimal switching

rate.

Theorem 3.3.1 (Minimum Price Changes)**.** *Let* $\mathcal{C}_{\mathsf{GenLB}} = (\mathcal{P}, \mathcal{Z}, d)$ *be the problem class used in Theorem 3.2.1 with* $\mathcal{P} = [\sqrt{2}/2, \sqrt{3}/2]$*,* $\mathcal{Z} = [1/3, 2/3]$*, and* $d(p; z) = \sqrt{z - p\,z}$*. If* ψ *is a policy in* Ψ_F *such that for all* $T \geq 1$ *and* $z \in \mathcal{Z}$*,*

$$\mathrm{Regret}(z, \mathcal{C}_{\mathsf{GenLB}}, T, \psi) \leq C\sqrt{T}\ ,$$

for some constant C *depending only on* $\mathcal{C}_{\mathsf{GenLB}}$*, then for all* $T \geq 1$*,*

$$\mathrm{Switch}(\mathcal{C}_{\mathsf{GenLB}}, T, \psi) \geq \frac{\log T}{2\log(1 + C^2 e^{20})}\ .$$

Geometric Intuition: Consider a switching-constrained pricing policy operating on the problem instance of Theorem 3.2.1, and suppose that this policy offers a fixed price P for some fixed number of time steps. The cumulative risk incurred by such a policy during this period is a linear function of the number of time steps, with slope given by the instantaneous risk $\mathbb{E}[r(p^*(z); z) - r(P; z)]$ incurred from offering price P. Thus, if we graph the cumulative risk of any policy in Ψ_F as a function of the number of time steps, then this graph will be a continuous piecewise-linear curve, whose individual segments correspond to the phases in which the different fixed prices were offered. In addition to the piecewise linear structure of the cumulative risk, we also know from Theorem 3.2.1 that the cumulative risk of any regret-optimal pricing policy in Ψ_F must be $\Theta(\sqrt{T})$. In light of these facts, we may regard Figure 3.1 as a schematic representation of the cumulative risk of a hypothetical pricing policy in Ψ_F.

Under this interpretation, we see that the smallest number of price changes necessary for a switching-constrained policy to be regret-optimal is equivalent to the minimum number of segments necessary for a continuous piecewise-linear curve

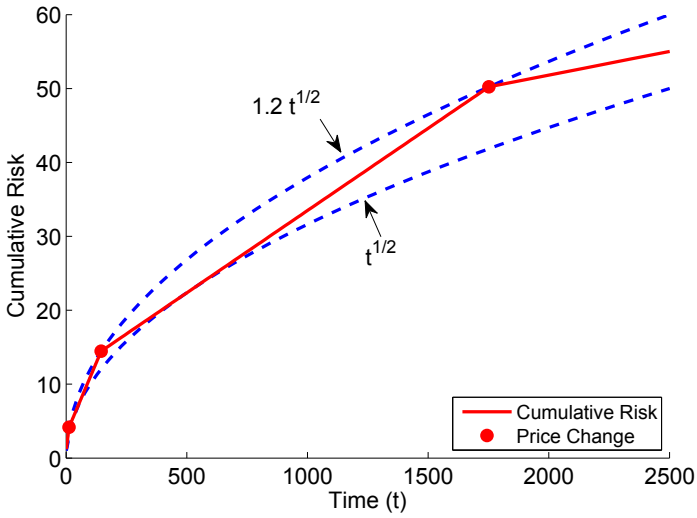


Figure 3.1: Cumulative risk of a switching constrained policy. During periods where the policy offers a fixed price, the cumulative regret increases linearly.

to stay bounded between two curves of the form $t \mapsto c\sqrt{t}$. Clearly, the number of segments needed must tend to infinity as $t \to \infty$, and in fact, the number is $\Omega(\log t)$. Based on this intuition, we now give a rigorous proof of Theorem 3.3.1.

Proof of Theorem 3.3.1. Consider an arbitrary policy $\psi \in \Psi_F$. Let $\{s_k \in \mathbb{Z}_+ : k = 1, 2, \ldots\}$ denote the deterministic sequence of time periods in which the policy ψ makes its price adjustments, where we set $s_1 = 1$. Note that if $P_1, P_2, \ldots$ denotes the sequence of prices under the policy ψ, then for all $k \in \mathbb{Z}_+$, we have

$$P_{s_k} = P_{s_k+1} = \cdots = P_{s_{k+1}-1} \ ,$$

with probability one. From Equation (3.1) in the proof of Theorem 3.2.1, we have that for all $k \in \mathbb{Z}_+$,

$$\mathbb{E}[r(p^*(Z); Z) - r(P_{s_k}; Z)] \ \geq \ \frac{c_1}{\mathbb{E}_\lambda[\mathrm{Regret}(Z, \mathcal{C}_{\mathsf{GenLB}}, s_k - 1, \psi)]} \geq \frac{c_1}{C\sqrt{s_k - 1}} \geq \frac{c_1}{C\sqrt{s_k}} \ ,$$

where $c_1 = \frac{(1/27)(1/26244)}{(30/26244)+36\pi^2} \geq 1/e^{20}$, and where the second inequality follows from our hypothesis that ψ is regret-optimal. Since the policy ψ uses the same price between periods s_k and $s_{k+1}-1$, it follows that

$$\sum_{t=s_k}^{s_{k+1}-1} \mathbb{E}[r(p^*(Z);Z)-r(P_t;Z)] = (s_{k+1}-s_k)\,\mathbb{E}[r(p^*(Z);Z)-r(P_{s_k};Z)] \geq (s_{k+1}-s_k)\,\frac{c_1}{C\sqrt{s_k}}\,.$$

On the other hand, we also have the following upper bound on the expected cumulative regret during periods s_k and $s_{k+1}-1$:

$$\begin{aligned}\sum_{t=s_k}^{s_{k+1}-1} \mathbb{E}[r(p^*(Z);Z) - r(P_t;Z)] &\leq \mathbb{E}_\lambda[\mathrm{Regret}(Z, \mathcal{C}_{\mathsf{GenLB}}, s_{k+1}-1, \psi)] \\ &\leq \mathbb{E}_\lambda[\mathrm{Regret}(Z, \mathcal{C}_{\mathsf{GenLB}}, s_{k+1}, \psi)] \\ &\leq C\sqrt{s_{k+1}}\ ,\end{aligned}$$

where the last inequality follows from our hypothesis that ψ is regret optimal. Combining the above two inequalities, we have that

$$(s_{k+1}-s_k)\,\frac{c_1}{C\sqrt{s_k}} \ \leq \ \sum_{t=s_k}^{s_{k+1}-1} \mathbb{E}[r(p^*(Z);Z) - r(P_t;Z)] \ \leq \ C\sqrt{s_{k+1}}\ ,$$

which implies that

$$\frac{C^2}{c_1} \ \geq \ \frac{s_{k+1}-s_k}{\sqrt{s_k s_{k+1}}} \ = \ \sqrt{\frac{s_{k+1}}{s_k}} - \sqrt{\frac{s_k}{s_{k+1}}} \ \geq \ \sqrt{\frac{s_{k+1}}{s_k}} - 1\ ,$$

where the last inequality follows from the fact that $s_k \leq s_{k+1}$. Let $\alpha = (1+C^2/c_1)^2$. It follows from the above inequality that $s_{k+1} \leq \alpha s_k$ for all k. Since $s_1 = 1$, we have that $s_{k+1} \leq \alpha^k$, which implies $k \geq \log_\alpha s_{k+1}$ for all $k \geq 1$.

Now, consider an arbitrary time period T. Suppose that $s_k \leq T < s_{k+1}$ for some k. Then,

$$\mathrm{Switch}(\mathcal{C}_{\mathsf{GenLB}}, T, \psi) = k \geq \log_\alpha s_{k+1} \geq \log_\alpha T = \frac{\log T}{\log \alpha}\ ,$$

and the desired result follows from the fact that $\log\alpha = 2\log(1 + C^2/c_1)$ and $1/c_1 \leq e^{20}$. □

3.4. Matching Upper Bounds

In this section, we construct a pricing policy that achieves $\mathcal{O}(\sqrt{T})$ regret uniformly over time, while switching prices at most $\mathcal{O}(\log T)$ times, matching the upper bounds presented in Sections 3.2 and 3.3. We begin by our analysis by consider the special case of a known time horizon.

3.4.1 A Motivating Example: The Horizon-Dependent Case

In this section, we consider a special case in which the time horizon is fixed to be some value T_0 which is known to the pricing policy in advance, and we consider the goal of achieving a regret guarantee *at the pre-specified time* T_0. We will see that when the time horizon is known in advance, there is a simple regret-optimal policy that adjusts the offer price only a *constant* number of times, independent of the time horizon T_0. While this section illustrates the benefits of having a priori knowledge of the time horizon, the primary goal is to highlight important design principles that will be used in the next section, when we develop pricing policies for more general settings.

To motivate the design of a pricing policy, recall that in Corollary 2.2.4, we saw that if the model parameter is $\mathbf{z}$, and we price the product at $p^*(\hat{\mathbf{z}})$ for some estimate $\hat{\mathbf{z}}$ of $\mathbf{z}$, then we have

$$r(p^*(\mathbf{z}); \mathbf{z}) - r(p^*(\hat{\mathbf{z}}); \mathbf{z}) \;\leq\; \frac{c_r\, L^2}{2}\, \|\mathbf{z} - \hat{\mathbf{z}}\|^2 \;.$$

Thus, given an estimate $\widehat{\mathbf{Z}}$ of the underlying parameter $\mathbf{z}$, it follows from Corollary 2.2.4 that the instantaneous regret is bounded above by the mean squared error of $\widehat{\mathbf{Z}}$. Recall also that by Lemma 2.3.7, the mean-squared error of the max-

imum likelihood estimate $\widehat{Z}_n$ based on t samples is bounded above by C_{mle}/t for some constant C_{mle} depending only on the underlying problem instance. Together, Corollary 2.2.4 and Lemma 2.3.7 suggest a simple pricing policy that achieves $\mathcal{O}(\sqrt{T})$ regret with only a constant number of prices changes. The policy takes advantage of the advance knowledge of the time-horizon to schedule all of its exploration "up-front," eliminating the need for frequent price adjustments. Under this policy, we first offer the exploration price $\bar{p}_1$ for $\lfloor\sqrt{T_0}\rfloor$ consecutive periods, followed by $\bar{p}_2$ for another $\lfloor\sqrt{T_0}\rfloor$ consecutive periods, followed by $\bar{p}_3$ for $\lfloor\sqrt{T_0}\rfloor$ periods, and so on, continuing in this fashion until we have offered the last exploration price $\bar{p}_n$. Note that during this exploration phase, we only change the price at most n times, independent of the time horizon T_0. Moreover, since the instantaneous regret in each time period is bounded by some constant C_0, the total cumulative regret during the exploration phase is $\mathcal{O}(nC_0\sqrt{T_0})$.

After concluding the exploration phase, we compute a maximum likelihood estimate $\widehat{\mathbf{Z}}$ from the $n\lfloor\sqrt{T_0}\rfloor$ exploration samples. We then use a single fixed price, correspond to the myopic price $p^*(\widehat{\mathbf{Z}})$, for the remaining time periods. It follows from Lemma 2.3.7 that the mean squared error of $\widehat{\mathbf{Z}}$ is $\mathcal{O}(1/\sqrt{T_0})$, and thus the expected instantaneous regret in each period is also $\mathcal{O}(1/\sqrt{T_0})$ by Corollary 2.2.4. Therefore, the cumulative regret incurred after the exploration phase will be $\mathcal{O}(\sqrt{T_0})$ because there are at most T_0 time periods remaining. This gives us the desired regret bound. It is clear that the total number of price changes is at most $n + 1$.

The above policy takes advantage of the known time horizon T_0, and it is designed specifically to yield small regret over T_0 periods. Clearly, this policy will perform poorly when we have a longer time horizon because of insufficient

exploration. In the next section, we will design a pricing policy that is truly regret-optimal, in that it satisfies a regret guarantee of $\mathcal{O}(\sqrt{T})$ *uniformly* for all time horizons T. Although we cannot hope that such a policy would change prices a constant number of times, it turns out, surprisingly, that the number of required price adjustments is still relatively small compared to the overall time horizon, matching the $\Omega(\log T)$ bound on price adjustments proved in Section 3.3.

3.4.2 A Regret-Optimal Policy with Minimum Price Adjustments

In this section, we describe a regret-optimal pricing policy – which we refer to as the DOUBLING policy – whose regret is bounded above by $\mathcal{O}(\sqrt{T})$ *uniformly* for all time horizon T. Using a standard doubling argument (Section 2.3, Cesa-Bianchi and Lugosi (2006)), the DOUBLING policy changes its offer price at pre-specified time periods that are (exponentially) far apart. To describe the exact time periods where the price changes occur, it is conceptually helpful to think of the DOUBLING policy as operating in cycles of (exponentially) increasing lengths, where each cycle consists of an exploration phase followed by an exploitation phase. Within each cycle, the policy changes the offer price only a constant a number of times, leading to a total number of price adjustments of $\mathcal{O}(\log T)$. A formal description of the policy is given below.

The following theorem establishes an upper bound on the cumulative regret and the expected number of price adjustments for the DOUBLING policy.

Theorem 3.4.1. *For any problem class $\mathcal{C}$ satisfying Assumptions 1 and 2, there exists a positive constant C_1 depending only on $\mathcal{C}$ such that for every $T \geq n^2$,*

$$\text{Regret}(\mathbf{z}, \mathcal{C}, T, \mathsf{Doubling}(\mathcal{C})) \leq C_1\sqrt{T} \qquad \textit{and} \qquad \text{Switch}(\mathcal{C}, T, \mathsf{Doubling}(\mathcal{C})) \leq (n+1)\log_2 T.$$

Policy Doubling($\mathcal{C}$)

Inputs: A problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ and exploration prices $\bar{\mathbf{p}} = (\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_n)$ satisfying Assumptions 1 and 2.

Description: For cycles $c = 1, 2, \ldots$

Exploration phase of cycle c: For $i = 1, 2, \ldots, n$, offer the exploration price $\bar{p}_i$ to $2^{\lfloor c/2 \rfloor}$ *consecutive* customers. For $m = 1, 2, \ldots, 2^{\lfloor c/2 \rfloor}$, let $Y_i(m)$ denote the response of the m^{th} customer when the exploration price $\bar{p}_i$ is offered, and let $\mathbf{Y}(m) = (Y_1(m), \ldots, Y_n(m)) \in \{0, 1\}^n$.

At the end of the exploration phase, let $\widehat{\mathbf{Z}}(c)$ denote the maximum likelihood estimate based on the customer selections $\mathbf{Y}(1), \mathbf{Y}(2), \ldots, \mathbf{Y}(2^{\lfloor c/2 \rfloor})$ from cycle c, that is,

$$\widehat{\mathbf{Z}}(c) = \arg\max_{\mathbf{z} \in \mathcal{Z}} \prod_{m=1}^{2^{\lfloor c/2 \rfloor}} Q^{\bar{\mathbf{p}}, \mathbf{z}}(\mathbf{Y}(m)) \ .$$

Exploitation phase of cycle c: Offer the price $p^*(\widehat{\mathbf{Z}}(c))$ for an additional 2^c periods.

Proof. Consider any arbitrary $T \geq 1$. There are at most $\log_2 T$ cycles from time one until the time horizon T. From the definition of the DOUBLING policy, the number of price changes in each cycle is at most $n+1$. Therefore, $\text{Switch}(\mathcal{C}, T, \text{Double}(\mathcal{C})) \leq (n+1) \log_2 T$. It remains to bound the regret.

Consider an arbitrary cycle c. During each period in the exploration phase, the instantaneous regret is bounded above by some constant C_0 that depends only on $\mathcal{C}$. The total regret during the exploration phase of cycle c is thus bounded by $n 2^{\lfloor c/2 \rfloor} C_0$. We will now bound the total regret during the exploitation phase. It follows from Corollary 2.2.4 and Lemma 2.3.7 that we have the following upper bound on the instantaneous regret in each period of the exploitation phase of cycle c.

$$\mathbb{E}_{\mathbf{z}} \left[r(p^*(\mathbf{z}); \mathbf{z}) - r(p^*(\widehat{\mathbf{Z}}(c)); \mathbf{z}) \right] \leq \frac{c_r L^2}{2} \mathbb{E}_{\mathbf{z}} \left[\left\| \widehat{\mathbf{Z}}(c) - \mathbf{z} \right\|^2 \right] \leq \frac{c_r L^2 C_{mle}}{2 \cdot 2^{\lfloor c/2 \rfloor}} \leq \frac{c_r L^2 C_{mle}}{2^{c/2}} \ ,$$

where the second inequality follows from the fact that each exploration price is of-

fered to $2^{\lfloor c/2 \rfloor}$ customers. This implies that the total regret during the exploitation phase is bounded above by

$$2^c \cdot \mathbb{E}_{\mathbf{z}} \left[r(p^*(\mathbf{z}); \mathbf{z}) - r(p^*(\widehat{\mathbf{Z}}(c)); \mathbf{z}) \right] \leq c_r L^2 C_{mle} 2^{c/2} \ .$$

Putting everything together, we have that the total regret in cycle c is bounded above by $nC_0 2^{\lfloor c/2 \rfloor} + c_r L^2 C_{mle} 2^{c/2} \leq (nC_0 + c_r L^2 C_{mle}) 2^{c/2}$. Let $K_0 = \lfloor \log_2(2T) \rfloor$. Note that $2^{K_0} \geq 2^{\log_2(2T)}/2 = T$. Thus, the number of cycles is at most K_0. Since the regret is non-decreasing, we have that

$$\begin{aligned}
\text{Regret}(\mathbf{z}, \mathcal{C}, T, \mathsf{Doubling}(\mathcal{C})) &\leq \text{Regret}\left(\mathbf{z}, \mathcal{C}, \sum_{c=1}^{K_0} \left(n 2^{\lfloor c/2 \rfloor} + 2^c \right), \mathsf{Doubling}(\mathcal{C}) \right) \\
&\leq (nC_0 + c_r L^2 C_{mle}) \sum_{c=1}^{K_0} 2^{c/2} \\
&\leq (nC_0 + c_r L^2 C_{mle}) \frac{2^{(K_0+1)/2}}{\sqrt{2} - 1} \\
&\leq (nC_0 + c_r L^2 C_{mle}) \frac{2\sqrt{T}}{\sqrt{2} - 1} \ ,
\end{aligned}$$

which is the desired result. □

3.5. Well-Separated Demand Curves

In the previous sections, we saw that the $\Theta(\log T)$ minimal switching rate for pricing policies in the general case was intrinsically linked to the optimal $\Theta(\sqrt{T})$ rate of regret. This suggests the following natural question: how would the switching behavior of a rate-optimal pricing strategy differ if we considered a model in which the optimal rate of regret was something other than $\Theta(\sqrt{T})$? In this section, we will investigate this question by considering the special class of "well-separated" demand families, under which the optimal rate of regret is $\Theta(\log T)$, as opposed to $\Theta(\sqrt{T})$ for the general case. We will see that, despite this large difference in the

optimal rate of regret, the minimal switching rate for this scenario is still $\Theta(\log T)$. Additionally, in this special case, we will prove bounds that are more fine-grained than those in the previous sections, which show that any pricing policy that adjusts its offer price at most $\tau - 1$ times must incur regret that is $\Omega(\tau T^{1/\tau})$ in the worst case, for $\tau \in \{1, \ldots, \lceil \log T \rceil\}$.

Recall that well-separated families of demand curves satisfy Assumptions 1 and 2 of Section 2.2, and have the additional property that for any two distinct parameter values, the willingness-to-pay distribution for one value stochastically dominates the willingness-to-pay distribution for the other value. This additional structural assumption ensures that we can estimate the parameters of the demand curve from customer responses to *any price*; this is in contrast to Assumption 2 for the general case, in which we were only guaranteed some fixed set of exploratory prices from which we could estimate the demand parameters. As we saw in Chapter 2, this additional property makes dynamic pricing easier, in that the optimal rate of regret $\Theta(\log T)$, as opposed to $\Theta(\sqrt{T})$ for the general case.

We now summarize our results for dynamic pricing under well-separated demand curves. In Section 3.5.1, we show that any regret optimal policy for the well-separated case must switch prices $\Omega(\log T)$ times in the worst-case. We deduce this lower bound on the minimal switching rate as a corollary to a more fine-grained result, which shows that in the well-separated case, any pricing policy that adjusts its offer price at most $\tau - 1$ times must incur regret that is $\Omega(\tau T^{1/\tau})$ in the worst case, for $\tau \in \{1, \ldots, \lceil \log T \rceil\}$. Finally, in Section 3.5.2, we construct a pricing policy that changes prices at most $\mathcal{O}(\log T)$ times in T time periods, and whose regret is $\mathcal{O}(\log T)$ across all problem instances. Thus, we classify that the optimal order of regret for the well-separated case is $\Theta(\log T)$, and that the

minimal switching rate is $\Theta(\log T)$.

3.5.1 Lower Bound on the Switching Rate

In this section, we establish that any pricing policy for the well-separated case that switches prices a finite number of times must incur polynomial regret, which is asymptotically worse than the optimal $\mathcal{O}(\log T)$ regret. As an immediate corollary, we deduce that the minimal switching rate is $\Omega(\log T)$.

To prove the desired lower bounds on the minimal switching rate, we will leverage the risk lower bounds for the well-separated case developed in Section 2.4.1, so let us first recall those results. Recall that the results of that section apply to the problem class $\mathcal{C}_{\mathsf{WellSepLB}} = (\mathcal{P}, \mathcal{Z}, d)$ given by $\mathcal{P} = [1/3, 1/2]$, $\mathcal{Z} = [2, 3]$, with $d(p; z) = 1 - (pz)/2$, and with a density $\lambda : \mathcal{Z} \to \mathbb{R}_+$ given by $\lambda(z) = 2\{\cos(\pi(z - 5/2))\}^2$. Recall that Theorem 2.4.5 established that for any policy ψ setting prices in $\mathcal{P}$ and any $T \geq 1$, if Z is a random variable taking values in $\mathcal{Z}$ with density λ, we have

$$\mathbb{E}_\lambda[\mathrm{Regret}(Z, \mathcal{C}_{\mathsf{WellSepLB}}, T, \psi)] \geq \frac{1}{405\pi^2} \log T \ .$$

Recall also that this theorem followed directly from Lemma 2.4.6, which states that for any $t \geq 1$,

$$\mathbb{E}\left[r(p^*(Z); Z) - r(P_{t+1}; Z)\right] \geq \frac{1}{405\pi^2} \cdot \frac{1}{t},$$

where P_{t+1} is the price offered by ψ at time $t+1$, and $\mathbb{E}[\,\cdot\,]$ denotes the expectation with respect to the joint distribution of P_t and the prior density λ of the parameter $Z \in \mathcal{Z}$. Finally, we note that it is straightforward to check that for any initial price $p_1 \in \mathcal{P}$, we have that

$$\mathbb{E}\left[r(p^*(Z); Z) - r(p_1; Z)\right] \geq \frac{1}{405\pi^2}.$$

With these results, we can proceed to prove a lower bound on the minimal switching rate for the well-separated case. Let $\Psi_F(\tau, T)$ denote the set of pricing policies that switch prices at most $\tau - 1$ times according to a predetermined schedule up to a time horizon T. The main result of the section is stated in the following theorem.

Theorem 3.5.1 (Risk Lower Bound for Deterministic-Switching Policies)**.** *Let $\mathcal{C}_{\mathsf{WellSepLB}}$ and $\lambda(\,\cdot\,)$ be the well-separated problem class and density used in Theorem 2.4.5. For any $T \geq 2$, any $\tau \in \{1, \ldots, \lceil \log T \rceil\}$, and any $\psi \in \Psi_F(\tau, T)$,*

$$\mathbb{E}_\lambda[\mathrm{Regret}(Z, \mathcal{C}_{\mathsf{WellSepLB}}, T, \psi)] \geq \frac{1}{3(405\pi^2)} \cdot \tau T^{1/\tau} \ .$$

In the following corollary, we show that worst-case minimal switching rate for well-separated families is at least $\Omega(\log T)$.

Corollary 3.5.2 (Minimum Price Changes for Well-Separated Families)**.** *If ψ is a policy in $\Psi_F(\tau, T)$ such that for all $T \geq 2$ and $z \in \mathcal{Z}$,*

$$\mathrm{Regret}(z, \mathcal{C}_{\mathsf{WellSepLB}}, T, \psi) \leq C \log T \ ,$$

for some constant C depending only on $\mathcal{C}_{\mathsf{WellSepLB}}$, then for any $T \geq 2$,

$$\mathrm{Switch}(\mathcal{C}_{\mathsf{WellSepLB}}, T, \psi) \geq \frac{1}{6C(405\pi^2)} \cdot \log T \ .$$

Proof. Let $c_3 = 1/\{3(405\pi^2)\}$. By Theorem 3.5.1, we have that for all $z \in \mathcal{Z}$,

$$C \log T \geq \mathrm{Regret}(z, \mathcal{C}_{\mathsf{WellSepLB}}, T, \psi) \geq c_3 \, \tau T^{1/\tau} = c_3 \, \tau e^{(\log T)/\tau} \ ,$$

which implies that

$$\frac{C}{c_3} \cdot \frac{\log T}{\tau} \geq e^{(\log T)/\tau} \geq \frac{1}{2} \left(\frac{\log T}{\tau} \right)^2 \ ,$$

where the last inequality follows from the fact that for any $x \geq 0$, $e^x \geq x^2/2$. Therefore, $\tau \geq (c_3/(2C)) \log T$, which is the desired result. □

We now proceed with the proof of Theorem 3.5.1. We will consider the following optimization problem. For any $T \geq 2$ and $\tau \in \{1, \ldots, \lceil \log T \rceil\}$, define

$$\Gamma^*(\tau, T) = \min \left\{ x_1 + \sum_{k=2}^{\tau} \frac{x_k}{\sum_{h=1}^{k-1} x_h} \;\middle|\; \sum_{k=1}^{\tau} x_k = T \text{ and } x_k \in \mathbb{Z}_+ \cup \{0\} \; \forall k \right\}. \quad (3.2)$$

We will prove Theorem 3.5.1 in two steps, by first establishing a lower bound on the risk in terms of $\Gamma^*(\tau, T)$, and then establishing a lower bound on $\Gamma^*(\tau, T)$ in terms of T and τ. The first step is addressed in the following Lemma.

Lemma 3.5.3 (Risk Lower Bound)**.** *For any $T \geq 2$, any $\tau \in \{1, \ldots, \lceil \log T \rceil\}$, and any $\psi \in \Psi_F(\tau, T)$,*

$$\mathbb{E}_\lambda[\text{Regret}(Z, \mathcal{C}_{\mathsf{WellSepLB}}, T, \psi)] \geq \frac{1}{405\pi^2} \cdot \Gamma^*(\tau, T) \ .$$

where λ is the density given by $\lambda(z) = 2\{\cos(\pi(z - 5/2))\}^2$.

Proof. Consider an arbitrary policy $\psi \in \Psi_F$. Let $\{s_k \in \{1, \ldots, T\} : k = 1, 2, \ldots \tau\}$ denote the time periods where the price switching under ψ occurs, where we set $s_1 = 1$, an define a time $s_{\tau+1} = T + 1$ for notational convenience. Recall that if $P_1, P_2, \ldots, P_T$ denotes the sequence of prices under the policy ψ, then for all $k \in 1, 2, \ldots, \tau$, we have that $P_{s_k} = P_{s_k+1} = \cdots = P_{s_{k+1}-1}$, with probability one. Let $\ell_k = s_{k+1} - s_k$ denote the length of the k^{th} phase, and for ease of notation, let $R_t(z) = r(p^*(z); z) - r(P_t; z)$ denote the instantaneous regret at time t when the underlying parameter is z. Now we have

$$\begin{aligned}
\mathbb{E}_\lambda[\text{Regret}(Z, \mathcal{C}_{\mathsf{WellSepLB}}, T, \psi)] &= \mathbb{E}\left[\sum_{t=1}^{T} R_t(Z)\right] = \mathbb{E}\left[\sum_{k=1}^{\tau} \sum_{h=s_k}^{s_{k+1}-1} R_h(Z)\right] \\
&= \mathbb{E}\left[\ell_1 \cdot R_{s_1}(Z) + \sum_{k=2}^{\tau} \ell_k \cdot R_{s_k}(Z)\right] \\
&= \ell_1 \cdot \mathbb{E}[R_{s_1}(Z)] + \sum_{k=2}^{\tau} \ell_k \cdot \mathbb{E}[R_{s_k}(Z)],
\end{aligned}$$

where Z is a random variable with density λ, and $\mathbb{E}[\cdot]$ denotes the expectation with respect to the joint distribution of Z and the distribution induced by ψ. To relate the last line to the objective function of the minimization problem (3.2), we must establish lower bounds on the terms $\mathbb{E}[R_{s_k}(Z)]$ for $k \in \{1, \ldots, \tau\}$. To do this, note that for the case $k \geq 2$, the price P_{s_k} offered at the beginning of the k^{th} phase is computed from at most $s_k - 1 = \sum_{h=1}^{k-1} \ell_h$ samples, so applying Lemma 2.4.6, we have that for $k \geq 2$,

$$\mathbb{E}[R_{s_k}(Z)] \geq \frac{1}{405\pi^2} \cdot \frac{1}{\sum_{h=1}^{k-1} \ell_h}. \tag{3.3}$$

For the case $k = 1$, we know by Lemma 2.4.6 that for any arbitrary initial price p_1, we have

$$\mathbb{E}_\lambda[r(p^*(Z); Z) - r(p_1; Z)] \geq \frac{1}{405\pi^2} \ .$$

Putting everything together, we have

$$\begin{aligned} \mathbb{E}_\lambda[\text{Regret}(Z, \mathcal{C}_{\mathsf{WellSepLB}}, T, \psi)] &= \frac{1}{405\pi^2} \cdot \left(\ell_1 + \sum_{k=2}^{\tau} \frac{\ell_k}{\sum_{h=1}^{k-1} \ell_h} \right) \\ &\geq \frac{1}{405\pi^2} \cdot \Gamma^*(\tau, T) \ , \end{aligned}$$

which proves the claim. □

Using this fact, we now establish a lower bound on $\Gamma^*(\tau, T)$ in terms of T and τ. This result is given in the following lemma whose proof follows from a standard dynamic programming technique, and can be found in Appendix B.4. Theorem 3.5.1 follows as an immediate corollary of Lemmas 2.4.6 and 3.5.4.

Lemma 3.5.4 (Dynamic Programming Lower Bound)**.** *For any $T \geq 2$ and any $\tau \in \{1, \ldots, \lceil \log T \rceil\}$,*

$$\Gamma^*(\tau, T) \geq \frac{\tau T^{1/\tau}}{3}.$$

3.5.2 A Matching Upper Bound

In this Section, we describe a pricing policy that achieves regret that is $\mathcal{O}(\tau T^{1/\tau})$, while adjusting its offer price at most $\tau - 1$ times, matching the lower bound stated in Section 3.5.1. Below is a description of the policy.

Policy $\mathsf{Well\text{-}Sep}(\mathcal{C}, T, \tau)$

Inputs: A time horizon $T \geq 2$, a well-separated problem instance $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ satisfying Assumptions 1 and 3, and a switching constraint $\tau \in \{1, \ldots, \lceil \log T \rceil\}$.

Initialization: Divide the time-line $\{1, \ldots, T\}$ into τ consecutive phases. For each $1 \leq c \leq \tau - 1$, let the length of phase c be $\ell_c = \lceil T^{c/\tau} \rceil$ time steps, and let the final phase τ contain the remaining time steps. For each $1 \leq c \leq \tau$, let s_c denote the first time period in phase c, and let $s_{\tau+1} = T + 1$.

Description: For phases $c = 1, \ldots, \tau$,

- If $c = 1$, offer an arbitrary initial price $P_1 \in \mathcal{P}$ for $\ell_c = \lceil T^{c/\tau} \rceil$ time periods. Let $\mathbf{Y}_{s_2-1} = (Y_1, \ldots, Y_{s_2-1})$ denote the vector of corresponding customer responses.
- If $c \geq 2$, compute the maximum-likelihood estimate

$$\widehat{Z}(c) = \arg\max_{z \in \mathcal{Z}} Q_{s_c-1}^{\mathsf{Well\text{-}Sep},z}(\mathbf{Y}_{s_c-1}),$$

  and offer $p^*(\widehat{Z}(c))$ for the entirety of phase c. Let $\mathbf{Y}_{s_{c+1}-1} = (Y_1, \ldots, Y_{s_{c+1}-1})$ denote the current vector of customer responses since time 1.

The main result of this section is stated in the following theorem.

Theorem 3.5.5 (Matching Upper Bound for Well Separated Demand Curves)**.** *For any well-separated problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$, $z \in \mathcal{Z}$, $T \geq 2$, and $\tau \in \{1, \ldots \lceil \log T \rceil\}$, there exists a constant C_2 depending only on the problem class $\mathcal{C}$ such that*

$$\mathrm{Regret}(z, \mathcal{C}, T, \mathsf{Well\text{-}Sep}(\mathcal{C}, T, \tau)) \leq C_2\, \tau\, T^{1/\tau} \quad \textit{and} \quad \mathrm{Switch}(\mathcal{C}, T, \mathsf{Well\text{-}Sep}(\mathcal{C}, T, \tau)) \leq \tau\ .$$

Remark 3.5.6. The definition of the policy Well-Sep$(\mathcal{C}, T, \tau)$, as well as the results of Theorem 3.5.5, apply to the case of a known time horizon T and a pre-specified number of switches τ. However, a straightforward modification of the Well-Sep policy yields a policy that operates *without foreknowledge* of the time-horizon, and achieves the optimal $\mathcal{O}(\log T)$ regret, while adjusting prices only $\mathcal{O}(\log T)$ times. Specifically, modifying the Well-Sep policy so that the length of the j^{th} phase is 2^j yields a policy that does not require knowledge of time horizon, and that switches prices $\Theta(\log T)$ times in T periods. Moreover, applying the regret bound of Theorem 3.5.5 with $\tau = \Theta(\log T)$ yields a regret guarantee of $\mathcal{O}(\log T)$, as desired.

The proof of Theorem 3.5.5 makes use of Theorem 2.4.7, which states that if $\widehat{Z}_t$ is a maximum-likelihood estimate computed by an online pricing policy based on t samples, then the mean squared error of $\widehat{Z}_t$ is bounded above by C_{mle}/t for some constant C_{mle} depending only on the underlying problem instance. With this fact, we now give the proof of Theorem 3.5.5.

Proof. Consider a fixed $T \geq 2$ and $\tau \in \{1, \ldots, \lceil \log T \rceil\}$. Let us assume without loss of generality that $\tau \geq 2$, or else the result holds trivially. Since there are τ phases, it is enough to show that the regret in each phase is bounded above by $C_2 \cdot T^{1/\tau}$. Denote the regret incurred in the c^{th} phase by $R_c(z)$, which is given by

$$R_c(z) = \sum_{t=s_c}^{s_{c+1}-1} \mathbb{E}_z[r(p^*(z); z) - r(p^*(\widehat{Z}(c)); z)] = \ell_c \cdot \mathbb{E}_z[r(p^*(z); z) - r(p^*(\widehat{Z}(c)); z)],$$

since the policy offers the price $p^*(\widehat{Z}(c))$ during each of the ℓ_c time periods in phase c. Recall that $s_{\tau+1} = T + 1$ for notational convenience.

To bound the regret in the first phase, note that the instantaneous regret incurred during any single time period is bounded above by some constant C_0

depending only on $\mathcal{C}$. It follows that the total regret incurred during the first phase satisfies

$$R_1(z) \leq C_0 \ell_1 = C_0 \lceil T^{1/\tau} \rceil \leq 2C_0 T^{1/\tau}.$$

To bound the regret in phases $2 \leq c \leq \tau$, note that by Corollary 2.2.4, Assumption 1 (c), and Theorem 2.4.7, we have

$$\begin{aligned}
\mathbb{E}_z[r(p^*(z); z) - r(p^*(\widehat{Z}(c)); z)] &\leq c_r \mathbb{E}_z[(p^*(z) - p^*(\widehat{Z}(c)))^2] \\
&\leq L^2 c_r \mathbb{E}_z[(z - \widehat{Z}(c))^2] \\
&\leq \frac{L^2 c_r C_{mle}}{s_c - 1} \\
&\leq \frac{2L^2 c_r C_{mle}}{s_c},
\end{aligned}$$

where the second to last inequality follows from the fact that $\widehat{Z}(c)$ is computed from $s_c - 1$ samples, and where the last inequality follows from the fact that for $c \geq 2$ we have $s_c \geq 2$, and so $1/(s_c - 1) \leq 2/s_c$. By the definition of the Well-Sep policy and the formula for geometric series, we have that

$$s_c = \sum_{k=1}^{c-1} \lceil T^{k/\tau} \rceil \geq \sum_{k=1}^{c-1} T^{k/\tau} = \frac{T^{c/\tau} - T^{1/\tau}}{T^{1/\tau} - 1} \geq T^{(c-1)/\tau} - 1 \geq \frac{T^{(c-1)/\tau}}{3},$$

where the last inequality holds because for each $2 \leq c \leq \tau$, $T^{(c-1)/\tau} = e^{(c-1)\log T/\tau} \geq e^{\log T/\lceil \log T \rceil} \geq e^{1/2}$. Also by the definition of the Well-Sep policy, we have $\ell_c \leq \lceil T^{c/\tau} \rceil$ for all $1 \leq c \leq \tau$, so it follows that

$$R_c(z) = \ell_c \cdot \mathbb{E}_z[r(p^*(z); z) - r(p^*(\widehat{Z}(c)); z)] \leq \lceil T^{c/\tau} \rceil \cdot \frac{6L^2 c_r C_{mle}}{T^{(c-1)/\tau}} \leq 12 L^2 c_r C_{mle} \cdot T^{1/\tau}.$$

Letting $C_2 = \max\{2C_0, 12L^2 c_r C_{mle}\}$ and putting everything together, we have

$$\text{Regret}(z, \mathcal{C}, T, \text{Well-Sep}(\mathcal{C}, T, \tau)) = \sum_{c=1}^{\tau} R_c(z) \leq C_2 \ \tau T^{1/\tau}.$$

□

3.6. Numerical Experiments

In this section, we compare the empirical performance of four pricing policies that have a wide variety of price-adjustment behavior, focusing exclusively on the general case of dynamic pricing discussed in Sections 3.2, 3.3, and 3.4. We consider the DOUBLING policy described in Section 3.4.2, and compare its performance against three alternative policies across a range of problem instances. We describe the alternative policies below.

1. **FP**: As a baseline for comparison, we consider a fixed-price policy FP that chooses a price uniformly at random from the pricing interval, and offers this price for the entire selling season. Note that this policy will perform zero price adjustments, and will have regret that is linear in T.

2. **MLE-C**: The second policy we consider is the MLE-CYCLE policy described in Section 2.3.2. Note that this policy is identical to the DOUBLING policy, except for the scheduling of the exploration and exploitation phases. In MLE-CYCLE, the length of the c^{th} cycle is $c + n$ time periods, consisting of an exploration phase in which n exploration prices are offered, and an exploitation phase in which the greedy price is offered for c periods. Thus, MLE-CYCLE updates its greedy price more frequently, potentially adjusting its price $\Theta(\sqrt{T})$ times up to time T, compared with $\Theta(\log T)$ times for the DOUBLING policy. Recall that MLE-CYCLE has a regret guarantee of $\mathcal{O}(\sqrt{T})$.

3. **CVP**: The third policy we consider is the Controlled Variance Pricing policy described in den Boer and Zwart (2010b). In time period t, the CVP policy computes a maximum-likelihood estimate of the unknown parameters based

on *all* previously observed samples, and then computes the optimal price with respect to these estimates. The policy then offers the estimated optimal price, with a small perturbation added in the case that this price is too close to the average of all past prices offered. Note that this policy potentially switches prices *in every time step*. The CVP policy has a regret guarantee of $\mathcal{O}(\sqrt{T}\log T)$.

In the implementation of both the DOUBLING and MLE-C policies, we use samples from *both* the exploration *and* exploitation phases to compute the estimates of the demand parameters, as we found this to improve performance in our empirical studies.

To summarize, the DOUBLING, MLE-C, and CVP policies all have a provable regret guarantees of $\mathcal{O}(\sqrt{T})$ (up to logarithmic factors), but these policies perform $\Theta(\log T)$, $\Theta(\sqrt{T})$, and $\Theta(T)$ potential price adjustments, respectively. Our goal in this section is to compare the finite-time regret and switching numbers of these policies, to empirically investigate the relationship between price adjustments and revenue loss, and to and provide numerical evidence to support that the DOUBLING policy can maintain competitive pricing performance while making a relatively small number of price adjustments.

3.6.1 Problem Class and Performance Measures

For our simulations, we focus on a linear demand problem class given by $\mathcal{P} = [\sqrt{2}/3, \sqrt{3}/2]$, $\mathcal{Z} = [\sqrt{2}/2, \sqrt{3}/2] \times [1/2, 3/4]$ and $d(p; \mathbf{z}) = z_1 - p\, z_2$. When implementing the policies on this class, both the DOUBLING and MLE-C policies depend on a choice of two exploration prices, which we fix to be $p_{min} = \sqrt{2}/3$ and $p_{max} = \sqrt{3}/2$. The CVP policy depends on a parameter that determines the

size of the price perturbation, and we set this parameter based on the suggestion contained in den Boer and Zwart (2010b).

To evaluate the performance of each policy, we generate random samples of the problem parameters $\mathbf{z}^1, \ldots, \mathbf{z}^m \in \mathcal{Z}$ (to be described later), and we consider two measure of average performance over this random sample of parameters. First, we consider the **Percentage Revenue Loss**, which is defined to be the average over the random sample of problem instances of the cumulative regret divided by the total optimal revenue. Thus, if $\mathbf{z}^1, \ldots, \mathbf{z}^m \in \mathcal{Z}$ is the sample of problem parameters, we have

$$\textbf{Percentage Revenue Loss}\,(T) \;=\; \frac{1}{m}\sum_{i=1}^{m} \frac{\sum_{s=1}^{t} r(p^*(\mathbf{z}^i);\mathbf{z}^i) - r(P_s^i;\mathbf{z}^i)}{t \,\cdot\, r(p^*(\mathbf{z}^i);\mathbf{z}^i)} \times 100\% \,.$$

Equivalently, this quantity describes the total amount of revenue lost by each policy with respect to the optimal policy, as a percentage of the total optimal revenue. We also consider the **Price Changes** performance measure, which is the average over the m problem instances of the number of price changes made by each policy up to time T.

3.6.2 First Simulation: The Lower Bound Distribution

For our first experiment, we investigate the performance of all four pricing policies when problem parameters are drawn from a distribution similar to the one used to prove the lower bound of Theorem 3.2.1. We generate 100 independent random samples $(\mathbf{z}^1, \ldots, \mathbf{z}^{100})$, by drawing independent random values z^i in the interval $[1/2, 3/4]$ according to the density $\lambda(z) = 8\left\{\cos(4\pi(x - 5/8))\right\}^2$, and then setting $\mathbf{z}^i = (\sqrt{z^i}, z^i)$.

In all experiments, the standard error of the figures reported in the **Percentage**

Table 3.1: Comparison of the Percentage Revenue Loss of Four Policies on the Lower Bound Instance

	Percentage Revenue Loss			
$T \times 10^3$	FP	DOUBLING	MLE-C	CVP
1	3.73 %	1.33 %	1.28 %	1.29 %
2	3.73 %	1.08 %	1.09 %	1.14 %
3	3.73 %	1.00 %	1.01 %	1.08 %
4	3.73 %	0.92 %	0.94 %	1.02 %
5	3.73 %	0.93 %	0.90 %	0.95 %

Table 3.2: Comparison of the Switching Performance of Four Policies on the Lower Bound Instance

	Price Changes			
$T \times 10^3$	FP	DOUBLING	MLE-C	CVP
1	0.0	27.0	129.0	987.8
2	0.0	30.0	183.0	1987.8
3	0.0	33.0	226.0	2987.8
4	0.0	33.0	261.0	3986.8
5	0.0	36.0	294.0	4986.8

Revenue Loss columns is less than 0.1% for all policies at all reported values of T. In all experiments, the standard error of the figures reported in the **Price**

Changes columns is 0.0 for the DOUBLING and MLE-C policies, and is less than 33 for the CVP policy, for all reported values of T.

We observe that the three competing heuristics all significantly outperform the fixed-price policy, and lose only a small percentage of the total optimal revenue. More importantly, we see that the percentage revenue loss of the three heuristic policies all decreases with the number of time steps, while the fixed-price policy obviously has a percentage revenue loss that does not improve with time. We note that the MLE-C policy and the CVP policy have average revenue loss that essentially the same as that of the DOUBLING policy, even though these policies switch many more times on average. As an extreme example, we note that at time $T = 5000$, the average percentage revenue loss of the CVP policy is within 0.02% of that of the DOUBLING policy, despite the CVP policy making over 4900 more price adjustments on average. This behavior is consistent with the insights gained in Section 3.4.2, which establish that rate-optimal pricing is possible with only $\mathcal{O}(\log T)$ price adjustments.

3.6.3 Second Simulation: General Distributions

A key result of this chapter is the $\Omega(\sqrt{T})$ regret lower bound of Section 3.2, and this regret lower bound utilized problem instances drawn from a specially chosen distribution. Thus, a natural question is whether the lower bound construction of Section 3.2 is pathological, and whether policies can perform significantly better on a more "natural" set of problem instances. To address this issue, we simulated the four policies on two alternative distributions of problem parameters, described below.

1. **Uniform**: We generate 100 independent $\mathbf{z}^i$ by drawing 100 independent samples z_1^i from the uniform distribution on $[\sqrt{2}/2, \sqrt{3}/2]$, and 100 independent random samples z_2^i from the uniform distribution on $[1/2, 3/4]$, and setting $\mathbf{z}^i = (z_1^i, z_2^i)$.

2. **Gaussian**: We generate 100 independent $\mathbf{z}^i$ by drawing 100 independent samples z_1^i from a normal distribution whose mean is the midpoint of $[\sqrt{2}/2, \sqrt{3}/2]$, and whose variance is the width of $[\sqrt{2}/2, \sqrt{3}/2]$, truncating so that all samples lie in the interval. We generate 100 independent random samples z_2^i for the interval $[1/2, 3/4]$ in a similar fashion, and set $\mathbf{z}^i = (z_1^i, z_2^i)$.

Table 3.3: Comparison of the Percentage Revenue Loss of Four Policies on the Uniform Instance

	Percentage Revenue Loss			
$T \times 10^3$	FP	DOUBLING	MLE-C	CVP
1	4.50 %	1.47 %	1.30 %	1.01 %
2	4.50 %	1.21 %	1.01 %	0.78 %
3	4.50 %	1.08 %	0.89 %	0.68 %
4	4.50 %	0.92 %	0.82 %	0.61 %
5	4.50 %	0.94 %	0.76 %	0.57 %

We see that in Tables 3.4 and 3.6, the performance of the four pricing policies follows the same trend observed in Table 3.2 for the lower bound problem instance.

Table 3.4: Comparison of the Switching Performance of Four Policies on the Uniform Instance

	Price Changes			
$T \times 10^3$	FP	DOUBLING	MLE-C	CVP
1	0.0	27.0	129.0	929.7
2	0.0	30.0	183.0	1914.5
3	0.0	33.0	226.0	2905.9
4	0.0	33.0	261.0	3904.9
5	0.0	36.0	294.0	4904.9

Table 3.5: Comparison of the Percentage Revenue Loss of Four Policies on the Gaussian Instance

	Percentage Revenue Loss			
$T \times 10^3$	FP	DOUBLING	MLE-C	CVP
1	4.07 %	1.23 %	1.21 %	1.16 %
2	4.07 %	0.99 %	0.98 %	0.92 %
3	4.07 %	0.88 %	0.87 %	0.81 %
4	4.07 %	0.75 %	0.80 %	0.75 %
5	4.07 %	0.79 %	0.76 %	0.70 %

This suggests that the problem distribution considered in regret lower bound of Section 3.2 is not a pathological special case, and that for a number of natural

Table 3.6: Comparison of the Switching Performance of Four Policies on the Gaussian Instance

	Price Changes			
$T \times 10^3$	FP	DOUBLING	MLE-C	CVP
1	0.0	27.0	129.0	966.3
2	0.0	30.0	183.0	1957.4
3	0.0	33.0	225.0	2946.7
4	0.0	33.0	261.0	3939.5
5	0.0	36.0	294.0	4936.7

heuristic policies, we can expect to observe similar performance across a wide range of problem distributions. Moreover, we see that the DOUBLING policy performs well against the competing heuristics in both simulations, achieving comparable pricing performance with a relatively small number of price adjustments.

3.7. Discussion

We considered the problem of pricing under demand uncertainty in the presence of price-adjustment constraints. We established that under a general parametric notion of demand uncertainty, the rate of regret of the optimal pricing strategy is $\Theta(\sqrt{T})$, and that any policy that adjusts its prices according to a pre-specified schedule must switch prices at least $\Omega(\log T)$ times to achieve this rate of regret. We also considered a special "well-separated" case of the dynamic pricing problem, in which the optimal rate of regret is $\Theta(\log T)$. We showed that in this scenario,

$\Omega(\log T)$ price adjustments are still necessary for rate-optimal pricing. We constructed pricing policies to achieve these rates, and showed empirically that these policies perform well with respect to alternative heuristics.

Chapter 4

Dynamic Resource Allocation

4.1. Introduction

In this chapter, we depart from the question of how a seller should *price* his goods, and consider the question of how a seller might best *distribute* his goods for sale, among a number of possible selling venues. We address this problem in the context of a general resource allocation problem, in which a decision maker possesses a fixed number of identical units of a resource, and must allocate his units of resource across a fixed number of venues, with the goal of maximizing the total units of resource consumed. As a concrete example, consider a newspaper vendor with a fixed stock of newspapers to sell. The vendor must sell the newspapers from a fixed set of locations (newsstands and newspaper vending machines), and wishes to maximize the total number of newspapers sold across all locations. At the beginning of each day, the vendor must decide how many copies of the paper to send to each venue. At the end of the day, the vendor may observe the number of newspapers purchased from each venue, which represents the minimum of the

demand for newspapers at that venue, and the number of newspapers allocated to that venue at the beginning of the day. Then, the vendor may take this information into account when deciding the allocation of newspapers to venues for the following day.

In our example, if the vendor had perfect information about the demand at each venue, then computing the optimal allocation each day would be straightforward (as we will see in detail in Section 4.5). However, if the vendor has some degree of uncertainty about the demand at each venue, then the resource allocation problem becomes more challenging. This will be the primary focus of our work: an online version of the resource allocation problem, in which a decision maker must allocate units of a product across the available venues *without* foreknowledge of the demand. We consider strategies that offer a sequence of allocations over multiple time periods, using observed consumption from previous time periods to improve future allocation decisions. As in Chapters 2 and 3, these policies must carefully balance exploration of the venues with best-guess optimal allocation, and we will see that with the appropriate strategy, a policy can achieve nearly the same long-run average performance as a policy which has fully information about the demands in advance.

As our main contribution, we describe a policy for the online resource allocation problem with independent and identically distributed demands, whose worst-case regret is nearly rate-optimal. To our knowledge, ours is the first work to describe a regret-optimal policy for the online resource allocation problem under any assumptions about the demands, advancing the state of the art for this problem (Agarwal et al. (2010) and Ganchev et al. (2010)). We also demonstrate that our policy performs well in numerical experiments, by evaluating them both on synthetic de-

mand distributions, and on a variety of demand distributions coming from usage data from a real-world vehicle-sharing network.

We now proceed to define our model primitives and notation for the single-period resource allocation problem.

4.2. The Resource Allocation Problem

For some positive integer n, suppose that there are n *locations* (or *venues*), which we denote by the integers $i = 1, 2, \ldots, n$. Let m be a positive integer specifying the number of available units of resource to be allocated, and suppose that associated with each location, there is a demand random variable D_i taking values in $\mathbb{Z}_+$ with an arbitrary distribution, and a capacity $c_i \in \mathbb{Z}_+$. Consider the single period optimization problem, which we refer to as the *Resource Allocation Problem* (RAP).

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} \mathbb{E}[\min\{D_i, x_i\}] \\ \text{s.t.} \quad & \sum_{i=1}^{n} x_i = m, \quad x_i \in \{0, 1, \ldots, c_i\} \ \forall \ i \ , \end{aligned} \tag{4.1}$$

In the context of the RAP, the objective function of the optimization problem (4.1) quantifies the total expected number of resources consumed across all venues, when the demand for venue i is D_i, and when the number of units allocated to venue i is x_i. We wish to optimize this objective over all allocations $(x_1, x_2, \ldots, x_n)$, such that the total number of units allocated is m, and such that the allocation to each venue is a non-negative integer value which does not exceed the capacity of that venue. With this interpretation, we list some example application of the RAP.

Example 4.2.1 (Revenue Maximization)**.** Suppose a retailer has m identical units

of a product to sell across n possible venues during a fixed selling period. Assume that the retailer charges a uniform price p for each unit of good across all venues, and suppose that the demand at venue i during the selling period is given by D_i. Suppose further that the goods are perishable, and retain no salvage value after the end of the selling period. Then if the retailer chooses to offer $x_i \geq 0$ units of the good at each venue i, with $\sum_{i=1}^{n} x_i = m$, then the total expected revenue generated by the seller is $p \sum_{i=1}^{n} \mathbb{E}[\min\{D_i, x_i\}]$. Thus, the solution to the RAP is the allocation of goods to locations that will maximize the sellers total expected revenue during the selling period.

Example 4.2.2 (Vehicle-Sharing)**.** In a vehicle-sharing operation, subscribers pay a yearly fee to have access to a fleet of communal vehicles, usually located within a single city or community. The vehicles are based at some number of fixed locations around the community, and users may drive the vehicles for limited periods of time on a first-come, first-serve basis. The goal of the manager in this scenario is to choose an allocation of vehicles to locations which maximizes total usage. As in the previous example, the observed usage at location i with demand D_i and an allocation of x_i vehicles is $\min\{D_i, x_i\}$, and the total usage across the systems is $\sum_{i=1}^{n} \min\{D_i, x_i\}$, and thus this is a natural resource allocation problem. We discuss this scenario in greater detail in Section 4.8, in which we evaluate our resource allocation policies using usage data obtained from a real-world vehicle sharing operation.

4.3. Literature Review

The RAP described in this paper is a special case of the well-studied *Simple Allocation Problem* (Ibaraki and Katoh (1988)). In the general case, the reward functions

$\mathbb{E}[\min\{D_i, x_i\}]$ may be replaced with arbitrary functions $f_i(x_i)$, and one may also consider additional constraints on the allocations allowed. A full review of the literature on the single-period simple allocation problem is beyond the scope of this dissertation; we refer the reader to Ibaraki and Katoh (1988), Hochbaum (1994), and Bretthauer and Shetty (1995), and references therein for a broad overview. To our knowledge, none of the traditional literature on the simple allocation problem addresses the case in which the reward functions are stochastic and unknown to the decision maker in advance, which is the primary focus of this work.

To our knowledge, the first work to analyze the ORAP is Ganchev et al. (2010). Here, the authors consider essentially the same problem as we consider here, under the name of "The Dark Pools Problem." The authors describe policies based on a modified Kaplan-Meier estimation technique, and demonstrate that with probability $1-\delta$, their policy converges to an allocation whose reward is at least $1-\epsilon$ times the reward of the optimal allocation. The authors do not explicitly analyze the regret of their policy; however, as noted by Agarwal et al. (2010), and as demonstrated in our numerical section, low-regret policies such as ours can out-perform theirs over a variety of problem instances.

Subsequent to Ganchev et al. (2010), Agarwal et al. (2010) consider this problem from a worst-case standpoint, i.e. with no distributional assumptions on the demands. In this setting, they design algorithms that exploit the concavity of the reward function to achieve regret that is $\widetilde{\mathcal{O}}(T^{2/3})$ with respect to the single best fixed allocation in hindsight. The authors also present a lower bound of $\Omega(\sqrt{T})$, based on arguments from Auer et al. (2003) on lower bounds for the standard bandits problem. In this chapter, we present the IndexGreedy policy, which enjoys a regret guarantee of $\widetilde{\mathcal{O}}(\sqrt{T})$, nearly match the lower bound presented in Agarwal

et al. (2010). While IndexGreedy is, to our knowledge, the first policy to achieve such a guarantee, it is not a strict improvement upon the policy presented in Agarwal et al. (2010), due to the stochastic assumption that we make on the reward functions. However, our analysis provides optimal policies for the case of independent and identically distributed reward functions, and this is the first work to provide matching upper and lower bounds for the ORAP under any assumptions on the reward functions. We also demonstrate through numerical experiments that our policies perform well with respect to those of Ganchev et al. (2010) and Agarwal et al. (2010) over a wide range of problem instances. Finding optimal algorithms for the worst-case version of this problem is still an open question.

Closely related to the ORAP is the stochastic multi-armed bandit problem, studied in Lai and Robbins (1985b) and Auer et al. (2002b), and in particular, its extension to the case of multiple plays, studied in Anantharam et al. (2002). In fact, one may view problem of allocating m units of resource to nc possible spaces as a variant of the problem of playing the best m of nc arms in a stochastic multi-armed bandit problem, with the added condition that the m spaces chosen by the policy must constitute a feasible allocation. We will discuss this connection in detail in Section 4.5.

There is a trivial reduction from the ORAP to the standard stochastic multi-armed bandit problem, in which each arm corresponds to a feasible allocation of the m units of resource among the n locations. However, this would result in a bandit problem with $\binom{m+n-1}{n}$ arms, and ignores the structure of the reward function. Our goal will be to design policies whose performance scales much better with the parameters m and n.

4.4. Model and Notation

In this section, we describe the Online Resource Allocation Problem. Recall from Section 4.2 that for some positive integer n, there are n *venues*, which we denote by the integers $i = 1, 2, \ldots, n$. Associated with each venue is a *capacity* $c_i \in \mathbb{Z}_+$, which indicates the maximum number of units of resource which may be allocated to that venue in any given time period. We let m be a positive integer denoting the total number of available units of resource to be allocated across all venues in a given time period.

The ORAP proceeds in time periods $t = 1, 2, \ldots, T$, and we assume that for each location $i \in [n]$, there is a sequence of independent and identically distributed demand random variables $\{D_i^t : t \geq 1\}$ taking values in $\mathbb{Z}_+$. We let $X_i^t \in \mathbb{Z}_+$ denote the number of units allocated to venue i in time period t, and we let $\mathbf{X}^t = (X_1^t, \ldots, X_n^t)$ denote the *allocation* in the t^{th} period. We say that an allocation $\mathbf{X}^t$ is *feasible* if, with probability one, $\sum_{i=1}^n X_i^t = m$, and $X_i^t \leq c_i$ for all $i \in [n]$. For a given allocation $\mathbf{X}^t$, we denote the corresponding observed rewards by $\mathbf{Y}^t = (Y_1^t, \ldots, Y_n^t)$, where $Y_i^t = \min\{D_i^t, X_i^t\}$. A *policy* ψ for the ORAP is a sequence of functions $\psi_1, \psi_2, \ldots$ such that ψ_t computes a feasible allocation $\mathbf{X}^t$ to offer in time period t, based only on the information $\{(\mathbf{X}^\ell, \mathbf{Y}^\ell) : \ell < t\}$.

For convenience of notation, for any positive integer k, we denote by $[k]$ the set of integers $\{1, \ldots, k\}$. For $i \in [n]$ and $s \in \mathbb{Z}_+$, we let $F(i, s) = \Pr\{D_i \geq s\}$ denote the tail probability of the demand D_i. An *instance* $\mathcal{I}$ of the ORAP is a tuple $\mathcal{I} = (n, m, c_1, \ldots, c_n, F(1, \cdot), \ldots, F(n, \cdot))$. For a location $i \in [n]$ and an integer $s \in [c_i]$, we will refer to the ordered pair (i, s) as a *space* associated with location i. We will say that a unit of resource has been allocated to space (i, s) under

allocation $(x_1, \ldots, x_n)$ if and only if $x_i \geq s$.

We measure the performance of a policy in terms of its *regret*, which is the difference between the reward generated by the policy, and the reward generated by an omniscient policy that knows the demand distributions in advance, and always offers an optimal allocation. Formally, the T-period cumulative regret of a policy ψ on a problem instance $\mathcal{I}$ is given by

$$\text{Regret}(\mathcal{I}, \psi, T) = \sum_{t=1}^{T} \sum_{i=1}^{n} \min\{D_i^t, x_i^*\} - \min\{D_i^t, X_i^t\} \ ,$$

where $(x_1^*, \ldots, x_n^*)$ denotes an optimal allocation for problem instance $\mathcal{I}$, and where $(X_1^t, \ldots, X_n^t)$ denotes the allocation chosen by the policy ψ at time t. Note that, in contrast to the previous two chapters, we define the regret in this section to be a *random variable*, rather than an expected value. We do this for convenience of notation, since the results in the following sections will establish bounds on the regret that hold with high probability. For the sake of simplicity, in the remainder of this work, we will consider the case when $c_i = c$ for all $i = 1, \ldots, n$; that is, when all location can accept at most c units of resource. It is straightforward to extend the analysis to the general case.

4.5. A Policy for the Online Resource Allocation Problem

To motive our policy for the ORAP, we will first consider the single-period, "offline" version of this problem, in which the distributions of the demands are known to the decision maker in advance. In this case, it is known that the following greedy policy computes an optimal allocation to the single-period problem (4.1) (see, for example, Ganchev et al. (2010)).

Policy Greedy$(F(1,\cdot),\ldots,F(n,\cdot))$

Inputs: A Resource Allocation Problem Instance, and tail probabilities $F(1,\cdot),\ldots,F(n,\cdot)$.
Initialization: Set $x_i = 0$ for $i \in [n]$.
Description: For $j = 1,\ldots,m$

- Set
$$i^* \in \arg\max_{k:x_k+1\leq c}\{F(k,x_k+1)\}$$
to be the index with the largest marginal increase in reward, given the current allocations x_k, and given the feasibility constraint c. (Ties are broken arbitrarily).
- Set $x_{i^*} \leftarrow x_{i^*}+1$, that is, allocate one additional unit of resource to location i^*.

The Greedy policy is appropriately named: since we have $F(k,x_k+1) = \Pr\{D_k \geq x_k+1\} = \mathbb{E}[\min\{D_k, x_k+1\}] - \mathbb{E}[\min\{D_k,x_k\}]$, the Greedy policy builds an allocation unit-by-unit, at each step allocating a single unit of resource to the space that will give the largest marginal increase in the objective function, given the current allocations and feasibility constraints. Let us now note two observations about the Greedy policy that will be useful in our analysis.

Observation 1. The Greedy policy for the single-period resource allocation problem is equivalent to the following policy: allocate one unit of resource to each of the m spaces (i,s) with the largest tail probabilities $\{F(i,s) : i \in [n], s \in [c]\}$,

This observation follows directly from the definition of the Greedy policy, and the fact that the tail probabilities $F(i,s)$ are non-increasing in s for all $i \in [n]$. Note that by this same fact, allocating to the m spaces (i,s) with the largest tail probabilities $\{F(i,s) : i \in [n], s \in [c]\}$ will always result in a feasible allocation.

Observation 2. If $(i_1,s_1),\ldots,(i_m,s_m)$ are the indices of the m spaces with the

largest tail probabilities $\{F(i,s) : i \in [n], s \in [c]\}$, then the reward achieved by the GREEDY policy is $\sum_{j=1}^{m} F(i_j, s_j)$.

This follows from the observation that in general, if we allocate x_i units to location i, then the total reward from this location is

$$\mathbb{E}[\min\{D_i, x_i\}] = \sum_{s=1}^{\infty} \Pr\{\min\{D_i, x_i\} \geq s\} = \sum_{s=1}^{x_i} \Pr\{D_i \geq s\} = \sum_{s=1}^{x_i} F(i,s).$$

To design a policy for the online version of the problem, in which the distributions of the demands are not known to the decision maker in advance, we might consider implementing the Greedy policy at each time step, with inputs $\{\widehat{F}(i,\cdot) : i \in [n]\}$ representing some estimates of the tail probabilities computed from previous observations. This is in fact the approach we will take in designing our policy; however, great care must be taken in choosing the estimates $\{\widehat{F}(i,\cdot) : i \in [n]\}$, as these values will determine the behavior of the policy.

To motivate our eventual choice of the estimates $\{\widehat{F}(i,\cdot) : i \in [n]\}$, let us first consider a natural choice for these values: the empirical estimates of the tail probabilities, computed from past observations. While simple, this choice would be poor for two primary reasons. First, such an approach would lead to the potential under-exploration of spaces. Indeed, if the empirical estimate $\widehat{F}(i,s)$ of the tail probability of a space (i,s) is computed from a small number of samples, then this estimate could potentially be far less that the true tail probability $F(i,s)$. Under the greedy policy, this low estimate could prevent the space (i,s) from receiving any further allocation, and thus the estimate $\widehat{F}(i,s)$ would never improve. If the space (i,s) happened to have a large associated tail probability, then the greedy policy would suffer large regret, due to its failure to allocate to (i,s) a sufficient number of times.

A second issue with this choice of the values $\{\widehat{F}(i,\cdot) : i \in [n]\}$ is that it ignores important structural information available to the decision maker. Indeed, the fact that the tail probabilities $F(i,s)$ are non-increasing in s is a crucial assumption for the optimality of the Greedy policy. Clearly, taking $\widehat{F}(i,s)$ to be the empirical estimate of the tail probability associate with each space (i,s) will result in estimates that are not non-increasing in s, which does not reflect the underlying structure of the true probabilities $F(i,s)$.

This line of reasoning suggests that the values $\{\widehat{F}(i,\cdot) : i \in [n]\}$ should posses the following properties. First, to prevent under-exploration, the estimates $\widehat{F}(i,s)$ should depend not only on the empirical tail probabilities, but also on the number of samples used to compute the estimate. We should choose values $\widehat{F}(i,s)$ that will be large if the space (i,s) is under-sampled, and will be small only if the true value $F(i,s)$ is small, *and* the space has been well-sampled. Secondly, the values $\{\widehat{F}(i,\cdot) : i \in [n]\}$ should reflect the non-increasing structure of the true tail probabilities. Combining these ideas leads to the IndexGreedy policy, which we describe below.

4.5.1 The IndexGreedy Policy

Below, we present the IndexGreedy policy for the ORAP, motivated by the discussion in the previous section. The idea of the policy is the following. In each time period, compute an index for each space $\{(i,s) : i \in [n], s \in [c]\}$, based on past observations, and then compute a greedy allocation based on these indices. Our choice of the index functions is based on standard upper confidence interval techniques (Auer et al. (2002b)), and these index functions are constructed to exploit the non-increasing structure of the reward functions.

Policy $\mathsf{IndexGreedy}(T)$

Inputs: A Resource Allocation Problem instance, and a time horizon T.
Outputs: A sequence of feasible allocations $\{(X_1^t, \ldots, X_n^t) : t \in [T]\}$.
Initialization: Set $\widehat{F}(i,s) = 0$ for $i \in [n]$, $s \in [c]$. Define $w(n,t) = 2\sqrt{(\log t)/n}$, and let $w(0,t) = 1$. **Description:** For $t = 1, 2, \ldots,$

- For each $(i,s) \in [n] \times [c]$, define an index

$$G^t(i,s) = \min\{\widehat{F}^t(i,r) + w(N^t(i,r), T) : r \in [s]\}.$$

- Perform a greedy allocation with respect to the values $G^t(i,s)$, that is, set

$$(X_1^t, \ldots, X_n^t) \leftarrow \mathsf{Greedy}(G^t(1,\cdot), \ldots, G^t(n,\cdot)).$$

- For each $i \in [n]$, observe responses $Y_i^t = \min\{D_i^t, X_i^t\}$.
- For each $(i,s) \in [n] \times [c]$, update

$$\mathcal{N}^t(i,s) = \{\ell \in [t] : X_i^\ell \geq s\} \qquad \text{and} \qquad N^t(i,s) = \left|\mathcal{N}^t(i,s)\right|.$$

For each $(i,s) \in [n]$ such that $N^t(i,s) > 0$, set

$$\widehat{F}^t(i,s) = \frac{1}{N^t(i,s)} \sum_{\ell \in \mathcal{N}^t(i,s)} \mathbf{1}\left\{Y_i^\ell \geq s\right\}.$$

In the description of the policy, the indices $G^t(i,s)$ correspond to the values $\widehat{F}(i,s)$ in the previous discussion. Note that the indices $G^t(i,s)$ are non-increasing in s by definition, and that they promote exploration of under-sampled spaces using standard confidence interval techniques (Auer et al. (2002b)).

Connection to the Stochastic Bandits Problem: In light of Observations 1 and 2, one may think of our problem in terms of the classical stochastic bandits problem with multiple plays (Anantharam et al. (2002)). To see this, consider the case when $c = 1$, that is, when each location can hold at most one unit of resource. For this case, in each time period, the policy must choose a subset of m out of n possible locations, and allocate a single unit of resource to each location in the

chosen subset. The observed reward for allocating one unit of resource to location i will be $\min\{D_i, 1\} = \mathbf{1}\{D_i \geq 1\}$, and thus the total expected reward generated by allocating to locations $i_1, \ldots, i_m$ will be $\sum_{\ell=1}^{m} \mathbb{E}[\min\{D_{i_\ell}, 1\}] = \sum_{\ell=1}^{m} F(i_\ell, 1)$. Thus, in this case, the problem is exactly the stochastic multi-armed bandit problem with multiple plays (Anantharam et al. (2002)), in which there are n arms whose rewards are Bernoulli($F(i, 1)$), and where in each time period, a policy must play m distinct arms, with the goal of playing the m arms with the highest expected rewards as frequently as possible.

Now let us consider the general case, when $c \geq 1$. Given the above discussion, we could view this problem as a stochastic bandits problem with nc arms corresponding to the spaces $\{(i, s) : i \in [n], s \in [c]\}$. In this case, there is no straightforward reduction from the ORAP to the stochastic bandits problem with multiple plays, because in the ORAP, the policy is not free play an arbitrary subset of m of the nc spaces. Instead, the policy is constrained to play only those subsets of m spaces which constitute a feasible allocation. As a consequence, the policy cannot observe the reward associated with allocating to space (i, s) unless the policy also allocates to all spaces $\{(i, r) : r \leq s\}$. On the other hand, the policy has additional information about the structure of the reward functions in the ORAP that is not present in the standard stochastic bandits case, namely, that the mean rewards $F(i, s)$ are non-increasing in s for each location $i \in [n]$. Despite these differences, this line of thinking, and the algorithms presented in Anantharam et al. (2002) for the bandits problem with multiple plays, serve as motivation in the design of our policies.

We now proceed to prove a regret upper bound for our policy.

4.6. Regret Upper Bound for the Policy

The goal of this section is to prove the following theorem.

Theorem 4.6.1. *For any problem instance $\mathcal{I}$ and any $T \geq 2$, with probability at least $1 - ncT^{-2}$, we have*

$$\mathrm{Regret}(\mathcal{I}, \mathsf{IndexGreedy}(T), T) \leq 8ncm\sqrt{T \log T}.$$

Before proceeding with the proof of the theorem, we will first state three lemmas that will be useful in the analysis. The first lemma shows that with high probability, the estimates $\widehat{F}^t(i,s)$ are always close to their mean values.

Lemma 4.6.2. *Let $w(n,t) = 2\sqrt{(\log t)/n}$. With probability at least $1 - ncT^{-2}$, we have that for any $T \geq 2$,*

$$\left|\widehat{F}^t(i,s) - F(i,s)\right| \leq w(N^t(i,s), T) \quad \forall\ i \in [n],\ s \in [c],\ t \in [T]. \tag{4.2}$$

Proof. This follows from a result of Garivier and Moulines (2008), which gives a Hoeffding-type concentration inequality for sample means involving a random number of summands. The proof is given in Appendix C.1. □

Let us denote by $\mathcal{E}$ the event that condition (4.2) holds. In light of Lemma 4.6.2, it suffices to show that the regret of the IndexGreedy policy is bounded above by $8ncm\sqrt{T \log T}$ on the event $\mathcal{E}$, and so we will restrict our attention to this event for the remainder of the analysis. The next lemma shows that the indices $G^t(i,s)$ maintained by the IndexGreedy policy always exceed the true tail probabilities $F(i,s)$, with high probability.

Lemma 4.6.3. *On the event $\mathcal{E}$, we have that $F(i,s) \leq G^t(i,s)$ for all $i \in [n]$, all $s \in [c]$, and all $t \in [T]$.*

Proof. Fix an arbitrary i, s, t. From the structure of the reward function, we have that $F(i,s) \leq F(i,r)$ for all $1 \leq r \leq s$. From Lemma 4.6.2, we have that $F(i,r) \leq \widehat{F}^t(i,r) + w(N^t(i,r),T)$ for all $1 \leq r \leq s$. It follows that $F(i,s) \leq \widehat{F}^t(i,r) + w(N^t(i,r),T)$ for all $1 \leq r \leq s$, and so

$$F(i,s) \leq \min\{\widehat{F}^t(i,r) + w(N^t(i,r),T) : r \in [s]\} = G^t(i,s).$$

□

The last lemma is an observation about the structure of the IndexGreedy policy.

Lemma 4.6.4. *The allocation of the* IndexGreedy *policy corresponds to allocating the largest m spaces according to $G^t(i,s)$.*

Proof. This follows directly from the definition of the IndexGreedy policy, and the fact that $G^t(i,s)$ is non-increasing in s for each fixed $i \in [n]$. □

We now move on to bound the regret of the policy. Recall that we refer to a pair (i,s) as a *space*, and the tail probability $F(i,s)$ as the *reward* of that space. To simplify notation, let us index the spaces $(i_1,s_1),\ldots,(i_{nc},s_{nc})$ in decreasing order of their rewards. For $j < k$, let $\Delta_{i,k} = F(i_j,s_j) - F(i_k,s_k) > 0$. Let $\mathcal{A}_t \subset [nc]$ denote the indices of the m spaces chosen by the IndexGreedy policy at time t; i.e. at time t, the policy allocates to spaces $\{(i_j,s_j) : j \in \mathcal{A}_t\}$. With this notation, we know by Observation 2 that the regret of the policy is given by

$$\sum_{t=1}^{T} \left(\sum_{\ell=1}^{m} F(i_\ell,s_\ell) - \sum_{j \in \mathcal{A}_t} F(i_j,s_j) \right).$$

To aid in the analysis of the regret, we will now define two auxiliary quantities. First, let $I_t : [m] \to \mathcal{A}_t$ be an ordering of the elements of $\mathcal{A}_t$ in decreasing order of

their indices $G^t(i,s)$. In other words, $I_t : [m] \to \mathcal{A}_t$ is a bijection with

$$G^t(i_{I_t(1)}, s_{I_t(1)}) \geq G^t(i_{I_t(2)}, s_{I_t(2)}) \geq \ldots \geq G^t(i_{I_t(m)}, s_{I_t(m)}).$$

Observe now that the regret of the policy may be written as

$$\sum_{t=1}^{T} \sum_{\ell=1}^{m} F(i_\ell, s_\ell) - F(i_{I_t(\ell)}, s_{I_t(\ell)}).$$

Now, we define a random variable $H^T(\ell, j)$ in terms of the orderings I_t as follows: for each $\ell \in [m]$ and each $j \in [nc]$, let

$$H^T(\ell, j) = \sum_{t=1}^{T} \mathbf{1}\left\{j = I_t(\ell)\right\},$$

Note that $H^T(\ell, j)$ counts the total number of time steps up to time T in which space j was played by the policy, *and* in which space j had the ℓ^{th} highest index among the spaces played by the policy in that time step.

The desired bound on the regret of the policy now follows from two lemmas. In Lemma 4.6.5, we prove a bound on the regret in terms of the random variables $H^T(\ell, j)$. Then, in Lemma 4.6.6, we prove an upper bound on each $H^T(\ell, j)$. Finally, we combine these two results to prove Theorem 4.6.1.

Lemma 4.6.5. *On the event $\mathcal{E}$, the regret of the* IndexGreedy *policy is bounded above by*

$$\sum_{\ell=1}^{m} \sum_{j=\ell+1}^{nc} \Delta_{\ell,j} H^T(\ell, j)$$

Proof. We have for the regret

$$\begin{aligned}\sum_{t=1}^{T}\sum_{\ell=1}^{m}(F(i_\ell,s_\ell)-F(i_{I_t(\ell)},s_{I_t(\ell)})) &= \sum_{t=1}^{T}\sum_{\ell=1}^{m}\sum_{j=1}^{nc}(F(i_\ell,s_\ell)-F(i_j,s_j))\mathbf{1}\left\{j=I_t(\ell)\right\}\\ &= \sum_{\ell=1}^{m}\sum_{j=1}^{nc}\sum_{t=1}^{T}(F(i_\ell,s_\ell)-F(i_j,s_j))\mathbf{1}\left\{j=I_t(\ell)\right\}\\ &= \sum_{\ell=1}^{m}\sum_{j=1}^{nc}(F(i_\ell,s_\ell)-F(i_j,s_j))H^T(\ell,j)\\ &\leq \sum_{\ell=1}^{m}\sum_{j=\ell+1}^{nc}(F(i_\ell,s_\ell)-F(i_j,s_j))H^T(\ell,j),\end{aligned}$$

where the third line follows from the fact that $H^T(\ell,j)=\sum_{t=1}^{T}\mathbf{1}\left\{j=I_t(\ell)\right\}$ by definition, and where the last inequality follows from the fact that $F(i_\ell,s_\ell)-F(i_j,s_j)\leq 0$ if $\ell\geq j$. Since we have $F(i_\ell,s_\ell)-F(i_j,s_j)=\Delta_{\ell,j}$ by definition, this proves the result. □

In the following lemma, we prove an upper bound on the $H^T(\ell,j)$ random variables. Combining this result with Lemma 4.6.5 will prove the theorem.

Lemma 4.6.6. *Given the orderings I_t of the elements of $\mathcal{A}_t$ in decreasing order of their upper confidence bounds $\{G^t(i_j,s_j):j\in\mathcal{A}_t\}$, then on the event $\mathcal{E}$, we have that*

$$H^T(\ell,j)\leq\min\left\{64\frac{\log T}{\Delta_{\ell,j}^2},T\right\},$$

for any $\ell\in[m]$, and any $j>\ell$.

The idea of the proof is to show that if we ever have $H^{t_0}(\ell,j)=64\frac{\log T}{\Delta_{\ell,j}^2}$ for some $t_0<T$, then for all $t\geq t_0$, there will be at least ℓ spaces whose indices will be larger than the index of (i_j,s_j) for all $t\geq t_0$. In this case, since the IndexGreedy policy allocates the m spaces with the largest index, and since the ordering I_t orders the spaces allocated by the policy in decreasing order of their index, we will have that

$H^t(\ell, j) = H^{t_0}(\ell, j)$ for all $t \geq t_0$, i.e. that $H^t(\ell, j)$ will never increase beyond its value at t_0.

Proof. Suppose by way of contradiction that for all sufficiently large t, $H^t(i_j, s_j) > 64 \log T/(\Delta^2_{\ell,j})$. Let t_0 denote the time step in which $H^{t_0}(i_j, s_j) = 64 \log T/(\Delta^2_{\ell,j})$. In this case, since we have the trivial inequality $N^{t_0}(i_j, s_j) \geq H^{t_0}(i_j, s_j)$, we know that $N^{t_0}(i_j, s_j) \geq 64 \log T/(\Delta^2_{\ell,j})$. By Lemma 4.6.2, we that for all t, $\widehat{F}^t(i_j, s_j) - F(i_j, s_j) \leq w(N^t(i_j, s_j), T)$, and so it follows that $\widehat{F}^t(i_j, s_j) + w(N^t(i_j, s_j), T) \leq F(i_j, s_j) + 2w(N^t(i_j, s_j), T)$. These facts together imply that for all $t \geq t_0$,

$$\begin{aligned} G^t(i_j, s_j) &= \min\{\widehat{F}^t(i_j, r) + w(N^t(i_j, r), T) : r \in [s_j]\} \\ &\leq \widehat{F}^t(i_j, s_j) + w(N^t(i_j, s_j), T) \\ &\leq F(i_j, s_j) + 2w(N^t(i_j, s_j), T) \\ &= F(i_j, s_j) + 4\sqrt{\frac{\log T}{N^t(i_j, s_j)}} \\ &\leq F(i_j, s_j) + \Delta_{\ell,j}/2 \ , \end{aligned}$$

where the first line follows from the definition of $G^t(i, s)$, the third line follows from the above implication, the fourth line follows from the definition of $w(n, t)$, and the last line follows from the assumption that $N^t(i_j, s_j) \geq 64 \log T/(\Delta^2_{\ell,j})$ for all $t \geq t_0$. On the other hand, on event $\mathcal{E}$, we have for all t

$$F(i_j, s_j) + \Delta_{\ell,j}/2 \ < \ F(i_j, s_j) + \Delta_{\ell,j} \ = \ F(i_\ell, s_\ell) \ \leq \ G^t(i_\ell, s_\ell),$$

where the equality follows from the definition of $\Delta_{\ell,j}$, and where the last inequality follows from Lemma 4.6.3. It follows from these estimates that for all $t \geq t_0$,

$$G^t(i_j, s_j) \ < \ G^t(i_\ell, s_\ell).$$

But note that since ℓ was arbitrary, and since $\Delta_{\ell',j} \geq \Delta_{\ell,j}$ for $\ell' \leq \ell$, we actually have the following result: for any $\ell \in [m]$ and any $j > \ell$, if $N^{t_0}(i_j, s_j) =$

$64 \log T/(\Delta_{\ell,j}^2)$, then

$$G^t(i_j, s_j) \;<\; G^t(i_{\ell'}, s_{\ell'}) \qquad \forall\; 1 \le \ell' \le \ell \quad \forall\; t \ge t_0.$$

But by the definition of the IndexGreedy policy, we know that if this ever occurs, then we will have $H^t(\ell, j) = H^{t_0}(\ell, j)$ for all $t \ge t_0$, i.e. $H^t(\ell, j)$ will never increase beyond $H^{t_0}(\ell, j)$, since there will be at least ℓ spaces whose indices $G^t(i_\ell, s_\ell)$ will be strictly larger than the index $G^t(i_j, s_j)$ of (i_j, s_j) for all $t \ge t_0$. □

Now using Lemma 4.6.6 and the above analysis, we can prove the main result.

Proof of Theorem 4.6.1. Using Lemma 4.6.6 and the above analysis, we have that on the event $\mathcal{E}$, the regret of the IndexGreedy policy is bounded above by

$$\begin{aligned}
\sum_{\ell=1}^{m} \sum_{j=\ell+1}^{nc} (F(i_\ell, s_\ell) - F(i_j, s_j)) H^t(\ell, j) &\le \sum_{\ell=1}^{m} \sum_{j=\ell+1}^{nc} \Delta_{\ell,j} \min\left\{ 64 \frac{\log T}{\Delta_{\ell,j}^2}, T \right\} \\
&= \sum_{\ell=1}^{m} \sum_{j=\ell+1}^{nc} \min\left\{ 64 \frac{\log T}{\Delta_{\ell,j}}, \Delta_{\ell,j} T \right\} \\
&\le \sum_{\ell=1}^{m} \sum_{j=\ell+1}^{nc} \sqrt{64 T \log(T)} \\
&\le 8ncm\sqrt{T \log T}.
\end{aligned}$$

□

Remark 4.6.7. In Agarwal et al. (2010), the authors give a lower bound for this problem (in the case where $m = c$) of $\Omega\left(\max\left\{\sqrt{mnT}, m\sqrt{\log(n)T}\right\}\right)$. The corresponding upper bound achieved by IndexGreedy in this case is $\mathcal{O}(m^2 n\sqrt{T \log T})$, and thus, our regret bound is tight up to a factor of $m^{2/3} n^{1/2} \sqrt{\log T}$. Additionally, IndexGreedy is the first policy to achieve a regret bound of $\widetilde{\mathcal{O}}(\text{poly}(m, n, c)\sqrt{T})$ for the online resource allocation problem.

4.7. IndexGreedy-B

In this section, we describe a variant of the IndexGreedy policy, which we call IndexGreedy-B, which shows superior performance in numerical simulations. To motivate the design of IndexGreedy-B, let us first recall the behavior of the original IndexGreedy policy. The IndexGreedy policy performs a greedy allocation of m units in each time step, with respect to an index G^t. Recall that for every space (i, s), the index $G^t(i, s)$ contained a term that forced the index to be large in the event that the space (i, s) had not been well-sampled. Thus, we can think of the IndexGreedy policy as "exploring" with all m units in each time period, since all m units are allocated with respect to the index G^t which promotes allocation to under-sampled spaces.

The IndexGreedy-B policy operates as follows. In each time period, IndexGreedy-B allocates $m-1$ of its units according to a purely greedy index S^t, where $S^t(i, s)$ is simply the empirical estimate of the tail probability of space (i, s) in time period t. Then, the policy allocates the remaining single unit of resource according to the index G^t, which promotes allocation to under-sampled spaces. Thus, we can think of IndexGreedy-B as "exploiting" with $m-1$ units of resource, and "exploring" with only a single unit of resource, whereas the original IndexGreedy policy explores with all m units in every time period.

A detailed description of IndexGreedy-B is given below, and we evaluate its performance in Section .

Policy IndexGreedyB(T)

Inputs: A Simple Allocation Problem instance, and a time horizon T.
Outputs: A sequence of feasible allocations $\{(X_1^t, \ldots, X_n^t) : t \in [T]\}$.
Initialization: Set $\widehat{F}(i,s) = 0$ for $i \in [n]$, $s \in [c]$. Define $w(n,t) = 2\sqrt{(\log t)/n}$, and let $w(0,t) = 1$.
Description: For $t = 1, 2, \ldots,$

- For each $(i,s) \in [n] \times [c]$, define indices

$$G^t(i,s) = \min\{\widehat{F}^t(i,r) + w(N^t(i,r), T) : r \in [s]\} \qquad S^t(i,s) = \min\{\widehat{F}^t(i,r) : r \in [s]\}.$$

- Perform a greedy allocation of $m-1$ units with respect to the values $S^t(i,s)$, that is, set

$$(X_1^t, \ldots, X_n^t) \leftarrow \mathsf{Greedy}(S^t(1,\cdot), \ldots, S^t(n,\cdot); m-1).$$

- Allocate the remaining unit of resource to location i^* satisfying

$$i^* \in \arg\max_{i \in [n]} G^t(i, X_i^t + 1).$$

- For each $i \in [n]$, observe responses $Y_i^t = \min\{D_i^t, X_i^t\}$.
- For each $(i,s) \in [n] \times [c]$, update

$$\mathcal{N}^t(i,s) = \{\ell \in [t] : X_i^\ell \geq s\} \qquad \text{and} \qquad N^t(i,s) = \left|\mathcal{N}^t(i,s)\right|.$$

  For each $(i,s) \in [n]$ such that $N^t(i,s) > 0$, set

$$\widehat{F}^t(i,s) = \frac{1}{N^t(i,s)} \sum_{\ell \in \mathcal{N}^t(i,s)} \mathbf{1}\left\{Y_i^\ell \geq s\right\}.$$

4.8. Numerical Experiments

In this section, we compare the empirical performance of the IndexGreedy policy against that of several other heuristics, which we list below.

1. Ganchev: This is Algorithm 2 of Ganchev et al. (2010), discussed in Section 4.1. This policy makes the same stochastic assumptions on the demand

that are made in this paper, and is similar to the IndexGreedy policy, in that it performs a greedy allocation in each time step with respect to some index computed from past samples. The index used by the GANCHEV policy is a Kaplan-Meier estimator, modified to promote exploration of under-explored venues.

2. AGARWAL: This is Algorithm 2 of Agarwal et al. (2010), discussed in Section 4.1. This policy makes no distributional assumptions on the sequence of demand values, and instead, exploits the concavity of the reward function to achieve a worst-case regret guarantee of $\widetilde{\mathcal{O}}(T^{2/3})$ against the best single fixed allocation in hindsight.
3. INDEXGREEDYB: This is the variant of INDEXGREEDY described in the previous section. INDEXGREEDYB maintains two indices for each space: the upper-confidence index of IndexGreedy, and also a sample-mean based index. At any given time period, IndexGreedyB allocates $m-1$ units greedily according to the sample-mean based index, and then allocates the remaining unit of resource according to the upper-confidence based index. Thus, at a high level, INDEXGREEDYB performs less exploration than INDEXGREEDY, and we will see that this modification significantly improves performance.

We compare the performance of these policies in two experiments, one involving synthetically generated demand values, and the other involving usage data from a vehicle-sharing network.

4.8.1 Synthetic Demand

In our first experiment, we evaluate the performance of the three policies on simulated demand random variables. For the first experiment, we set the number

of locations $n = 5$, the capacity at each location $c = 10$, and the total number of units of resource to be $m = 10$. We fix a time horizon $T = 2000$, and then generate an ensemble of 100 problem instances in the following way. For $i = 1, \ldots, 100$, we generate a parameter vector $(\lambda_1^i, \ldots, \lambda_n^i)$ by selecting each value λ_j^i uniformly at random from the interval $[0, 10]$. We then define demand distributions $D_i \sim \text{Poisson}(\lambda_i)$, and we execute the three policies on the ensemble of problem instances. We measure the performance of the policies in this section in terms of the T-period percentage optimal reward, which we define to be

$$T \mapsto \frac{1}{m} \sum_{\ell=1}^{m} \frac{\sum_{t=1}^{T} \sum_{i=1}^{n} \mathbb{E}[\min\{D_i^{\ell,t}, X_i^t\}]}{T \cdot \sum_{i=1}^{n} \mathbb{E}[\min\{D_i^{\ell,t}, x_i^*\}]} \times 100\% \ .$$

In Table 4.1, we compare the Percentage Optimal Reward for each of the policies for several values of T. The standard error is less than 0.2% for all reported values. Note that while the percentage optimal reward of all policies tends towards 100%,

Table 4.1: Comparison of the Percentage Optimal Reward of the heuristics over the ensemble.

	Percentage Optimal Reward			
$T \times 10^2$	Ganchev	Agarwal	IndexGreedy	IndexGreedy-B
4	96.2 %	94.4 %	98.2 %	99.2 %
8	96.3 %	96.5 %	98.9 %	99.5 %
12	96.3 %	97.3 %	99.2 %	99.7 %
16	96.3 %	97.7 %	99.3 %	99.7 %
20	96.4 %	98.0 %	99.4 %	99.8 %

the IndexGreedy and IndexGreedy-B policies outperform the other two heuristics on

this ensemble of problem instances, with IndexGreedy-B showing the best performance out of all the heuristics. We also note that the gap in performance between IndexGreedy and IndexGreedy-B is the largest in the earlier time periods, which we attribute to the greater amount of exploration performed by IndexGreedy with respect to IndexGreedy-B. Specifically, IndexGreedy uses a relatively large number of allocations in the early time periods to explore sub-optimal but under-sampled spaces, while IndexGreedy-B avoids this issue with a less aggressive exploration rule.

In the following section, we evaluate the performance of the heuristics using usage data from a real-world vehicle-sharing operation.

4.8.2 Demand from a Vehicle-Sharing Network

In this section, we evaluate the empirical performance of the policies on data from a vehicle-sharing network. We first describe the operation of the vehicle-sharing operation in greater detail.

Vehicle-Sharing Operation: Here we summarize the logistics of the vehicle-sharing operation from which we get our usage data. The car-sharing company operates a fleet of 13 communal vehicles, with each vehicle assigned to a specific physical location within the city. Users gain access to the vehicles by paying a yearly subscription fee, plus a usage fee based on hours and milage. To use a vehicle, customers place a reservation for a specific vehicle / location through an online reservation system. The user must return the vehicle to its original location at the end of their reservation period. If a total of x_i vehicles are available at location i during a given time period, and if the demand for vehicles at location i is D_i, then the vehicle sharing operation observes a total usage of $\min\{D_i, x_i\}$. The goal of the vehicle sharing operation is to find an allocation of vehicles to locations

that will maximize total usage, and thus, this is naturally modeled as a resource allocation problem.

Usage Data and Model Fitting: Since we are unable to actually implement our policies in a real world scenario, we instead will use usage data to calibrate a demand model, and then will simulate our policies on this model. We obtained a year's worth of usage data from the above vehicle-sharing operation. Each record in the data set specifies an instance in which a given vehicle was used, and contains the location of the vehicle, the time the reservation was placed, the time the vehicle was checked-out, and the time the vehicle was returned.

While the logistics describe by this usage data are not fully captured by our model, we note several observations that allow us to simplify this usage data. First, we note that nearly all of the reservations are made within two hours of the usage time of the vehicle. Secondly, nearly all of the trips are short in duration (under six hours). Based on these observations, we fit a model to the vehicle-usage data set in the following way. First, we filter the data to include only trips with reservation lead time less than two hours, and total trip duration less the six hours. Then, we divide the time-line into two-hour buckets, and define the usage for a particular location and time bucket to be the number of vehicles that were in use at that location during that time bucket. This gives us a derived data set of censored demand information for each location, and we fit Poisson demand random variables to this data using MLE for censored observations.

To describe the specifics of the MLE, consider a fixed location, and for a time period t, let $x^t \in \mathbb{Z}_+$ denote the number of vehicles allocated to the location during time period t. Let $U^t \in \mathbb{Z}_+$ denote the observed usage at the location during time period t. We make the assumption that the true demand at the location during

each time period is given by a demand random variable $D^t \sim \text{Poisson}(\lambda)$, so that $U_t = \min\{D^t, x^t\}$. The likelihood of observing a sequence of usages $U_1, U_2, \ldots, U_T$ is given by

$$L(\lambda) = \prod_{t=1}^{T} \Pr_P\{U^t; \lambda\}^{\mathbf{1}\{U^t < x^t\}} F_P(x^t; \lambda)^{\mathbf{1}\{U^t \geq x^t\}}.$$

where $\Pr_P(\cdot; \lambda)$ denotes the Poisson probability mass function with real parameter $\lambda > 0$, and $F_P(x; \lambda) = \sum_{i=x}^{\infty} \Pr_P(i; \lambda)$. For each venue, we use numerical methods to find a maximizer of the corresponding likelihood function over the set $\{\lambda \in \mathbb{R} : \lambda > 0\}$.

In Table 4.2, we summarize the Poisson demand model fitted to the derived vehicle usage data. For each location, the capacity indicates the number of vehicles which were made available at that location by the vehicle-sharing operation. (In contrast to our online policies, the existing vehicle-sharing operation kept the number of vehicles available at a given location fixed for the entire year). The parameter λ indicates the rate of the Poisson demand random variable fitted to the censored data, as described above. In other words, if λ_i is the fitted parameter for location i, then the demand for vehicles at location i during a given time period is given by a Poisson random variable with parameter λ_i.

Table 4.2: Summary of fitted demand model.

Summary of Fitted Demand Model

Location:	1	2	3	4	5	6	7	8	9
Capacity:	3	2	2	1	1	1	1	1	1
λ:	0.349	0.358	0.518	0.159	0.157	0.163	0.221	0.088	0.125

We execute the three policies on the Poisson demand random variables fitted to the vehicle-sharing data. To understand how the policies will perform under an increased load on the system, we also simulate the heuristics on Poisson demand with parameters that are 3 times and 5 times the original fitted parameter values. To make the simulation more realistic, we constrain the policies to adjust their allocation only once per week, rather than once for every two-hour time block. In the tables below, we listed the Percentage Optimal Reward of the heuristics. The standard error in all figures reported is less than 0.2%.

Table 4.3: Comparison of the Percentage Optimal Reward of the heuristics for the fitted values of λ.

	Percentage Optimal Reward			
$T \times 10^2$	Ganchev	Agarwal	IndexGreedy	IndexGreedy-B
5	87.0 %	94.6 %	85.5 %	99.2 %
10	93.2 %	94.9 %	91.6 %	99.3 %
15	95.3 %	95.0 %	94.1 %	99.4 %
20	96.3 %	95.0 %	95.5 %	99.4 %
25	96.9 %	95.1 %	96.3 %	99.4 %
30	97.4 %	95.2 %	96.9 %	99.5 %

First, we note that in all three simulations, IndexGreedy-B outperforms all other heuristics at all time periods, which is consistent with the simulation of the previous section. We note that on this particular problem instance, the Ganchev policy

Table 4.4: Comparison of the Percentage Optimal Reward of the heuristics for 3× the fitted values of λ.

	Percentage Optimal Reward			
$T \times 10^2$	Ganchev	Agarwal	IndexGreedy	IndexGreedy-B
5	85.3 %	89.0 %	81.1 %	98.1 %
10	92.5 %	89.6 %	90.4 %	98.8 %
15	94.5 %	89.8 %	93.6 %	99.2 %
20	96.2 %	90.1 %	95.2 %	99.4 %
25	96.9 %	90.2 %	96.2 %	99.5 %
30	97.4 %	90.3 %	96.8 %	99.6 %

slightly outperforms IndexGreedy; this is in contrast to the results on the synthetic data, which showed that when averaged over a large number of problem instances, the performance of IndexGreedy is better. Finally, we note that in nearly all cases, the performance of a particular heuristic at a particular time decreases as the load of the systems increases. Intuitively, as the demand for vehicles increases, there is a greater difference in reward between a poor allocation and an optimal allocation. In other words, when demand is high, there is a greater opportunity cost for offering a sub-optimal allocation.

4.8.3 A Second Fitted Demand Model

In the previous section, we adopted a simple approach to fitting a demand model to the vehicle-sharing usage data, which preserved the i.i.d. assumptions on the

Table 4.5: Comparison of the Percentage Optimal Reward of the heuristics for 5× the fitted values of λ.

	Percentage Optimal Reward			
$T \times 10^2$	Ganchev	Agarwal	IndexGreedy	IndexGreedy-B
5	86.1 %	86.0 %	79.3 %	98.4 %
10	93.1 %	87.2 %	89.5 %	99.0 %
15	95.4 %	87.6 %	93.0 %	99.3 %
20	96.5 %	87.9 %	94.7 %	99.5 %
25	97.2 %	88.1 %	95.8 %	99.6 %
30	97.7 %	88.4 %	96.5 %	99.6 %

demand that are made in this chapter. While the model derived from this approach accurately reflects the amount of usage of each vehicle, it does not capture the dependence between time periods of vehicle usage. Indeed, since vehicles are used in contiguous blocks of time, then the usage in a given time period is dependent on usage in previous time periods. This effect was mitigated in the previous section by filtering out longer trips and bucketing the time periods into large blocks; however, it leaves open the question of how the heuristics would perform on more realistic demand sequences.

In this section, we adopt an alternative stochastic model for vehicle demand, which does *not* satisfy the i.i.d. assumption on demand, and which better reflects the usage patters of vehicles observed in the data set. During a given time period t, at a given location i, we assume that a random number D_i^t of customers wish

to begin a trip using a vehicle stationed at location i, where $D_i^t \sim \text{Poisson}(\lambda_i)$ for some parameter λ_i. We assume that each of the D_i^t customers wishes to use the vehicle for a specific length of time, and this length is also random with an arbitrary distribution. We can easily simulate vehicle demand using this model: in each time period, we draw a random number of customers D_i^t, and for each customer, we draw a random trip duration, and these data together contribute to the demand for vehicles at the given location from time period t, to some future time period at which all of the trips beginning in time period t have ended.

To fit the parameters λ_i, we count for each time period the number of trips *started* in that time period (as opposed to the number of vehicles in use during that time period, as done in the previous section), and assume this number is a random variable with distribution $\text{Poisson}(\lambda_i)$. If there are no vehicles available at a given location and time, we are unable to observe whether a customer wishes to begin a trip, so we ignore time periods in which a location has no available vehicles for the purposes of estimation. (The proportion of time periods in which no vehicles are available at a given location is relatively small, and so the estimates produced in this way still provide a reasonable approximation to the number of customer arrivals). To generate a random trip length for a given location, we simply draw a length from the empirical distribution of trip lengths observed in the data.

In Table 4.6, we plot the total cumulative reward of the four heuristics, averaged over 25 demand sequences randomly generated according to the model above. Note that in this simulation, one time period corresponds to a fifteen-minute segment, as we do not perform the two-hour bucketing done in the previous section. Here, one unit of reward corresponds to satisfying one unit of demand in a single time period. As in the previous section, all policies are constrained to updated their

allocation once per week, and the standard error in all figures reported is less that 80.

Table 4.6: Comparison of the Average Cumulative Reward of the heuristics for the simulated demand sequence.

	Average Cumulative Reward			
$T \times 10^3$	Ganchev	Agarwal	IndexGreedy	IndexGreedy-B
1	1076	1577	1071	1645
3	4224	4827	3927	4965
5	7462	7997	7134	8213
7	10744	11201	10397	11492
9	14119	14482	13777	14870

We note that, although the demand in this model is not i.i.d across time periods, the IndexGreedy-B policy still outperforms the other heuristics. We also see that the Agarwal policy performs nearly as well as the IndexGreedy-B policy in this simulation; this is to be expected, since the Agarwal policy is designed to perform well on demand sequences that do not satisfy the i.i.d. assumption. These results suggest that even for demand sequences that are not perfectly independent between time periods, we can still expect reasonably good performance from the index-type strategies considered in this chapter.

4.9. Discussion

In this chapter, we considered the online resource allocation problem, and introduced the IndexGreedy policy for this problem. We established a regret upper bound for IndexGreedy that matches a known regret lower bound for this problem, demonstrating that IndexGreedy has regret which is rate-optimal. To our knowledge, this is the first regret-optimal policy for the online resource allocation problem. We demonstrated that the IndexGreedy policy performs well in numerical simulations, both on synthetic data, and on a demand simulator calibrated on a real-world data set from a vehicle-sharing operation. We also described a variant of our policy, called IndexGreedyB, which exhibits significantly better performance in numerical experiments by performing more judicious exploration of venues. Designing policies with strong regret guarantees for the case of non-stationary demand distributions is a compelling open question.

Appendix A

Proofs from Chapter 2

A.1. Proofs from Section 2.3.1

The proof of Lemmas 2.3.3 and 2.3.4 will make use of the following properties of the problem class $\mathcal{C}$ define in the statement of Theorem 2.3.1.

Lemma A.1.1 (Properties of $\mathcal{C}$)**.** *For all $p \in \mathcal{P}$ and $z \in \mathcal{Z}$,*

1. $p^*(z) = (1 + 2z)/(4z)$
2. $p^*(z_0) = 1$ *for* $z_0 = 1/2$.
3. $d(p^*(z_0); z) = 1/2$ *for all* $z \in \mathcal{Z}$
4. $r(p^*(z); z) - r(p; z) \geq \frac{1}{3}(p^*(z) - p)^2$
5. $|p^*(z) - p^*(z_0)| \geq \frac{1}{4}|z - z_0|$
6. $|d(p; z) - d(p; z_0)| \leq |p^*(z_0) - p|\ |z - z_0|$

Proof. Property 1 follows from checking first and second order optimality conditions of the revenue function $r(p; z) = pd(p; z)$. Properties 2 and 3 follow by simple

calculations using the formulas for $p^*(z)$ and $d(p;z)$. Property 4 follows from the fact that $r'(p^*(z);z) = 0$ and $r''(p;z) = -2z \leq -2/3$ for all $(p,z) \in \mathcal{P} \times \mathcal{Z}$. Property 5 follows from an application of the Mean Value Theorem, and the fact that $\frac{d}{dz}p^*(z) = -1/(4z^2) \leq -1/4$ for all $z \in \mathcal{Z}$. Finally, Property 6 follows from the calculation

$$\begin{aligned}|d(p;z) - d(p;z_0)| = |1/2 + z - pz - 1/2 - z_0 + pz_0| &= |z - z_0| \cdot |1 - p| \\ &= |z - z_0| \cdot |p^*(z_0) - p|\end{aligned}$$

since $p^*(z_0) = 1$ by construction. □

The proof of Lemma 2.3.3 also makes use of the following standard results, which gives an upper bound on the KL-divergence between two Bernoulli distributions.

Lemma A.1.2 (Corollary 3.1 in Taneja and Kumar (2004))**.** *Suppose B_1 and B_2 are distributions of Bernoulli random variables with parameters q_1 and q_2, respectively, with $q_1, q_2 \in (0,1)$. Then*

$$\mathcal{K}(B_1; B_2) \leq \frac{(q_1 - q_2)^2}{q_2(1 - q_2)}.$$

A.1.1 Proof of Lemma 2.3.3

Consider a policy ψ setting prices in $\mathcal{P} = [3/4, 5/4]$ and some $s \geq 1$. To prove the lemma, we appeal to the Chain Rule for KL divergence (Theorem 2.5.3, Cover and Thomas (1999)), which states that

$$\mathcal{K}(Q_t^{\psi,z_0}; Q_t^{\psi,z}) = \sum_{s=1}^{t} \mathcal{K}(Q_s^{\psi,z_0}; Q_s^{\psi,z} | \mathbf{Y}_{s-1}),$$

where each term in the sum is the conditional KL divergence, defined as

$$\mathcal{K}(Q_s^{\psi,z_0}; Q_s^{\psi,z} | \mathbf{Y}_{s-1}) \;=\; \sum_{\mathbf{y}_s \in \{0,1\}^s} Q_s^{\psi,z_0}(\mathbf{y}_s) \log\left(\frac{Q_s^{\psi,z_0}(y_s|\mathbf{y}_{s-1})}{Q_s^{\psi,z}(y_s|\mathbf{y}_{s-1})}\right).$$

In light of this fact, we may prove the inequality of the lemma as follows. First, show that the conditional KL divergence in each time period is bounded above by the instantaneous regret in that time period (times some additional terms), and then apply the Chain Rule to show that the total KL divergence is bounded above by the cumulative regret (times additional terms).

To proceed along these lines, let $p_s = \psi(\mathbf{y}_{s-1})$. We have

$$\begin{aligned}
& \mathcal{K}(Q_s^{\psi,z_0}; Q_s^{\psi,z} | \mathbf{Y}_{s-1}) \\
= & \sum_{\mathbf{y}_s \in \{0,1\}^s} Q_s^{\psi,z_0}(\mathbf{y}_s) \log\left(\frac{Q_s^{\psi,z_0}(y_s|\mathbf{y}_{s-1})}{Q_s^{\psi,z}(y_s|\mathbf{y}_{s-1})}\right) \\
= & \sum_{\mathbf{y}_{s-1} \in \{0,1\}^{s-1}} Q_{s-1}^{\psi,z_0}(\mathbf{y}_{s-1}) \sum_{y_s \in \{0,1\}} Q_s^{\psi,z_0}(y_s|\mathbf{y}_{s-1}) \log\left(\frac{Q_s^{\psi,z_0}(y_s|\mathbf{y}_{s-1})}{Q_s^{\psi,z}(y_s|\mathbf{y}_{s-1})}\right) \\
\leq & \frac{1}{d(p_s;z)\,(1-d(p_s;z))} \sum_{\mathbf{y}_s \in \{0,1\}^{s-1}} Q_{s-1}^{\psi,z_0}(\mathbf{y}_{s-1})\,(d(p_s;z_0) - d(p_s;z))^2 \ , \\
\leq & \frac{3}{16} \sum_{\mathbf{y}_s \in \{0,1\}^{s-1}} Q_{s-1}^{\psi,z_0}(\mathbf{y}_{s-1})\,(d(p_s;z_0) - d(p_s;z))^2 \ .
\end{aligned}$$

The first line follows from the definition of conditional KL divergence. The second line follows from an algebraic manipulation using the relation $Q_s^{\psi,z_0}(\mathbf{y}_s) = Q_s^{\psi,z_0}(y_s|\mathbf{y}_{s-1}) Q_s^{\psi,z_0}(\mathbf{y}_{s-1})$, and the fact that $Q_s^{\psi,z_0}(\mathbf{y}_{s-1}) = Q_{s-1}^{\psi,z_0}(\mathbf{y}_{s-1})$. The third line follows from Lemma A.1.2 and the fact that

$$Q_s^{\psi,z_0}(y_s|\mathbf{y}_{s-1}) = d(p_s;z_0)^{y_s}(1-d(p_s;z_0))^{1-y_s},$$

and the fourth line follows from the fact that $d(p;z) \in [1/4, 3/4]$ for all $p \in \mathcal{P}$ and $z \in \mathcal{Z}$.

By Property 6 in Lemma A.1.1, we have that $(d(p_s;z_0) - d(p_s;z))^2 \leq (z_0 - z)^2\ (p^*(z_0) - p_s)^2$, which implies

$$\begin{aligned}
\mathcal{K}(Q_s^{\psi,z_0}; Q_s^{\psi,z} | \mathbf{Y}_{s-1}) & \leq \frac{3}{16}\,(z_0 - z)^2 \sum_{\mathbf{y}_s \in \{0,1\}^{s-1}} Q_{s-1}^{\psi,z_0}(\mathbf{y}_{s-1})\,(p^*(z_0) - p_s)^2 \\
& = \frac{3}{16}\,(z_0 - z)^2\, \mathbb{E}_{z_0}\left[(p^*(z_0) - P_s)^2\right] \ .
\end{aligned}$$

Summing over all s and using the Chain Rule for KL-divergence, we have that

$$\begin{aligned}\mathcal{K}(Q_t^{\psi,z_0}; Q_t^{\psi,z}) &= \sum_{s=1}^{t} \mathcal{K}(Q_s^{\psi,z_0}; Q_s^{\psi,z} | \mathbf{Y}_{s-1}) \;\leq\; \frac{3}{16}(z_0 - z)^2 \sum_{s=1}^{t} \mathbb{E}_{z_0}\left[(p^*(z_0) - P_s)^2\right] \\ &\leq \frac{9}{16}(z_0 - z)^2 \sum_{s=1}^{t} \mathbb{E}_{z_0}\left[r(p^*(z_0); z_0) - r(P_s; z_0)\right] \\ &\leq \frac{9}{16}(z_0 - z)^2 \,\mathrm{Regret}(z_0, \mathcal{C}, t, \psi)\ ,\end{aligned}$$

where the last inequality follows from Property 4 in Lemma A.1.1. This concludes the proof.

We now proceed to the proof of Lemma 2.3.4. The proof of this lemma uses the following standard result on the minimal error of a two-hypothesis test, which is derived from Theorem 2.2 of Tsybakov (2009).

Lemma A.1.3 (Theorem 2.2, Tsybakov (2009))**.** *Let Q_0 and Q_1 be two probability distributions on a finite space $\mathcal{Y}$, with $Q_0(y), Q_1(y) > 0$ for all $y \in \mathcal{Y}$. Then for any function $J : \mathcal{Y} \to \{0, 1\}$,*

$$Q_0\{J = 1\} + Q_1\{J = 0\} \geq \frac{1}{2} e^{-\mathcal{K}(Q_0; Q_1)},$$

where $\mathcal{K}(Q_0; Q_1)$ denotes the KL divergence of Q_0 and Q_1.

A.1.2 Proof of Lemma 2.3.4

Let $z_0 = 1/2$ be as in Lemma A.1.1, and fix a time horizon $T \geq 2$. Let $z_1 = z_0 + \frac{1}{4}T^{-1/4}$, and define two intervals $C_{z_0} \subset \mathcal{P}$ and $C_{z_1} \subset \mathcal{P}$ by

$$C_{z_0} = \left\{p : |p^*(z_0) - p| \leq \frac{1}{48T^{1/4}}\right\} \text{ and } C_{z_1} = \left\{p : |p^*(z_1) - p| \leq \frac{1}{48T^{1/4}}\right\}.$$

Note that C_{z_0} and C_{z_1} are disjoint, since Property 5 in Lemma A.1.1 gives that $|p^*(z_0) - p^*(z_1)| \geq \frac{1}{4}|z_0 - z_1| = \frac{1}{16T^{1/4}}$. It follows from Property 4 in Lemma A.1.1

that for each $z \in \{z_0, z_1\}$, if $p \in \mathcal{P} \setminus C_z$, then the instantaneous regret is at least $\frac{1}{3(48^2)\sqrt{T}}$ because

$$r(p^*(z); z) - r(p; z) \geq \frac{1}{3}\left(p - p^*(z)\right)^2 \geq \frac{1}{3(48)^2\sqrt{T}} = \frac{1}{3(48)^2 \cdot \sqrt{T}} \ .$$

Let $P_1, P_2, \ldots$ denote the sequence of prices under the policy ψ. Then,

$$\begin{aligned}
&\text{Regret}\left(z_0, \mathcal{C}, T, \psi\right) + \text{Regret}\left(z_1, \mathcal{C}, T, \psi\right) \\
&\geq \sum_{t=1}^{T-1} \mathbb{E}_{z_0}\left[r(p^*(z_0); z_0) - r(P_{t+1}; z_0)\right] + \mathbb{E}_{z_1}\left[r(p^*(z_1); z_1) - r(P_{t+1}; z_1)\right] \\
&\geq \frac{1}{3(48)^2 \cdot \sqrt{T}} \sum_{t=1}^{T-1} \Pr_{z_0}\left\{P_{t+1} \notin C_{z_0}\right\} + \Pr_{z_1}\left\{P_{t+1} \notin C_{z_1}\right\} \\
&\geq \frac{1}{3(48)^2 \cdot \sqrt{T}} \sum_{t=1}^{T-1} \Pr_{z_0}\left\{J_{t+1} = 1\right\} + \Pr_{z_1}\left\{J_{t+1} = 0\right\} \ ,
\end{aligned}$$

where for all $t \geq 1$, $J_{t+1} = \mathbf{1}\left\{P_{t+1} \in C_{z_1}\right\}$ is a binary random variable that takes the value of 1 when P_{t+1} is in C_{z_1}, and zero otherwise. The second inequality follows from the fact that when $J_{t+1} = 1$, we have $P_{t+1} \in C_{z_1} \subset \mathcal{P} \setminus C_{z_0}$, and thus $P_{t+1} \notin C_{z_0}$, so that $\Pr_{z_0}\left\{J_{t+1} = 1\right\} \leq \Pr_{z_0}\left\{P_{t+1} \notin C_{z_0}\right\}$. Now a standard result on the minimum error in a simple hypothesis test (Lemma A.1.3) implies that for all t,

$$\Pr_{z_0}\left\{J_{t+1} = 1\right\} + \Pr_{z_1}\left\{J_{t+1} = 0\right\} \geq \frac{1}{2} e^{-\mathcal{K}\left(Q_t^{\psi, z_0}; Q_t^{\psi, z_1}\right)}.$$

Now putting things together and summing over t, we have

$$\begin{aligned}
\text{Regret}(z_0, \mathcal{C}, T, \psi) + \text{Regret}(z_1, \mathcal{C}, T, \psi) &\geq \frac{1}{3(48)^2\sqrt{T}} \cdot \frac{1}{2} \sum_{t=1}^{T-1} e^{-\mathcal{K}\left(Q_t^{\psi, z_0}; Q_t^{\psi, z_1}\right)} \\
&\geq \frac{1}{3(48)^2\sqrt{T}} \cdot \frac{T-1}{2} e^{-\mathcal{K}\left(Q_T^{\psi, z_0}; Q_T^{\psi, z_1}\right)} \\
&\geq \frac{\sqrt{T}}{12(48^2)} e^{-\mathcal{K}\left(Q_T^{\psi, z_0}; Q_T^{\psi, z_1}\right)}.
\end{aligned}$$

where the second inequality follows from the standard fact that $\mathcal{K}\left(Q_t^{\psi, z_0}; Q_t^{\psi, z_1}\right)$ is non-decreasing in t (see, for example, Theorems 2.5.3 and 2.6.3 in Cover

and Thomas (1999)), and the third inequality follows from the fact that $(T-1)/(2\sqrt{T}) \geq \sqrt{T}/4$ for all $T \geq 2$. This completes the proof.

A.2. Proof of Lemma 2.3.7

The proof of Lemma 2.3.7 is a direct application of the following standard result on the finite-sample mean-squared error of a maximum-likelihood estimator.

Theorem A.2.1 (Tail Inequality for MLE based on IID Samples, Theorem 36.3 in Borovkov (1998))**.** *Let $\mathcal{Z} \subset \mathbb{R}^n$ be compact and convex, and let $\{Q^{\mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$ be family of distributions on a discrete sample space $\mathcal{Y}$ parameterized by $\mathcal{Z}$. Suppose Y is a random variable taking value in $\mathcal{Y}$ with distribution $Q^{\mathbf{z}}$, and the following conditions hold.*

(i) The family $\{Q^{\mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$ is identifiable.

(ii) For some $s > k$, $\sup_{\mathbf{z}\in\mathcal{Z}} \mathbb{E}_{\mathbf{z}} \left[\|\nabla \log Q^{\mathbf{z}}(Y)\|^s\right] = \gamma < \infty$.

(iii) The function $\mathbf{z} \mapsto \sqrt{Q^{\mathbf{z}}}$ is differentiable on $\mathcal{Z}$.

(iv) The Fisher information matrix*, whose $(i,j)^{\text{th}}$ entry is given by $\mathbb{E}_{\mathbf{z}}\left[-\frac{\partial^2}{\partial z_i \partial z_j} \log Q^{\mathbf{z}}(\mathbf{Y})\right]$, is positive definite.*

Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d. random variables taking value in $\mathcal{Y}$ with distribution $Q^{\mathbf{z}}$, and let $\widehat{\mathbf{Z}}(t) = \arg\max_{\mathbf{z}\in\mathcal{Z}} \prod_{\ell=1}^{t} Q^{\mathbf{z}}(Y_\ell)$ denote the maximum likelihood estimate based on t i.i.d. samples. Then, there exists a constants $\eta_1 > 0$ and $\eta_2 > 0$ depending only on s, k, $Q^{\mathbf{z}}$ and $\mathcal{Z}$ such that for any $t \geq 1$ and any $\epsilon \geq 0$,

$$\Pr_{\mathbf{z}}\left\{\left\|\widehat{\mathbf{Z}}(t) - \mathbf{z}\right\| \geq \epsilon\right\} \leq \eta_1 \, e^{-t\eta_2 \, \epsilon^2} \ .$$

To apply Theorem A.2.1 to our setting, we first check that the hypothesis hold

for the family $\{Q^{\bar{\mathbf{p}},\mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$ for the exploration prices $\bar{\mathbf{p}}$ satisfying Assumption 2. For any problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$, the parameter set $\mathcal{Z}$ is compact and convex, by assumption. Conditions (i) and (iv) hold by Assumption 2, so it is enough to check conditions (ii) and (iii). To verify condition (ii), recall that for any $\mathbf{y} \in \{0,1\}^k$,

$$Q^{\bar{\mathbf{p}},\mathbf{z}}(\mathbf{y}) = \prod_{\ell=1}^{k} d(\bar{p}_\ell; \mathbf{z})^{y_\ell} (1 - d(\bar{p}_\ell; \mathbf{z}))^{1-y_\ell},$$

where $d : \mathcal{P} \times \mathcal{Z} \to [d_{min}, d_{max}]$ is smooth, with $d_{min}, d_{max} \in (0,1)$. Thus, we have

$$\nabla \log Q^{\bar{\mathbf{p}},\mathbf{z}}(\mathbf{y}) = \nabla \sum_{\ell=1}^{k} \log Q^{\bar{p}_\ell,\mathbf{z}}(y_\ell) = \sum_{\ell=1}^{k} y_\ell \nabla \log d(\bar{p}_\ell; \mathbf{z}) + (1-y_\ell) \nabla \log(1 - d(\bar{p}_\ell; \mathbf{z})),$$

and it follows that

$$\left\| \nabla \log Q^{\bar{\mathbf{p}},\mathbf{z}}(\mathbf{y}) \right\| \leq \sum_{\ell=1}^{k} \left\| \nabla \log d(\bar{p}_\ell; \mathbf{z}) \right\| + \left\| \nabla \log(1 - d(\bar{p}_\ell; \mathbf{z})) \right\|.$$

Now since $d(\bar{p}; \cdot)$ is a smooth function that is bounded away from zero and one, we have that $\nabla \log d(\bar{p}_\ell; \mathbf{z})$ and $\nabla \log(1 - d(\bar{p}_\ell; \mathbf{z}))$ are smooth functions on the compact set $\mathcal{Z}$ for each ℓ, and it follows that there exists a constant $\bar{D}_3$ depending only on the problem instance $\mathcal{C}$ such that $\|\nabla \log Q^{\bar{\mathbf{p}},\mathbf{z}}(\mathbf{y})\| \leq \bar{D}_3$. It follows that with probability one, we have $\|\nabla \log Q^{\bar{\mathbf{p}},\mathbf{z}}(\mathbf{y})\|^s \leq \bar{D}^s$, which is the desired result.

To verify condition (iii), note that $Q^{\mathbf{p},\mathbf{z}}(\mathbf{y})$ is smooth on $\mathcal{P} \times \mathcal{Z}$, since $Q^{\mathbf{p},\mathbf{z}}(\mathbf{y})$ is a product of smooth functions on $\mathcal{P} \times \mathcal{Z}$. We also have that $Q^{\mathbf{p},\mathbf{z}}(\mathbf{y})$ is bounded away from zero, since $Q^{\mathbf{p},\mathbf{z}}(\mathbf{y}) \geq (d_{min})^k$, so it follows that $\mathbf{z} \mapsto \sqrt{Q^{\mathbf{p},\mathbf{z}}(\mathbf{y})}$ is differentiable on $\mathcal{Z}$ for any $\mathbf{p} \in \mathcal{P}^k$. Thus, we also have that $\mathbf{z} \mapsto \sqrt{Q^{\bar{\mathbf{p}},\mathbf{z}}(\mathbf{y})}$ is differentiable on $\mathcal{Z}$.

Now the result of Lemma 2.3.7 follows from a direct application of this theorem. Since the estimator $\widehat{\mathbf{Z}}(c)$ is formed from c i.i.d. samples, we have by Theorem A.2.1

$$\mathbb{E}_{\mathbf{z}}\left[\left\| \widehat{\mathbf{Z}}(c) - \mathbf{z} \right\|^2 \right] = \int_0^\infty \Pr_{\mathbf{z}}\left\{ \left\| \widehat{\mathbf{Z}}(c) - \mathbf{z} \right\|^2 \geq u \right\} du \leq \int_0^\infty \eta_1 e^{-c\eta 2u} \, du = \frac{\eta_1}{c\eta_2}.$$

Taking $C_{mle} = \eta_1/\eta_2$ proves the claim.

A.3. Proof of Lemma 2.4.6

The proof of Lemma 2.4.6 depends on van Trees' inequality, which we state below.

Lemma A.3.1 (van Trees' Inequality, Gill and Levit (1995))**.** *For a closed interval $\mathcal{Z} \subset \mathbb{R}$, let $\{Q^z : z \in \mathcal{Z}\}$ be a family of distributions on a discrete sample space $\mathcal{Y}$, and let Z be a random variable taking values in $\mathcal{Z}$ with density $\lambda : \mathcal{Z} \to \mathbb{R}_+$. Suppose that the following conditions hold:*

1. *For each $y \in \mathcal{Y}$, the function $z \mapsto Q^z(y)$ is absolutely continuous on $\mathcal{Z}$.*
2. *λ is absolutely continuous on $\mathcal{Z}$, and $\lambda \to 0$ at the endpoints of $\mathcal{Z}$.*
3. *$\mathbb{E}_z \left[\frac{d}{da} \log Q^z(Y)\right] = 0$*

where $\mathbb{E}_z$ denotes expectation of the random variable Y having the distribution Q^z. Then, for any smooth function $g : \mathcal{Z} \to \mathbb{R}$ and any function $\hat{g} : \mathcal{Y} \to \mathbb{R}$,

$$\mathbb{E}[(\hat{g}(Y) - g(Z))^2] \geq \frac{(\mathbb{E}[\frac{d}{dz} g(Z)])^2}{\mathbb{E}\left[\left(\frac{d}{dz} \log Q^Z(Y)\right)^2\right] + \mathbb{E}\left[\left(\frac{d}{da} \log \lambda(Z)\right)^2\right]}\,, \tag{A.1}$$

where $\mathbb{E}[\,\cdot\,]$ denotes the expectation with respect to the joint distribution of Q^z and λ.

To apply the above result to our setting, recall the problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ defined in Theorem 2.4.5, which has $\mathcal{P} = [1/3, 1/2]$, $\mathcal{Z} = [2, 3]$, and $d(p; z) = 1 - (pz)/2$. For any policy ψ setting prices in $\mathcal{P}$ and any $t \geq 1$, we define the sample space to be $\mathcal{Y} = \{0, 1\}^t$, and we consider the family of distributions $\left\{Q_t^{\psi,z} : z \in \mathcal{Z}\right\}$, where $Q_t^{\psi,z} : \{0, 1\}^t \to [0, 1]$ is the distribution of customer decisions induced by the policy ψ up to time t. That is,

$$Q_t^{\psi,z} = \prod_{\ell=1}^{t} (1 - (p_\ell z)/2)^{y_\ell} ((p_\ell z)/2)^{1-y_\ell}.$$

A convenient choice of the density $\lambda(z) : [2,3] \to \mathbb{R}_+$ is $\lambda(z) = 2\{\cos(\pi(z-5/2))\}^2$.

To check that the hypotheses of Lemma A.3.1 hold under these assumptions, note that Conditions 1 and 2 of Lemma A.3.1 follow immediately from our construction. Condition 3 is also satisfied because

$$\mathbb{E}_z\left[\frac{d}{dz}\log Q_t^{\psi,z}(\mathbf{Y}_t)\right] = \sum_{\mathbf{y}_t\in\{0,1\}^t}\left(\frac{\frac{d}{dz}Q_t^{\psi,z}(\mathbf{y}_t)}{Q_t^{\psi,z}(\mathbf{y}_t)}\right)Q_t^{\psi,z}(\mathbf{y}_t) = \frac{d}{dz}\sum_{\mathbf{y}_t\in\{0,1\}^t}Q_t^{\psi,z}(\mathbf{y}_t)$$
$$= \frac{d}{dz}(1) = 0\ .$$

By checking first and second order optimality conditions, it is straightforward to check that $p^*(z) = 1/z$, and so $p^*(z)$ is a smooth function of z on $\mathcal{Z}$. Therefore all of the conditions of Lemma A.3.1 are satisfied, and we can apply van Trees' Inequality to our problem.

To complete the proof, we will now compute the values on the right-hand side of van Trees' inequality (Equation A.1) for our specific problem. Since $p^*(z) = 1/z$, we have that $\frac{d}{dz}p^*(z) = 1/z^2 \geq 1/9$ for all $z \in \mathcal{Z}$. It follows that $(\mathbb{E}[\frac{d}{dz}p^*(z)])^2 \geq 1/81$. Recalling that $\lambda(z) = 2\{\cos(\pi(z-5/2))\}^2$, it is straightforward to compute that

$$\mathbb{E}\left[\left(\frac{d}{dz}\log\lambda(Z)\right)\right] = 8\pi^2\int_2^3\{\sin(\pi(z-5/2))\}^2\,dz = 4\pi^2.$$

Finally, for any $\mathbf{z} \in \mathcal{Z}$, we may compute that

$$\mathbb{E}_z\left[\left(\frac{d}{dz}\log Q_t^{\psi,z}(\mathbf{Y}_t)\right)\middle|\mathbf{Y}_{t-1} = \mathbf{y}_{t-1}\right] = \frac{p}{z(2-pz)} \leq \frac{(1/2)}{2(2-3/2)} = \frac{1}{2},$$

where the last inequality follows from the fact that $p \in \mathcal{P}$ and $z \in \mathcal{Z}$. Applying the Chain Rule for Fisher Information (Lemma A.5.2), we have

$$\mathbb{E}_z\left[\left(\frac{d}{dz}\log Q_t^{\psi,z}(\mathbf{Y}_t)\right)^2\right] \leq \frac{t}{2}.$$

Since $P_{t+1} = \psi_{t+1}(\mathbf{Y}_t)$, we may apply Lemmas A.3.1 to get

$$\mathbb{E}[(p^*(z) - P_{t+1})^2] \geq \frac{(1/81)}{4\pi^2 + t/2} \geq \frac{1}{81(4\pi^2+1/2)}\cdot\frac{1}{t} \geq \frac{1}{405\pi^2}\cdot\frac{1}{t},$$

which is the desired result.

A.4. Proof of Theorem 2.4.7

In contrast to the general case, MLE-GREEDY forms an estimate of the unknown parameter based on samples which are *not* i.i.d. Thus, we need to develop a new bound for our estimate. The proof is motivated by techniques for establishing finite-sample deviation inequalities for maximum likelihood estimators; see, for example, Borovkov (1998), Theorem 33.3.

The analysis depends on estimates of the *Hellinger distance*. For any $t \geq 1$ and $\mathbf{y}_{t-1} \in \{0,1\}^{t-1}$, we define the conditional Hellinger distance

$$H^{\mathcal{G}}(z, u|\mathbf{y}_{t-1}) = \sum_{y_t \in \{0,1\}} \left(\sqrt{Q_t^{\mathcal{G},z}(y_t|\mathbf{y}_{t-1})} - \sqrt{Q_t^{\mathcal{G},z+u}(y_t|\mathbf{y}_{t-1})} \right)^2 ,$$

for all pairs $z \in \mathcal{Z}$ and $u \in \mathcal{Z} - z$. Note that $Q_t^{\mathcal{G},z}(y_t|\mathbf{y}_{t-1})$ denotes the probability that $Y_t = y_t$ conditioned on the event that $\mathbf{Y}_{t-1} = \mathbf{y}_{t-1}$, when the policy $\mathcal{G}$ is used and the parameter is z.

Lemma A.4.1 (Hellinger Distance Lower Bound)**.** *There exists a constant c_H depending only on the problem class $\mathcal{C} = (\mathcal{P}, \mathcal{Z}, d)$ such that for any $t \geq 1$ and any $\mathbf{y}_{t-1} \in \{0,1\}^{t-1}$, and for all pairs $z \in \mathcal{Z}$ and $u \in \mathcal{Z} - z$,*

$$H^{\mathcal{G}}(z, u|\mathbf{y}_{t-1}) \geq c_H \cdot u^2.$$

Proof. By Corollary 4.3 of Taneja and Kumar (2004), we have the following lower bound on the conditional Hellinger distance in terms of the KL divergence.

$$H^{\mathcal{G}}(z, u|\mathbf{y}_{t-1}) \geq \frac{\sqrt{d_{min}}}{2} \mathcal{K}(Q_t^{\psi,z}(\,\cdot\,|\mathbf{y}_{t-1}); Q_t^{\psi,z+u}(\,\cdot\,|\mathbf{y}_{t-1})) = \frac{\sqrt{d_{min}}}{2} \mathcal{K}(Q^{p_t,z}; Q^{p_t,z+u}),$$

where $p_t = \psi_t(\mathbf{y}_{t-1})$. So, to prove the desired lower bound, it is enough to prove a quadratic lower bound on the function $u \mapsto \mathcal{K}(Q^{p_t,z}; Q^{p_t,z+u})$. To do this, first note that

$$\frac{\partial^2}{\partial u^2}\mathcal{K}(Q^{p_t,z}; Q^{p_t,z+u}) = \frac{\partial^2}{\partial u^2}\mathbb{E}_z\left[\log\left(\frac{Q^{p_t,z}(Y)}{Q^{p_t,z+u}(Y)}\right)\right] = \mathbb{E}_z\left[-\frac{\partial^2}{\partial u^2}\log Q^{p_t,z+u}(Y)\right],$$

and by Assumption 3, this term is bounded below by $c_f > 0$ for all $p_t \in \mathcal{P}$ and all $z \in \mathcal{Z}$. Also, we have that

$$\frac{\partial}{\partial u}\mathcal{K}(Q^{p_t,z}; Q^{p_t,z+u})\Big|_{u=0} = -\mathbb{E}_z\left[\frac{\partial}{\partial u}\log Q^{p_t,z+u}(Y)\Big|_{u=0}\right] = 0$$

by a straightforward calculation. It follows from a standard result that

$$\mathcal{K}(Q^{p_t,z}; Q^{p_t,z+u}) \geq \frac{c_f}{2}u^2$$

for all $u \in \mathcal{Z} - z$. Taking $c_H = c_f\sqrt{d_{min}}/4$ proves the claim.

□

For all pairs $z \in \mathcal{Z}$ and $u \in \mathcal{Z} - z$, let the likelihood ratio $X_t^{\mathcal{G},z}(u)$ and the conditional likelihood ratio $X_t^{\mathcal{G},z}(u\big|\mathbf{Y}_{t-1})$ be defined by

$$X_t^{\mathcal{G},z}(u) = \frac{Q_t^{\mathcal{G},z+u}(\mathbf{Y}_t)}{Q_t^{\mathcal{G},z}(\mathbf{Y}_t)} \quad \text{and} \quad X_t^{\mathcal{G},z}(u\big|\mathbf{Y}_{t-1}) = \frac{Q_t^{\mathcal{G},z+u}(Y_t \mid \mathbf{Y}_{t-1})}{Q_t^{\mathcal{G},z}(Y_t \mid \mathbf{Y}_{t-1})}\ .$$

The following lemma gives an upper bound on a moment of the likelihood ratio.

Lemma A.4.2 (Likelihood Ratio Moment Inequality)**.** *For all pairs $z \in \mathcal{Z}$ and $u \in \mathcal{Z} - z$, and $t \geq 1$, we have*

$$\mathbb{E}_z\left[\sqrt{X_t^{\mathcal{G},z}(u|\mathbf{Y}_{t-1})}\ \middle|\ \mathbf{Y}_{t-1}\right] \leq e^{-c_H u^2/2}\ ,$$

with probability one, and

$$\mathbb{E}_z\left[\sqrt{X_t^{\mathcal{G},z}(u)}\right] \leq e^{-c_H t u^2/2}.$$

Proof. To establish the first inequality, note that for all $\mathbf{y}_{t-1} \in \{0,1\}^{t-1}$,

$$\begin{aligned}
\mathbb{E}_z\left[\sqrt{X_t^{\mathcal{G},z}(u|\mathbf{Y}_{t-1})}\;\middle|\;\mathbf{Y}_{t-1}=\mathbf{y}_{t-1}\right] &= \sum_{y_t\in\{0,1\}}\sqrt{\frac{Q_t^{\mathcal{G},z+u}(y_t\mid\mathbf{y}_{t-1})}{Q_t^{\mathcal{G},z}(y_t\mid\mathbf{y}_{t-1})}}\cdot Q_t^{\mathcal{G},z}(y_t\mid\mathbf{y}_{t-1}) \\
&= \sum_{y_t\in\{0,1\}}\sqrt{Q_t^{\mathcal{G},z+u}(y_t\mid\mathbf{y}_{t-1})}\sqrt{Q_t^{\mathcal{G},z}(y_t\mid\mathbf{y}_{t-1})} \\
&= 1-\frac{H(z,u|\mathbf{y}_{t-1})}{2} \le e^{-H(z,u|\mathbf{y}_{t-1})/2} \le e^{-c_H u^2/2}\,,
\end{aligned}$$

which gives the desired result. Note that the last equality follows from the definition of $H(z,u|\mathbf{y}_{t-1})$ which shows that

$$\begin{aligned}
H^{\mathcal{G}}(z,u|\mathbf{y}_{t-1}) &= \sum_{y_t\in\{0,1\}}\left(\sqrt{Q_t^{\mathcal{G},z}(y_t|\mathbf{y}_{t-1})}-\sqrt{Q_t^{\mathcal{G},z+u}(y_t|\mathbf{y}_{t-1})}\right)^2 \\
&= 2\left(1-\sum_{y_t\in\{0,1\}}\sqrt{Q_t^{\mathcal{G},z}(y_t|\mathbf{y}_{t-1})}\sqrt{Q_t^{\mathcal{G},z+u}(y_t|\mathbf{y}_{t-1})}\;\right)
\end{aligned}$$

We will establish the second inequality of Lemma A.4.2 by induction on t. The case when $t=1$ follows immediately from the above calculation. So, assume the claim holds for $t-1$, that is,

$$\mathbb{E}_z\left[\sqrt{X_{t-1}^{\mathcal{G},z}(u)}\right] \le e^{-(t-1)c_H u^2/2}$$

Now, by definition, we have that

$$\begin{aligned}
\mathbb{E}_z\left[\sqrt{X_t^{\mathcal{G},z}(u)}\;\right] &= \mathbb{E}_z\left[\sqrt{X_{t-1}^{\mathcal{G},z}(u)}\cdot\sqrt{X_t^{\mathcal{G},z}(u|\mathbf{Y}_{t-1})}\;\right] \\
&= \mathbb{E}_z\left[\sqrt{X_{t-1}^{\mathcal{G},z}(u)}\cdot\mathbb{E}_z\left[\sqrt{X_t^{\mathcal{G},z}(u|\mathbf{Y}_{t-1})}\;\middle|\;\mathbf{Y}_{t-1}\right]\right] \\
&\le e^{-c_H u^2/2}\cdot\mathbb{E}_z\left[\sqrt{\mathbb{E}_z[X_{t-1}^{\mathcal{G},z}(u)]}\;\right] \le e^{-tc_H u^2/2}\,,
\end{aligned}$$

where the first inequality follows from the first part of Lemma A.4.2, and the final inequality follows from the inductive hypothesis. This completes the proof. □

Here is the proof of Theorem 2.4.7.

Proof. Consider an arbitrary $z \in \mathcal{Z}$. For all $u \in \mathcal{Z} - z$, let $L_t^{\mathcal{G},z}(u) = -\log X_t^{\mathcal{G},z}(u)$. By Assumption 4, $L_t^{\mathcal{G},z}(u)$ is globally convex in u. Moreover, it is easy to verify that $L_t^{\mathcal{G},z}(0) = 0$. It follows from the definition of $\widehat{Z}(t)$ that

$$\widehat{Z}(t) = \arg\max_{v \in \mathcal{Z}} Q_t^{\mathcal{G},v}(\mathbf{Y}_t) = z + \arg\max_{u \in \mathcal{Z}-z} X_t^{\mathcal{G},z}(u) = z + \arg\min_{u \in \mathcal{Z}-z} L_t^{\mathcal{G},z}(u)$$

Therefore, for any $\delta \in \mathcal{Z} - z$, if $|\widehat{Z}(t) - z| > |\delta|$, then the minimizer of $L_t^{\mathcal{G},z}(\,\cdot\,)$ must be outside the interval $[-\delta, \delta]$, which implies that either $L_t^{\mathcal{G},z}(\delta) \leq 0$ or $L_t^{\mathcal{G},z}(-\delta) \leq 0$. Hence, for any $\delta \in \mathcal{Z} - z$, we have that

$$\Pr_z\{|\widehat{Z}(t) - z| \geq |\delta|\} \leq \Pr_z\{L_t^{\mathcal{G},z}(\delta) \leq 0\} + \Pr_z\{L_t^{\mathcal{G},z}(-\delta) \leq 0\} \ .$$

By Markov's Inequality and Lemma A.4.2, it follows that

$$\begin{aligned} \Pr_z\{L_t^{\mathcal{G},z}(\delta) \leq 0\} &= \Pr_z\{X_t^{\mathcal{G},z}(\delta) \geq 1\} = \Pr_z\left\{\sqrt{X_t^{\mathcal{G},z}(\delta)} \ \geq \ 1\right\} \\ &\leq \mathbb{E}_z\left[\sqrt{X_t^{\mathcal{G},z}(\delta)}\right] \ \leq \ e^{-tc_H\delta^2/2} \ . \end{aligned}$$

A similar argument shows that $\Pr_z\{L_t^{\mathcal{G},z}(-\delta) < 0\} \leq e^{-tc_H\delta^2/2}$, which implies that for any $\delta \in \mathcal{Z} - z$,

$$\Pr_z\{|\widehat{Z}(t) - z| > |\delta|\} \leq 2\, e^{-tc_H\delta^2/2} \ .$$

Thus, for any $0 < \epsilon \leq \max\{|x| : x \in \mathcal{Z} - z\}$, we have that

$$\Pr_z\{|\widehat{Z}(t) - z| > \epsilon\} \leq 2e^{-tc_H\epsilon^2/2} \ .$$

On the other hand, if $\epsilon > \max\{|x| : x \in \mathcal{Z} - z\}$, then $\Pr_z\{|\widehat{Z}(t) - z| > \epsilon\} = 0$ by definition. This gives the desired result.

The upper bound on the mean squared error follows immediately because

$$\mathbb{E}_z[(\widehat{Z}(t) - z)^2] = \int_0^\infty \Pr_z\{(\widehat{Z}(t) - z)^2 > u\}\ du \ \leq \ 2\int_0^\infty e^{-tc_H\ u/2}\ du \ = \ \frac{4}{c_H} \cdot \frac{1}{t}$$

□

A.5. Proofs of Auxiliary Results

A.5.1 Proof of Remark 2.4.1

By Lemma A.1.2, we have that

$$(d(p;z) - d(p;z+u))^2 \geq d_{min}(1-d_{max})\mathcal{K}(Q^{p,z};Q^{p,z+u}) \ .$$

Now by applying the arguments of Lemma A.4.1 and using Assumption 3, we have that

$$\mathcal{K}(Q^{p,z};Q^{p,z+u}) \geq \frac{c_f}{2}u^2.$$

Choosing $c_d = d_{min}(1-d_{max})c_f/2$ establishes the inequality.

A.5.2 Chain Rule for Fisher Information

It is a standard result (e.g. Cover and Thomas, 1999, Exercise 11.19) that for distributions satisfying mild regularity assumptions (which are satisfied in our model), the Fisher information may also be written as

$$\mathbb{E}_z\left[\left(\frac{d}{dz}\log Q_t^{\psi,z}(\mathbf{Y}_t)\right)^2\right] = -\mathbb{E}_z\left[\frac{d^2}{dz^2}\log Q_t^{\psi,z}(\mathbf{Y}_t)\right],$$

So it follows that

$$\begin{aligned}\mathbb{E}_z\left[\left(\frac{d}{dz}\log Q_t^{\psi,z}(\mathbf{Y}_t)\right)^2\right] &= -\mathbb{E}_z\left[\frac{d^2}{dz^2}\log\prod_{\ell=1}^{t}Q_t^{\psi,z}(Y_\ell \mid \mathbf{Y}_{\ell-1})\right] \\ &= \sum_{\ell=1}^{t}-\mathbb{E}_z\left[\frac{d^2}{dz^2}\log Q_t^{\psi,z}(Y_\ell \mid \mathbf{Y}_{\ell-1})\right] \\ &= \sum_{\ell=1}^{t}\mathbb{E}_z\left[\left(\frac{d}{dz}\log Q_t^{\psi,z}(Y_\ell \mid \mathbf{Y}_{\ell-1})\right)^2\right] \ .\end{aligned}$$

This completes the proof.

Appendix B

Proofs from Chapter 3

B.1. Proof of Lemma 3.2.2

It is easy to check that for any $p \in \mathcal{P}$ and $z \in \mathcal{Z}$, we have $p^*(z) = 1/(2\sqrt{z})$, $r'(p^*(z); z) = 0$ and $r''(p; z) = -2z$. These facts together imply that for any $z \in \mathcal{Z}$,

$$r(p^*(z); z) - r(p; z) \geq \frac{1}{2} \inf_{z \in \mathcal{Z}} |r''(p; z)| \, (p^*(z) - p)^2 \geq (1/3)(p^*(z) - p)^2. \qquad \text{(B.1)}$$

Thus, to establish the inequality of Lemma 3.2.2, it suffices to establish a lower bound on the mean squared error $\mathbb{E}[(p^*(Z) - P_t)^2]$. Recall that $P_t = \psi_t(\mathbf{Y}_{t-1})$ is a function of the random variable $\mathbf{Y}_{t-1}$, or in other words, an *estimator* of the unknown parameter z. Note also that the optimal price $p^*(z)$ is an absolutely continuous function of z. With this interpretation, we may appeal to a standard result on the minimum mean-squared Bayes risk of an arbitrary estimator (van Trees' inequality, Theorem 2, Gill and Levit (1995)) to conclude that

$$\mathbb{E}[(p^*(Z) - P_{t+1})^2] \geq \frac{\left(\mathbb{E}[\frac{d}{dz} p^*(Z)]\right)^2}{\mathbb{E}\left[\left(\frac{d}{dz} \log Q_t^{\psi,Z}(\mathbf{Y}_t)\right)^2\right] + \mathbb{E}\left[\left(\frac{d}{dz} \log \lambda(Z)\right)^2\right]}, \qquad \text{(B.2)}$$

To establish the desired inequality, we note that since $\frac{d}{dz}p^*(z) = -1/(4z^{3/2}) \leq -1/3$ for all $z \in [1/3, 2/3]$, we have $\left(\mathbb{E}\left[\frac{d}{dz}p^*(Z)\right]\right)^2 \geq 1/9$. Also, it is easy to check that

$$\mathbb{E}\left[\left(\frac{d}{dz}\log\lambda(Z)\right)^2\right] = 216\pi^2\int_{1/3}^{2/3}\sin(3\pi(z-1/2))^2\,dz = 36\pi^2.$$

Combining these estimates with (B.1) and (B.2) proves the lemma.

B.2. Proof of Lemma 3.2.3

To prove the inequality of Lemma 3.2.3, we will use Lemma A.5.2, which states that

$$\mathbb{E}_{z_0}\left[\left(\frac{d}{dz}\log Q_t^{\psi,z}(\mathbf{Y}_t)\big|_{z=z_0}\right)^2\right] = \sum_{\ell=1}^{t}\mathbb{E}_{z_0}\left[\left(\frac{d}{dz}\log Q_t^{\psi,z}(Y_\ell|\mathbf{Y}_{\ell-1})\big|_{z=z_0}\right)^2\right] .$$

Using the chain rule, it suffices to show that for each $1 \leq \ell \leq t$, the instantaneous regret satisfies

$$\mathbb{E}_{z_0}\left[\left(\frac{d}{dz}\log Q_t^{\psi,z}(Y_\ell|\mathbf{Y}_{\ell-1})\big|_{z=z_0}\right)^2\right] \leq 30\,\mathbb{E}_{z_0}[r(p^*(z_0);z_0) - r(P_\ell;z_0)] .$$

To establish this inequality, recall from the proof of Lemma 3.2.2 that for the lower bound instance of Section 3.2, given by $\mathcal{P} = [\sqrt{2}/2, \sqrt{3}, 2]$, $\mathcal{Z} = [1/3, 2/3]$, and $d(p;z) = \sqrt{z} - pz$, we have $p^*(z) = 1/(2\sqrt{z})$. Also, recall that for this lower bound instance, we have by definition that for any $1 \leq \ell \leq t$,

$$Q_t^{\psi,z}(y_\ell = 1|\mathbf{y}_{\ell-1}) = \sqrt{z} - p_\ell z \qquad \text{and} \qquad Q_t^{\psi,z}(y_\ell = 0|\mathbf{y}_{\ell-1}) = 1 - \sqrt{z} + p_\ell z ,$$

where $p_\ell = \psi_\ell(\mathbf{y}_{\ell-1})$. Using these facts, fix an arbitrary $1 \leq \ell \leq t$ and $\mathbf{y}_{\ell-1} \in \{0,1\}^{\ell-1}$, and let $p_\ell = \psi_\ell(\mathbf{y}_{\ell-1})$. Then we have

$$\begin{aligned}
&\mathbb{E}_{z_0}\left[\left(\frac{d}{dz}\log Q_t^{\psi,z}(Y_\ell|\mathbf{Y}_{\ell-1})\big|_{z=z_0}\right)^2 \;\middle|\; \mathbf{Y}_{\ell-1}=\mathbf{y}_{\ell-1}\right] \\
&= \sum_{y_\ell\in\{0,1\}} \frac{\left(\frac{d}{dz}Q_t^{\psi,z}(y_\ell|\mathbf{y}_{\ell-1})\big|_{z=z_0}\right)^2}{Q_t^{\psi,z_0}(y_\ell|\mathbf{y}_{\ell-1})} \\
&= \frac{\left(\frac{d}{dz}(\sqrt{z}-p_\ell z)\big|_{z=z_0}\right)^2}{(\sqrt{z_0}-p_\ell z_0)} + \frac{\left(\frac{d}{dz}(1-\sqrt{z}+p_\ell z)\big|_{z=z_0}\right)^2}{(1-\sqrt{z_0}+p_\ell z_0)} \\
&= \left(\frac{1}{2\sqrt{z_0}}-p_\ell\right)^2\left(\frac{1}{(\sqrt{z_0}-p_\ell z_0)(1-\sqrt{z_0}+p_\ell z_0)}\right) \\
&\leq \frac{1}{d_{min}(1-d_{max})}\left(\frac{1}{2\sqrt{z_0}}-p_\ell\right)^2,
\end{aligned}$$

where the first line follows from the definition of the conditional expectation, the second line follows from the definition of the measure $Q_t^{\psi,z}(\,\cdot\,|\mathbf{y}_{\ell-1})$, and the remaining lines follow from simplification. Now using the fact that $p^*(z_0) = 1/(2\sqrt{z_0})$, we have

$$\begin{aligned}
\frac{1}{d_{min}(1-d_{max})}\left(\frac{1}{2\sqrt{z_0}}-p_\ell\right)^2 &= \frac{1}{d_{min}(1-d_{max})}(p^*(z_0)-p_\ell)^2 \\
&\leq 3(5)(2)\;\left(r(p^*(z_0);z_0)-r(p_\ell;z_0)\right) \\
&= 30\,\mathbb{E}_{z_0}\left[r(p^*(z_0);z_0)-r(P_\ell;z_0)\;\middle|\;\mathbf{Y}_{\ell-1}=\mathbf{y}_{\ell-1}\right],
\end{aligned}$$

where the second line follows from (B.1) and the fact that $d(p,z) \in [1/5, 1/2]$ for all $(p,z) \in \mathcal{P}\times\mathcal{Z}$, and the third line follows from the fact that $P_\ell = \psi_\ell(\mathbf{Y}_{\ell-1})$.

Since the above inequality is true for all $1 \leq \ell \leq t$ and all $\mathbf{y}_{\ell-1} \in \{0,1\}^{\ell-1}$, applying $\mathbb{E}_{z_0}[\,\cdot\,]$ to both sides gives

$$\mathbb{E}_{z_0}\left[\left(\frac{d}{dz}\log Q_t^{\psi,z}(Y_\ell|\mathbf{Y}_{\ell-1})\big|_{z=z_0}\right)^2\right] \leq 30\;\mathbb{E}_{z_0}[r(p^*(z_0);z_0)-r(P_\ell;z_0)]\,,$$

which is the desired result.

B.3. Proof of Lemma 3.2.4

Since $r''(p;z) = -2z$ and $\mathcal{Z} = [1/3, 2/3]$, we have that

$$\mathbb{E}_\lambda[\text{Regret}(Z, \mathcal{C}_{\mathsf{GenLB}}, 1, \psi)] \geq \frac{1}{3}\,\mathbb{E}_\lambda[(p_1 - p^*(Z))^2] = \frac{1}{3}\int_{1/3}^{2/3}\left(p_1 - \frac{1}{2\sqrt{z}}\right)^2 \lambda(z)\; dz$$

Consider $\delta = 1/54$. We know that $(p_1 - 1/(2\sqrt{z}))^2 \geq \delta^2$ whenever $|p_1 - 1/(2\sqrt{z})| \geq \delta$. Since $\frac{d}{dz}1/(2\sqrt{z}) = -1/(4z^{3/2}) \leq -1/3$ for all $z \in [1/3, 2/3]$, we can conclude that the set $I(p_1, \delta) = \{z \in [1/3, 2/3] : |p_1 - 1/(2\sqrt{z})| < \delta\}$ is an interval whose length $|I(p_1, \delta)|$ is bounded above by 6δ. So

$$\begin{aligned}
\frac{1}{3}\int_{1/3}^{2/3}\left(p_1 - \frac{1}{2\sqrt{z}}\right)^2 \lambda(z)\; dz &\geq \frac{\delta^2}{3}\int_{[1/3,2/3]\setminus I(p_1,\delta)} \lambda(z)\; dz = \frac{\delta^2}{3}\Big(1 - \int_{I(p_1,\delta)} \lambda(z)\; dz\Big) \\
&\geq \frac{\delta^2}{3}\Big(1 - \Big(|I(p_1,\delta)| \cdot \sup_{z\in[1/3,2/3]} \lambda(z)\Big)\Big) \\
&\geq \frac{\delta^2}{3}(1 - (6\delta)6),
\end{aligned}$$

where we use the fact that $\lambda(z) \leq 6$ for $z \in [1/3, 2/3]$. Since $\delta = 1/54$, we have

$$\mathbb{E}_\lambda[\text{Regret}(Z, \mathcal{C}_{\mathsf{GenLB}}, 1, \psi)] \geq \frac{1}{3(54^2)}\left(1 - \frac{36}{54}\right) = \frac{1}{26244}\;.$$

B.4. Proof of Lemma 3.5.4

We will first establish that for any $T \geq 2$ and any $\tau \in \{1, \ldots, \lceil \log T \rceil\}$, we have

$$\Gamma^*(\tau, T) \geq \tau T^{1/\tau} - \tau + 1 \geq \tau(T^{1/\tau} - 1).$$

We can lower bound $\Gamma^*(\tau, T)$ by considering its continuous relaxation defined by:

$$\Upsilon^*(\tau, T) = \min\left\{x_1 + \sum_{k=2}^{\tau} \frac{x_k}{\sum_{h=1}^{k-1} x_h} \;\Big|\; \sum_{h=1}^{\tau} x_h = T \quad \text{and} \quad x_k \in \mathbb{R}_+ \quad \forall k\right\}\;.$$

We will now show that $\Upsilon^*(\tau, T) = \tau T^{1/\tau} - \tau + 1$. The structure of the objective function associated with $\Upsilon^*(\tau, T)$ lends itself to a standard dynamic programming

approach. So, for $j = 2, \ldots, \tau$, let $V_j : [0, T] \to \mathbb{R}_+$ be defined by: for any $s \in [0, T]$,

$$V_j(s) = \min \left\{ \sum_{k=j}^{\tau} \frac{x_k}{s + \sum_{h=j}^{k-1} x_h} \;\middle|\; s + \sum_{h=j}^{\tau} x_h = T \quad \text{and} \quad (x_j, x_{j+1}, \ldots, x_\tau) \in \mathbb{R}_+^{\tau-j+1} \right\} .$$

By definition, we have that $V_\tau(s) = (T - s)/s = (T/s) - 1$. By our construction,

$$V_2(s) = \min \left\{ \sum_{k=2}^{\tau} \frac{x_k}{s + \sum_{h=2}^{k-1} x_h} \;\middle|\; s + \sum_{h=2}^{\tau} x_h = T \quad \text{and} \quad (x_2, x_3, \ldots, x_\tau) \in \mathbb{R}_+^{\tau-1} \right\} ,$$

and $\Upsilon^*(\tau, T) = \min_{q \in [0,T]} \{q + V_2(q)\}$. By definition of $V_j(\cdot)$, it is easy to verify that we have the following dynamic programming equation: for $j < \tau$,

$$V_j(s) = \min_{x_j \in \mathbb{R}_+ : s + x_j \le T} \left\{ \frac{x_j}{s} + V_{j+1}(s + x_j) \right\}$$

We will prove by induction that $V_j(s) = (\tau - j + 1) \left(\left(\frac{T}{s} \right)^{1/(\tau-j+1)} - 1 \right)$ for all $j = 2, 3, \ldots, \tau$. The result trivially holds in the base case because $V_\tau(s) = (T/s) - 1$, which is the desired result. Suppose that the result is true for $V_{j+1}(\cdot)$, we will show that it also holds for $V_j(\cdot)$. Note that

$$\begin{aligned} V_j(s) &= \min_{x_j \in \mathbb{R}_+ : s + x_j \le T} \left\{ \frac{x_j}{s} + V_{j+1}(s + x_j) \right\} \\ &= \min_{x_j \in \mathbb{R}_+ : s + x_j \le T} \left\{ \frac{x_j}{s} + (\tau - j) \left(\left(\frac{T}{s + x_j} \right)^{1/(\tau-j)} - 1 \right) \right\} \\ &= \min_{x_j \in \mathbb{R}_+ : s + x_j \le T} \left\{ \frac{s + x_j}{s} + (\tau - j) \left(\frac{T}{s + x_j} \right)^{1/(\tau-j)} - (\tau - j + 1) \right\} \end{aligned}$$

To solve the above problem, note that the minimizer of the following unconstrained optimization problem is given by

$$\arg\min_{q \ge 0} \left\{ \frac{q}{s} + (\tau - j) \left(\frac{T}{q} \right)^{1/(\tau-j)} \right\} = s^{(\tau-j)/(\tau-j+1)} \cdot T^{1/(\tau-j+1)} ,$$

which can be found by a simple analysis. Thus, the minimizer $x_j^*(s)$ for the above dynamic programming equation is given by

$$x_j^*(s) = s^{(\tau-j)/(\tau-j+1)} \cdot T^{1/(\tau-j+1)} - s ,$$

which is a feasible solution. This implies that

$$V_j(s) = \left(\frac{T}{s}\right)^{1/(\tau-j+1)} + (\tau-j)\left(\frac{T}{s}\right)^{1/(\tau-j+1)} - (\tau-j+1) = (\tau-j+1)\left(\left(\frac{T}{s}\right)^{1/(\tau-j+1)} - 1\right) ,$$

which completes the induction. It follows that

$$\Upsilon^*(\tau, T) = \min_{q\in[0,T]} \{q + V_2(q)\} = \min_{s\in[0,T]} \left\{ q + (\tau-1)\left(\left(\frac{T}{q}\right)^{1/(\tau-1)} - 1\right)\right\} \;=\; \tau T^{1/\tau} - \tau + 1 ,$$

where the last inequality follows from the same argument as before.

To finish the proof, note that for any $T \geq 2$ and $\tau \in \{1, \ldots, \lceil \log T \rceil\}$, we have $T^{1/\tau} = e^{\log T/\tau} \geq e^{1/2}$, so that $(T^{1/\tau} - 1) \geq (1 - e^{-1/2})T^{1/\tau} \geq \{T^{1/\tau}\}/3$. It follows that

$$\Gamma^*(\tau, T) \geq \tau(T^{1/\tau} - 1) \geq \frac{\tau T^{1/\tau}}{3} ,$$

which is the desired result.

Appendix C

Proofs from Chapter 4

C.1. Proof of Lemma 4.6.2

To prove the Lemma, we will use the following result from Garivier and Moulines (2008).

Lemma C.1.1 (Theorem 18, Garivier and Moulines (2008))**.** *Let* $\{V_t : t \geq 1\}$ *be a sequence of i.i.d. random variables with common mean* μ *and with* $V_t \in [0,1]$ *with probability one for all* $t \geq 1$*. Let* $\{\mathcal{F}_t : t \geq 1\}$ *be a filtration with* $\sigma(V_1, \ldots, V_t) \subset \mathcal{F}_t$ *and* V_{t+1} *independent from* $\mathcal{F}_t$ *for all* $t \geq 1$*. Let* $\{B_t : t \geq 1\}$ *be a sequence of Bernoulli random variables with* $B_t \in \mathcal{F}_{t-1}$ *for all* t*. Let* $N(t) = \sum_{i=1}^{t} B_i$*. Then*

$$\Pr\left\{\left|\frac{1}{N(t)}\sum_{i=1}^{t} V_i B_i - \mu > \sqrt{\frac{\delta}{N(t)}}\right|\right\} \leq 8\lceil \log t \rceil e^{-1.99\delta}.$$

To apply this result, note that with probability one, $\mathbf{1}\{Y_i \geq s\}\mathbf{1}\{X_i \geq s\} = \mathbf{1}\{D_i \geq s\}\mathbf{1}\{X_i \geq s\}$. This means that with probability one, for any fixed $i \in [n]$

and $s \in [c]$, we have

$$\widehat{F}^t(i,s) = \frac{1}{N^t(i,s)} \sum_{\ell \in \mathcal{N}^t(i,s)} \mathbf{1}\left\{Y_i^\ell \geq s\right\} = \frac{1}{N^t(i,s)} \sum_{j=1}^{t} \mathbf{1}\left\{Y_i^j \geq s\right\} \mathbf{1}\left\{X_i^j \geq s\right\}$$
$$= \frac{1}{N^t(i,s)} \sum_{j=1}^{t} \mathbf{1}\left\{D_i^j \geq s\right\} \mathbf{1}\left\{X_i^j \geq s\right\}.$$

Applying Lemma C.1.1 with $V_j = \mathbf{1}\left\{D_i^j \geq s\right\}$, $B_j = \mathbf{1}\left\{X_i^j \geq s\right\}$, and $\mathcal{F}_j = \sigma(\{D_i^j : i \in [n], j \in [t]\})$, and recalling that $w(n,t) = 2\sqrt{(\log t)/n}$, we have

$$\Pr\left\{\left|\widehat{F}^t(i,s) - \Pr\{D_i \geq s\}\right| \geq w(N^t(i,s),T)\right\} \leq 8\lceil \log t \rceil e^{-1.99(4\log T)} \leq 8\lceil \log t \rceil T^{-7} \leq T^{-3},$$

where the last inequality follows from the fact that $1 \leq t \leq T$ and $T \geq 2$. Taking the union bound over all $t \in [T]$, $i \in [n]$, and $s \in [c]$ proves the claim.

Bibliography

Agarwal, A., P. Bartlett, M. Dama. 2010. Optimal allocation strategies for the dark pool problem. *Working paper, UC Berkeley* .

Agrawal, R. 1995. The continuum-armed bandit problem. *SIAM Journal of Control and Optimization* **33**(6) 1926–1951.

Anantharam, V., P. Varaiya, J. Walrand. 2002. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part I: Iid rewards. *Automatic Control, IEEE Transactions on* **32**(11) 968–976.

Araman, V., R. Caldenty. 2005. Dynamic pricing for non-perishable products with demand learning. *Operations Research* **57**(5) 1169–1188.

Auer, P., N. Cesa-Bianchi, P. Fischer. 2002a. Finite-time analysis of the multi-armed bandit problem. *Machine Learning* **47**(2) 235–256.

Auer, P., N. Cesa-Bianchi, P. Fischer. 2002b. Finite-time analysis of the multi-armed bandit problem. *Machine learning* **47**(2) 235–256.

Auer, P., N. Cesa-Bianchi, Y. Freund, R.E. Schapire. 2003. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing* **32**(1) 48–77.

Auer, P., R. Ortner, C. Szepesvári. 2007. Improved rates for the stochastic

continuum-armed bandit problem. *20th Conference on Learning Theory (COLT)* 454–468.

Aviv, Y., A. Pazgal. 2002. Pricing of short life-cycle products through active learning. *Working paper, Washington University* .

Ben-Akiva, M., S. Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. The MIT Press: Cambridge, MA.

Bertsimas, D., G. Perakis. 2003. Dynamic pricing: A learning approach. *Mathematical and Computational Models for Congestion Charging*, *Applied Optimization*, vol. 101. Springer, New York, 45–79.

Besbes, O., A. Zeevi. 2008. Dynamic pricing without knowing the demand function: Risk bounds and near optimal algorithms. *To appear in Operations Research* .

Besbes, O., A. Zeevi. 2009. On the minimax complexity of pricing in a changing environment. *To appear in Operations Research* .

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

Bitran, G., S. Mondschein. 1997. Periodic pricing of seasonal products in retailing. *Management Science* **43** 64–79.

Borovkov, A. 1998. *Mathematical Statistics*. Gordon and Breach Science Publishers.

Bretthauer, K.M., B. Shetty. 1995. The nonlinear resource allocation problem. *Operations Research* **43**(4) 670–683.

Broadie, M., D. Cicek, A. Zeevi. 2009. General bounds and finite-time improvement for the kiefer-wolfowitz stochastic approximation algorithm. *To appear in Operations Research* .

Brynjolfsson, E., M. Smith. 2000. Frictionless commerce? a comparison of internet and conventional retailers. *Management Science* **46**(4) 563–585.

Carvalho, A., M. Puterman. 2005. Learning and pricing in an internet environment with binomial demands. *Journal of Revenue & Pricing Management* **3**(4) 320–336.

Çelik, S., A. Muharremoglu, S. Savin. 2009. Revenue management with costly price adjustments. *Operations Research* **57**(5) 1206–1219.

Cesa-Bianchi, N., G. Lugosi. 2006. *Prediction, learning, and games*. Cambridge Univ Pr.

Cope, E. 2006. Bayesian strategies for dynamic pricing in e-commerce. *Naval Research Logistics* **54**(3) 265–281.

Cope, E. 2009. Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Transactions on Automatic Control* **54** 1243–1253.

Cover, T., J. Thomas. 1999. *Elements of Information Theory*. J. Wiley, Hoboken.

den Boer, A., B. Zwart. 2010a. Simultaneously learning and optimizing using controlled variance pricing. *Working paper, Centrum Wiskunde & Informatica and the University Amsterdam* .

den Boer, A., B. Zwart. 2010b. Simultaneously learning and optimizing using controlled variance pricing. *working paper, CWI* .

Fabian, V. 1967. Stochastic approximation of minima with improved asymptotic speed. *Annals of Mathematical Statistics* **38**(1) 191–200.

Farias, V., B. Van Roy. 2007. Dynamic pricing with a prior on market response. *To appear in Operations Research* .

Feng, Y., G. Gallego. 1995. Optimal starting times for end-of-season sales and optimal stopping times for promotional fares. *Management Science* **46** 1371–1391.

Ganchev, K., Y. Nevmyvaka, M. Kearns, J.W. Vaughan. 2010. Censored exploration and the dark pool problem. *Communications of the ACM* **53**(5) 99–107.

Garivier, A., E. Moulines. 2008. On upper-confidence bound policies for non-stationary bandit problems. *Preprint. http://arxiv. org/abs/0805.3415* .

Gill, R., B. Levit. 1995. Applications of the van trees inequality: A Bayesian Cramér-Rao bound. *Bernoulli* **1**(1) 59–79.

Goldenshluger, A., A. Zeevi. 2008. Performance limitations in bandit problems with side observations. *To appear in IEEE Transactions on Information Theory* .

Goldenshluger, A., A. Zeevi. 2009. Woodroofe's one-armed bandit problem revisited. *Annals of Applied Probability* **19**(4) 1603–1633.

Harrison, J., N. Keskin, A. Zeevi. 2010. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Working Paper, Stanford University* .

Hochbaum, D.S. 1994. Lower and upper bounds for the allocation problem and

other nonlinear optimization problems. *Mathematics of Operations Research* **19**(2) 390–409.

Ibaraki, T., N. Katoh. 1988. Resource allocation problems: Algorithmic approaches .

Kiefer, J., J. Wolfowitz. 1967. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics* **23**(3) 462–466.

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

Knuth, D. 1997. *The Art of Computer Programming, Volume 1: Fundamental Algorithms*. Addison-Wesley.

Lai, T., H. Robbins. 1985a. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* **6**(1) 4–22.

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

Levy, D., M. Bergen, S. Dutta, R. Venable. 1997. The magnitude of menu costs: direct evidence from large us supermarket chains. *Quarterly Journal of Economics* **112** 791–825.

Levy, D., S. Dutta, M. Bergen, R. Venable. 1998. Price adjustment at multiproduct retailers. *Managerial and Decision Economics* **19** 81–120.

Lobo, M., S. Boyd. 2003. Pricing and learning with uncertain demand. *Working paper, Duke University* .

Netessine, S. 2006. Dynamic pricing of inventory/capacity with infrequent price changes. *Eur. J. Oper. Res.* **174**(1) 553–580.

Talluri, K., G. van Ryzin. 2004. *The Theory and Practice of Revenue Management*. Springer, New York.

Taneja, I., P. Kumar. 2004. Relative information of type s, Csiszár's f-divergence, and information inequalities. *Information Sciences* **166**(1–4) 105–125.

Tsybakov, A. 2009. *Introduction to Nonparametric Estimation*. Springer, New York.

Weiss, R., A. Mehrotra. 2001. Online dynamic pricing: Efficiency, equity, and the futuer of e-commerce. *Virginia Journal of Law and Technology* **6**(2). `http://www.vjolt.net/vol6/issue2/v6i2-a11-Weiss.html`.

Zbaracki, M., M. Ritson, D. Levy, S. Dutta, M. Bergen. 2004. Managerial and customer costs of price adjustment: Direct evidence from industrial markets. *Review of Economics and Statistics* **86**(2) 514–533.