The Practical Value of Field Experiments

Jimmy Q. Li[*], Paat Rusmevichientong[†], Duncan Simester[§], John N. Tsitsiklis[*], and Spyros I. Zoumpoulis[*]

[*]Laboratory for Information and Decision Systems, Massachusetts Institute of Technology

[†]Marshall School of Business, University of Southern California

[§]MIT Sloan School of Management, Massachusetts Institute of Technology

September 4, 2013

The authors thank Eric Anderson, Eric Bradlow, John Liechty, Olivier Toubia, Catherine Tucker, Martin Wainwright, and Juanjuan Zhang for their helpful comments and suggestions. The paper has also greatly benefited from comments by the anonymous reviewers and the Associate Editor. This research was partially supported by NSF grants CMMI-0856063 and CMMI-1158658.

Abstract

In many situations, the capabilities of firms are better suited to conducting and analyzing field experiments than to analyzing sophisticated demand models. However, the practical value of using field experiments to optimize marketing decisions remains relatively unstudied. We investigate category pricing decisions that require estimating a large matrix of cross-product demand elasticities and ask: how many experiments are required as the number of products in the category grows? Our main result demonstrates that if the categories have a favorable structure then we can learn faster and reduce the number of experiments that are required: the number of experiments required may grow just logarithmically with the number of products. These findings potentially have important implications for the application of field experiments. Firms may be able to obtain meaningful estimates using a practically feasible number of experiments, even in categories with a large number of products. We also provide a relatively simple mechanism that firms can use to evaluate whether a category has a structure that makes it feasible to use field experiments to set prices. We illustrate how to accomplish this using either a sample of historical data or a pilot set of experiments. We also discuss how to evaluate whether field experiments can help optimize other marketing decisions, such as selecting which products to advertise or promote.

1 Introduction

The increased availability of demand data has been widely reported and many firms have been investigating how best to use "Big Data" to improve their marketing decisions. One option is to conduct analysis on historical data. However, the analytical sophistication required to generate insight from historical data is not always within a firm's capabilities. An alternative is to conduct field experiments and use the demand data as a feedback mechanism. As long as there is a high quality control group, analyzing the results of a field experiment can be as simple as a comparison of two means. In many cases, the capabilities of firms are better suited to conducting and analyzing field experiments than to analyzing sophisticated demand models.

While firms may be able to conduct field experiments, they generally incur costs to do so, and the practical value of using field experiments to improve marketing decisions remains relatively unstudied. We investigate this issue by considering settings in which firms must estimate the elasticity of demand in response to price changes. We ask how many experiments are required to estimate these elasticities as the number of products grows.

Using experiments to optimize marketing decisions may be relatively straightforward when there are few products. Experimentally manipulating variables can allow retailers to quickly optimize their decisions using just a handful of experiments. However, in large categories containing many products with interdependent demands, the problem is more challenging.[1] The number of parameters to estimate grows quickly with the number of products, and so the number of field experiments required may be impractically large.

We consider a large set of n products and assume that there may be complementary or substitute relationships between them. As a result, varying the price of one product may affect the demand of not just that item but also other products sold by the firm. As the number of products (n) increases, the number of parameters to estimate grows at the rate of n^2 (and may grow even faster for nonlinear models). On the other hand, if each experiment reveals the demand for each item, we learn n pieces of information from each experiment. This suggests that the number of experiments required to learn all of the parameters will grow at least linearly with the number of products.

Our main result shows that if the problem has a favorable structure, we can learn faster and reduce the number of experiments that are required. In particular, we will show that if the number of complementary or substitute relationships affecting any one product is bounded, then the number of required experiments grows instead logarithmically with the number of products. This result holds even if the firm is not sure which of the products have complementary or substitute relationships, as long as there is a limit on the number

[1] Interdependencies between products are now well-documented. For example, Anderson and Simester (2001) report that placing "sale" signs on products can increase demand for those products by up to 60%, but can decrease sales of other products by similar amounts. Manchanda et al. (1999) report own-price elasticities for laundry detergent and fabric softener of -0.40 and -0.70 (respectively). The cross-price elasticities are -0.06 (the price of softener on demand for detergent) and -0.12. For cake mix and frosting, the own-price elasticities are -0.17 and -0.21, while the cross-price elasticities are -0.11 (frosting price on cake mix demand) and -0.15.

of cross-product relationships that each product has. We also obtain a similar result if the joint impact of own- and cross-product effects on any single product is bounded.

We also provide a practical method for evaluating whether a product category has a favorable structure that makes it feasible to use field experiments to set category prices. In particular, we propose a method for estimating the bounds on the number of interdependencies between products. The method can be implemented using either a pilot set of experiments, or using historical data. We illustrate this using both simulations and a sample of actual data from the cold remedies category. Our empirical results suggest that within the cold remedies category, the number of complementary or substitute relationships grows sublinearly with the number of products, and therefore that the elasticity parameters can indeed be feasibly estimated using relatively few experiments.

These findings potentially have important implications for the application of field experiments in settings where there is a large number of parameters to estimate. Because the number of required experiments grows logarithmically rather than linearly with the number of products, firms may be able to obtain meaningful estimates from a realistic number of experiments, even in categories where the number of complementary or substitutable products is large.

Although we focus on pricing decisions in this paper, the range of marketing decisions on which firms can experiment is broad. Experiments may be used to choose which products to promote, as well as to optimize the length of product lines and to choose creative copy and media plans. We discuss how to extend our results to make promotional decisions, and in the Conclusions section discuss possible extensions to other types of marketing decisions.

1.1 Related Work

The feasibility of learning a large number of parameters through experimentation is relatively unstudied, particularly in social science settings. However, the topic does relate to at least two literatures.

First, there is the line of research on optimal experimental design. In the marketing literature, there is work focusing on efficient experimental design for conjoint studies (see Louviere et al. 2000, Chapter 5; and Louviere et al. 2004 for reviews of this literature). Recent contributions to this literature have focused on adaptively designing experiments (Toubia et al. 2003) or on optimal designs when customers' utility functions depart from a standard compensatory specification (see for example Hauser et al. 2010, Liu and Arora 2011). An often used measure of the efficiency of an experimental design is the D-error: $\det[I(\theta \mid X)]^{-1/m}$, where I is the information matrix, θ are the unobserved parameters, X is the experimental design matrix, and m is the dimension of I. The information matrix is calculated from the variance of the first-order derivatives of the log-likelihood with respect to θ (Huber and Zwerina 1996). Optimizing this criterion with respect to X yields locally optimized designs for any θ. Because θ is not known when designing the experiments, Bayesian approaches can be used to minimize the D-error over the prior distribution of the parameter values (Sandor and Wedel 2001).

When each experiment generates an explicit reward or cost, an alternative formulation

of the experimental design problem is as a multi-armed bandit problem, where the objective is to choose a sequence of experiments to maximize the total reward over some time horizon. In this context, each experiment can be thought of as choosing and pulling an arm of the multi-armed bandit, and the reward could be sales, advertising click-through rates, or some other measure. Because we learn the reward distribution of each arm of the bandit only after pulling it, there exists a trade-off between *exploiting* the best arm currently known by pulling it every time and *exploring* new arms in search of something even better. In the classic bandit model, the reward distributions of each arm are assumed to be independent, and so anything learned from pulling one arm does not reveal anything about a different arm. As a result, when there is a large number of parameters (and therefore a large number of arms), many pulls, or experiments, are required to learn the reward distributions of all the arms. Recent work has proposed an alternative model in which the arms have statistically dependent reward distributions, and therefore pulling one arm also gives information about other arms. In this setting, the correlation between payoffs of different arms allows for faster learning, even when the number of arms is very large (Dani et al. 2008, Mersereau et al. 2009).

This focus on the information learned from experiments is a common feature of both this literature and the research in this paper. However, we do not focus on identifying optimal experimental designs. Instead we use random experimental designs, which ensure independence between experiments and allow us to apply a series of results that rely on this independence. Because it will generally be possible to improve on these designs, our guarantees on the information learned will continue to hold when optimal designs are used.

We investigate the practical value of field experiments by studying the number of experiments required. Other studies have investigated the required size of field experiments. For example, Lewis and Rao (2012) conducted a set of 25 field experiments involving large display advertising campaigns, each one including over 500,000 unique users and totaling over \$2.8M worth of impressions. Even with such large experiments, the data generated little meaningful information about the ROI of the campaigns, demonstrating that in settings where the effect sizes are small and the response measures are highly stochastic, very large field experiments may be required to generate information.

The second related literature is that on estimation and learning under assumptions of sparsity. Beginning with variable selection in regressions, research has focused on determining which subset of potential predictors should be included in the "best" model. This can equivalently be thought of as selecting a subset of predictors to be zero, thereby giving rise to a sparse model. Various approaches have been proposed, including the use of regularization, such as the "Lasso" of Tibshirani (1996), and the Stochastic Search Variable Selection procedure developed in George and McCulloch (1993).

More recently, the assumption of sparse structures has been used to show that if an unknown vector $\mathbf{x} \in \mathbb{R}^N$ is sparse, then it can be recovered using measurements of the form $\mathbf{y} = \Phi\mathbf{x}$, even with much fewer than N measurements. Results in the field, which is often referred to as "compressive sensing," generally characterize conditions on (i) the sparsity index (i.e., the number of nonzero entries of $\mathbf{x}$), (ii) the number of measurements, and (iii)

the ambient dimension N, in order to guarantee recovery of $\mathbf{x}$. We refer the reader to Candès (2006) for a short survey, and to Candès and Tao (2005), Candès et al. (2006) for a deeper treatment.

More directly relevant to our work are the results on information-theoretic limits of sparsity recovery in Wainwright (2009). For a noisy linear observation model based on sensing matrices drawn from the standard Gaussian ensemble, a set of both sufficient and necessary conditions for asymptotically perfect recovery is derived. Our theoretical findings are best thought of as an application of Wainwright (2009) results. Although this application required some theoretical developments, these are best considered adaptations and extensions rather than fundamentally new developments. The exception is the estimation of the sparsity parameters in Section 4 and the investigation of how these parameters vary with the size of the problem (the number of products). This is the first paper that we know of that addresses these issues.

Originating from and motivated by applications in signal processing, coding theory, and statistics, compressive sensing results have also a variety of other relevant applications. Previous applications related to marketing include Farias et al. (2013), which introduces a paradigm for choice modeling where the problem of selecting an appropriate model of choice (either explicitly, or implicitly within a decision making context) is itself automated and data-driven. For this purpose, the sparsest choice model consistent with observed data is identified.

In this work, we leverage sparsity to obtain a dramatic improvement in the rate of learning. If each product is substitutable by or complementary with a limited number of other products (and therefore the matrix capturing the substitution and complementarity effects is sparse), we show that the number of required experiments grows logarithmically with the number of products.

1.2 Overview

We consider pricing decisions for a firm with a large assortment of products. The firm would like to know how price changes will affect demand. We propose a model for the demand function, which tells us the quantities demanded under any pricing decision. In order to learn the parameters of this function, we perform experiments by varying the prices of certain products and observing the quantities demanded. Because each experiment is costly to run, the firm would like to learn the parameters using as few experiments as possible.

The experiments that we contemplate include both a treatment group and a control group. The construction of these groups will vary depending upon the nature of the firm. For a direct marketing firm, the groups may be constructed by randomly assigning individual customers to the two groups. For a bricks and mortar retailer, the groups might be constructed by randomly assigning stores. In a business to business setting, the firm might randomly assign regions, or distributors and resellers. We assume that the results of the experiment are analyzed by aggregating the customers in each group and comparing the mean response between the two groups. Essentially all firms are capable of performing this

aggregate analysis (as long as they can vary prices and measure the response).[2] This also ensures that the error terms are Gaussian.

Our findings can also apply in settings where the firms vary prices across different time periods. Demand in the different time periods could in principle be adjusted to account for seasonality or day-of-week differences (before submitting the data to our model), perhaps using demand for a sample of unrelated products or demand in different stores. We caution that we will assume that errors are independent between experiments (though not between products in the same experiment), and this independence assumption may be threatened when a common set of measures is used to adjust for seasonality. The independence assumption is more likely to hold when randomization occurs separately for each experiment, and when the control group provides an accurate control for any intervening events (such as seasonality).

We also caution that our results are *not* well suited to experiments where firms randomly assign *products* to treatment and control groups if the demands for those products are possibly related. For example, a firm may vary prices on half of the items in a product category and leave the other half of the prices unchanged. Recall that the goal of the paper is to investigate how a firm can estimate the entire matrix of cross-price elasticities and so the second half of the products cannot function as controls. There are other reasons to be concerned about this experimental design. Unless the cross-price elasticities are zero, then the experimental manipulation of prices in the treatment group of products will confound the demands in the control group.

We recognize that it is possible to augment experimental data with more complex econometric analysis (for example, Manchanda et al. 1999). This raises an interesting but distinct topic: what is the value of sophisticated analysis in evaluating experimental data? This question is beyond the scope of the present work. Instead, our results can be interpreted as describing the "information" that is revealed by experimental data. Conditions under which experimental data are more informative are likely to yield better estimates both when using simple comparisons and when augmenting the data with sophisticated econometric analysis.

The rest of this paper is structured as follows: In Section 2, we propose a model for demand that captures the effects of cross-product demand elasticities. In Section 3, we develop a method for estimating the demand function and provide bounds on the number of experiments required to achieve accurate estimates. In Section 4, we propose a method for estimating how sparse the price elasticities are, which provides a practical way for managers to evaluate whether it is feasible to set prices using field experiments. We also investigate how sparsity is affected by the size of the category. In Section 5, we present simulation results that illustrate the rate at which we acquire information, as the number of products and number of experiments vary. Finally, in Section 6, we conclude and describe directions for extensions and future research.

[2]Even though direct marketing firms often could analyze experimental results at the individual customer-level, in our experience most firms simply aggregate the results and compare the mean response between treatment and control groups.

2 Model

In this section, we introduce our model for demand. Throughout this paper, we consider each experiment as a comparison between two conditions. The first condition is a control under which the firm takes "standard" actions; in the second treatment condition, the firm varies prices. For ease of exposition (and without loss of generality), we will assume that prices are set at a "baseline" level in the control condition.

2.1 Modeling Own- and Cross-Price Elasticities

The response in demand to a firm's action is difficult to predict because there are multiple effects at play due to cross-product substitute and complementary relationships. In the following sections, we present a model that captures these effects.

2.1.1 Individual and Pairwise Effects

Changing the price of product i may have two effects:

(i) It may change demand for the product itself.

(ii) It may also affect the demand for other products through substitution away from the focal product or complementarity with the focal product.

For the first effect, we introduce a quantity a_{ii} to indicate the *percentage change* in demand for product i if the price of product i itself is increased by 100%.[3] For the second effect, we first consider a pair of products in isolation. Intuitively, there are three possible scenarios:

1. If products i and j are substitutes, decreasing the price of j may decrease the demand for i if customers substitute purchases of j for purchases of i.

2. If i and j are complements, decreasing the price of j may increase the demand for i as more demand for j leads to more demand for i.

3. If i and j are unrelated, then varying the price of j has no effect on the demand for i.

For each pair of products i and j, we introduce a quantity a_{ij} to indicate the *percentage change* in demand for product i if the price of product j is increased by 100%. The quantity a_{ij} would be positive, negative, and zero, in cases 1, 2, and 3 above, respectively.

[3] This is not to say that in our experiments, we propose increasing prices by 100%.

2.1.2 Cumulative Effects

We are interested in settings in which there are dozens of products with hundreds of interactions at play. If multiple prices are varied simultaneously, how do these changes combine and interact to produce an overall effect on demand?

To capture the cumulative effects, we propose a linear additive model of overall substitution and complementarity effects. Specifically, to calculate the overall percentage change in demand for product i, we take all of the products j whose prices are varied and sum together each one's individual effect on the demand for i.

Let Δq_i be the overall percentage change in the demand for i, and let us express the percentage change in the price of product j from the baseline as

$$x_j = \frac{x_j^t - x_j^b}{x_j^b},$$

where x_j^t and x_j^b are the treatment and baseline prices, respectively, of product j. We denote the number of products by n. Then, by our model, we can write the overall percentage change in demand for i as

$$\Delta q_i = \sum_{j=1}^{n} a_{ij} x_j.$$

By assuming a linear model, we are implicitly assuming that the elasticities are the same at all points on the demand curve. Although this may be appropriate for small price changes, it is unlikely to be true for large price changes. We can ensure that price changes are small by bounding the size of the price changes in the experiments. More generally, we can interpret our linear model as an approximation to the true model in the neighborhood around the baseline levels of price and demand, in the spirit of a first-order Taylor approximation. The model also assumes additive separability in the impact of the multiple price changes on the demand for product i. This is convenient for analytical tractability. In Appendix A, we show that it is relatively straightforward to extend our findings to a multiplicative demand model. More generally, recall that our goal is to approximate the (somewhat abstract) concept of information learned from a field experiment. Although the actual parameter estimates may be influenced by this specification of the demand model, we believe our approximations of the rate of learning are likely to be robust.

In some cases a firm may want to focus on improving just a subset of prices in the category. This could occur if some items sell relatively low volumes and optimizing these prices is not a priority (or if the retail prices are set by the manufacturer of the brand). This may also arise if too many experiments are required to optimize all of the prices in the category, and so the firm would just like to focus on just those prices that it considers most important.[4] We can easily accommodate this possibility by identifying the products that the firm does not want to experiment with, and collapsing these products into a single "other" product. Sales of this "other" product is simply the sales of the products within it.

[4]We thank an anonymous reviewer for this suggestion.

We could also construct a price index for the "other" product by averaging the prices of the corresponding items (because the firm does not want to experiment with these prices, the value of the corresponding x_j's will always equal zero). This allows the firm to focus on a subset of products in the category, while continuing to take into account the impact on sales across the entire category.

We can further simplify notation by collecting all of the pairwise effects as elements of a matrix $\mathbf{A}$, where (as suggested by the notation) the entry in the ith row and jth column, a_{ij}, gives the percentage change in demand for product i in response to a 100% increase in the price of product j.[5] Similarly, we can collect price variation decisions into a vector $\mathbf{x}$ whose jth element x_j is equal to the percentage change in the price of product j from the baseline, and we can also collect the overall percentage change in demand for each product into a vector $\Delta\mathbf{q}$.

The overall percentage change in each product's demand due to price changes $\mathbf{x}$ is therefore given by the product

$$\Delta\mathbf{q} = \mathbf{A}\mathbf{x}.$$

The elements a_{ij} of the matrix $\mathbf{A}$ may be positive (indicating a substitute relationship between i and j), negative (indicating a complementary relationship), or zero (indicating no relationship).[6]

We emphasize that the matrix $\mathbf{A}$ captures *percentage changes* in demand. To calculate actual demand quantities, we also need a baseline level of demand for each product. Recall that we assume there is a fixed set of firm actions, corresponding to the control condition, which achieves a certain level of demand. We let this be the baseline demand and denote it by the vector $\mathbf{q}^b$. The overall change in demand for a product in response to the price changes is then given by the product of the baseline demand and the percentage change in demand.

[5] We do not impose symmetry (i.e., $a_{ij} = a_{ji}$) or transitivity (i.e., $a_{ij} > 0, a_{jk} > 0 \Rightarrow a_{ik} > 0$) on the $\mathbf{A}$ matrix for two reasons. First, there are examples where these constraints are intuitively unlikely to hold (e.g., price decreases on cameras may increase battery sales but not vice versa, violating symmetry; price decreases on milk may increase sales of cereal, and price decreases on cereal may increase sales of soymilk, but price decreases on milk may not increase sales of soymilk, violating transitivity). Second, neither symmetry nor transitivity is a necessary assumption for our analysis, and imposing these constraints would only make our results weaker and less applicable. Instead, we want the space of "allowable" $\mathbf{A}$ matrices to be as large as possible. Furthermore, if the true $\mathbf{A}$ matrix is indeed symmetric or transitive, then because our method gives accurate estimates, the estimated matrix would also be close to symmetric or transitive with high probability.

[6] We also assume that the matrix $\mathbf{A}$ is constant. It is possible that there may be time dependencies or seasonal effects that could lead to changes in the $\mathbf{A}$ matrix. The model could accommodate these possibilities as long as these dynamics are known so that we can continue to estimate a static set of parameters. If the parameters themselves change in a manner that is not known, then the results of an experiment performed at time t may not provide much information about the value of the parameters in future periods. Note that this limitation is obviously not specific to our model.

2.2 Noiseless Model

Let $\mathbf{q}^t$ be the vector of actual demand levels in response to a decision $\mathbf{x}$, which we refer to as the *treatment* demand level. We then have the following equation for our model:

$$\mathbf{q}^t = \mathbf{q}^b + \mathbf{q}^b \circ (\Delta\mathbf{q}) = \mathbf{q}^b \circ (\mathbf{e} + \mathbf{A}\mathbf{x}), \tag{1}$$

where $\circ$ denotes component-wise multiplication, and $\mathbf{e}$ is the vector of all 1's. In words, price changes $\mathbf{x}$ will cause some percentage change in demand through the elasticity matrix $\mathbf{A}$, which when combined with the baseline demand $\mathbf{q}^b$ give the observed treatment demand $\mathbf{q}^t$. Note that this model has the desired property that when prices are the same as the baseline prices (i.e., $\mathbf{x} = \mathbf{0}$), the treatment demand is the same as the baseline demand (i.e., $\mathbf{q}^t = \mathbf{q}^b$) because there is effectively no treatment.

We can also rewrite Equation (1) as

$$\Delta\mathbf{q} = \frac{\mathbf{q}^t - \mathbf{q}^b}{\mathbf{q}^b} = \mathbf{A}\mathbf{x}, \tag{2}$$

where the division is performed component-wise. The left-hand-side gives the percentage change in demand for each product, and the right-hand-side gives the model of how that change is caused by the decision vector. This arrangement emphasizes the fact that $\mathbf{A}$ captures the percentage change in demand. It also suggests a way of learning $\mathbf{A}$: for each experiment, choose a decision vector $\mathbf{x}$, observe the resulting $\mathbf{q}^b$ and $\mathbf{q}^t$, and calculate $\Delta\mathbf{q}$. This gives a system of linear equations from which we can recover $\mathbf{A}$, ideally using as few experiments as possible.

2.3 Noisy Model

In reality, the demand function is not captured perfectly by Equation (1), and the demand that we observe will also be subject to measurement noise. Therefore, Equation (1) gives only an idealized model. To capture the presence of error, we introduce an additive noise term $\mathbf{w}$, which is a vector of random variables $(w_1, w_2, \ldots, w_n)$. Our complete model is then given by

$$\mathbf{q}^t = \mathbf{q}^b \circ (\mathbf{e} + \mathbf{A}\mathbf{x} + \mathbf{w}), \tag{3}$$

which can also be written as

$$\Delta\mathbf{q} = \frac{\mathbf{q}^t - \mathbf{q}^b}{\mathbf{q}^b} = \mathbf{A}\mathbf{x} + \mathbf{w}. \tag{4}$$

Equations (3) and (4) are analogous to Equations (1) and (2) with the additional noise vector $\mathbf{w}$. The observed treatment demand is modeled as a deviation from the baseline demand due to price changes and noise.

2.3.1 Statistics of the Noise Terms

For our analysis, we make the following assumptions on the noise terms:

Assumption 1 (Zero-mean, sub-Gaussian noise, i.i.d. across experiments)**.** *For any experiment, each w_i has zero mean and is sub-Gaussian with parameter c for some constant $c \geq 0$. Furthermore, the random variables w_i are independent and identically distributed across different experiments.*

We assume that the noise terms have zero mean, and therefore that our model has no systematic bias. We also assume that the noise terms across different experiments are independent and identically distributed. However, we do not assume that the noise terms are independent between different products within the same experiment. In other words, each experiment gets an independent draw of $(w_1, \ldots, w_n)$ from a single joint distribution in which the w_i's can be dependent. Indeed, the noise terms within the same experiment may be correlated across products (e.g., between products within the same category). Fortunately our analysis does not require independence at this level.

Sub-Gaussian random variables are a generalization of Gaussian random variables, in the sense that their distributions are at least as concentrated around their means as Gaussian distributions.

Definition 1. A random variable X is *sub-Gaussian* with parameter $\sigma > 0$ if

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp(\sigma^2\lambda^2/2), \quad \forall \lambda \in \mathbb{R}.$$

A sub-Gaussian random variable X with parameter σ satisfies the following concentration bound:

$$\mathbf{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right), \quad \forall \epsilon \geq 0.$$

As suggested by the notation, the parameter σ plays a role similar to that of the standard deviation for Gaussian random variables. Examples of sub-Gaussian random variables with parameter σ include Gaussian random variables with standard deviation σ and bounded random variables supported on an interval of width 2σ. Therefore, by using sub-Gaussian noise terms, we encompass many possible distributions. In all cases, sub-Gaussianity assures us that the noise will be concentrated around its mean.

2.4 High-dimensional Problems

Now that we have presented our model, we reiterate the high-dimensional nature of the problem in more specific terms. In our model, with n products, $\mathbf{A}$ would be an $n \times n$ square matrix, and hence there would be n^2 unknown parameters to be estimated. Even with 50 products, a reasonable number for many product categories, there would be 2500 parameters. In order to estimate all of these parameters accurately, we expect to need to perform many experiments.

Table 1: Summary of notation

Term	Description
$\mathbf{A}$	A matrix capturing the substitution and complementarity effects – the element a_{ij} represents the effect on the demand for product i due to a 100% increase in the price of product j
$\mathbf{x}^t$	A vector of *treatment* prices
$\mathbf{x}^b$	A vector of *baseline* prices
$\mathbf{x}$	A decision vector, whose entries are percentage changes in price from the baseline
$\mathbf{w}$	The random error or noise vector
$\mathbf{q}^t$	The observed *treatment* demand
$\mathbf{q}^b$	The *baseline* demand, which is assumed to be known from the control condition
$\hat{\mathbf{A}}$	An estimate of the true matrix $\mathbf{A}$
n	The number of products
s	The number of experiments

Unfortunately, each experiment is costly to the firm in terms of not only time and resources needed to run it, but also opportunity costs. Therefore, our goal is to estimate the parameters accurately and make good decisions using as few experiments as possible.

Although we are faced with a difficult problem, our main insight is that even though there are many products, each one is likely to interact with only a small fraction of the remaining products. In terms of our model, this means that the $\mathbf{A}$ matrix is likely to have many entries equal to zero. Our main result shows that if $\mathbf{A}$ exhibits this sparse structure, we can greatly reduce the number of experiments needed to learn $\mathbf{A}$ and to find a good decision vector $\mathbf{x}$, even if the locations of the nonzero terms are not *a priori* known.

2.5 Summary of Baseline Model

Before we present our results, we first review the baseline model that we will be considering. Our demand model is given by the following equation:

$$\Delta\mathbf{q} = \frac{\mathbf{q}^t - \mathbf{q}^b}{\mathbf{q}^b} = \mathbf{A}\mathbf{x} + \mathbf{w}.$$

The functional form $\Delta\mathbf{q} = \mathbf{A}\mathbf{x} + \mathbf{w}$ is convenient for analytical tractability. However, our analysis does not place any limitations on how $\Delta\mathbf{q}$ is defined. Indeed, we could use different variations, including alternatives that ensure symmetry in the measures of demand increases and decreases. Table 1 summarizes the relevant terms of our model.

3 Estimating the Matrix A

In order to find an optimal set of firm actions, we will first estimate the substitute and complementary relationships between products, which are modeled by the matrix $\mathbf{A}$. In this section, we describe a general technique for estimating $\mathbf{A}$, introduce our sparsity assumptions, present bounds on the number of experiments needed to learn $\mathbf{A}$ accurately, and discuss our results.

3.1 Random Experimental Design

Our goal is to learn $\mathbf{A}$ as quickly as possible and so we would like to design experiments (i.e., $\mathbf{x}$ vectors) that give as much information as possible. One approach is to design decision vectors deterministically in order to maximize some orthogonality measure between decision vectors. However, because we do not make any assumptions about how the locations or values of the entries of $\mathbf{A}$ are distributed, for any deterministic design, there will be classes of $\mathbf{A}$ matrices for which the design is poor.

As an alternative, we use random experiments: the decision of how much to change the price of a particular product for a given experiment will be a random variable. Moreover, if we make these decisions independently across products and across experiments, we achieve approximate orthogonality between all of our experiments. By using randomization, we are able to take advantage of the extensive body of probability theory and prove that we can learn every element of $\mathbf{A}$ to high accuracy with high probability, for *any* $\mathbf{A}$ matrix. Next, we describe our estimation procedure in more detail.

3.2 Unbiased Estimators, Convergence, and Concentration Bounds

For each parameter a_{ij}, we define a statistic y_{ij} that is a function of the random decision vector and the resulting (random) observed demands. This statistic is therefore also a random variable, and we design it so that its mean is equal to a_{ij}. In other words, we find an unbiased estimator for each parameter.

If we perform many independent experiments and record the statistic y_{ij} for each one, the law of large numbers tells us that the sample mean of these statistics converges to the true mean, which is exactly the parameter a_{ij} that we are trying to estimate. This sample mean is a random variable, and its probability distribution will become more and more concentrated around a_{ij} as we collect more samples (i.e., perform more experiments). To get a sense of the speed of convergence, we calculate a bound on the concentration of the distribution around a_{ij} after each additional sample. This bound will in turn allow us to prove results on the number of experiments needed to achieve accurate estimates with high confidence.

3.3 Uniformly ϵ-accurate Estimates

Our goal is to learn the $\mathbf{A}$ matrix accurately to within a certain bound with high probability. To be precise, let $\hat{a}_{ij}$ be our estimate of a_{ij}, an arbitrary element in the matrix $\mathbf{A}$. We adopt

a conservative criterion, which requires

$$\mathbf{P}\left(\max_{i,j}|\hat{a}_{ij}-a_{ij}|\geq\epsilon\right)\leq\delta,$$

where $\epsilon>0$ is the tolerance in our estimates and $1-\delta\in(0,1)$ is our confidence. In other words, we would like the probability that our estimates deviate substantially from their true values to be low, no matter what the true $\mathbf{A}$ matrix is. Because of the maximization over all entries in the matrix, we require that every single entry meets this criterion. Hence, we refer to this as the *uniform ϵ-accuracy* criterion. This notion of error is known as "probably approximately correct" in the machine learning field, which also aims to learn accurately with high probability (see Valiant 1984).

Ideally we would like both ϵ and δ to be small so that we have accurate estimates with high probability, but in order to achieve smaller ϵ and δ, intuitively we would need to run more experiments to gather more data. Our first objective is to determine, for a given number of products n and fixed accuracy and confidence parameters ϵ and δ, how many experiments are needed to achieve those levels uniformly. This answer in turn tells us how the number of experiments needed scales with the number of products.

3.3.1 Interpretation and Discussion

As has been described, uniform ϵ-accuracy is an intuitive measure of accuracy. It is also a conservative measure because it requires every entry of $\mathbf{A}$ to be accurate. Alternatively, we can consider other criteria, such as bounding the root-mean-square error:

$$\mathbf{P}\left(\sqrt{\frac{1}{n^2}\sum_{i,j=1}^{n}(\hat{a}_{ij}-a_{ij})^2}\geq\epsilon\right)\leq\delta.$$

This is a relaxation of the uniform ϵ-accuracy criterion: if estimates $\hat{a}_{ij}$ satisfy uniform ϵ-accuracy, then they also satisfy the RMSE criterion. Therefore, any positive results on the speed of learning under uniform ϵ-accuracy also hold under weaker criteria, such as the RMSE criterion. Our results then give a worst-case upper bound, in the sense that the number of experiments required to achieve a weaker criterion would be no more than the number of experiments required to achieve the stricter uniform ϵ-accuracy criterion.[7]

3.4 Asymptotic Notation

In order to judge different learning models, we compare how many experiments are needed to achieve uniform ϵ-accuracy. Because we are interested in the regime where the number of

[7] A similar point can be made about the method used to design the experiments and estimate the parameters. Improvements on our random experimental design and our relatively simple comparisons of the treatment and control outcomes should lead to further improvements in the amount of information learned and therefore decrease the number of experiments required to achieve uniform ϵ-accuracy.

products is large, we focus on how quickly the number of experiments needed increases as the number of products increases. To capture the scale of this relationship, we use standard asymptotic notation (see Appendix B for a detailed description).

3.5 Estimation of General A Matrices

We first consider the problem of estimating general **A** matrices, without any assumptions of additional structure. Following the technique outlined in Section 3.2, our precise estimation procedure is the following:

1. Perform independent experiments. For each experiment, use a random, independent decision vector $\mathbf{x}$, where for each product, x_j is distributed uniformly on $[-\rho, \rho]$, where $0 < \rho < 1$. Observe the resulting vector of changes in demand $\Delta\mathbf{q}$.

2. For the tth experiment and for each a_{ij}, compute the statistic

$$y_{ij}(t) \triangleq \beta \cdot \Delta q_i \cdot x_j,$$

   where $\beta \triangleq 3/\rho^2$.

3. After s experiments, for each a_{ij} compute the sample mean

$$\hat{a}_{ij} = \frac{1}{s} \sum_{t=1}^{s} y_{ij}(t),$$

   which is an unbiased estimate of a_{ij}.

The following theorem gives a bound on the accuracy of this estimation procedure after s experiments.

Theorem 1 (Estimation accuracy with sub-Gaussian noise for general **A** matrices)**.** *Under Assumption 1, for any $\epsilon \geq 0$,*

$$\mathbf{P}\left(\max_{i,j} |\hat{a}_{ij} - a_{ij}| \geq \epsilon \right) \leq 2n^2 \exp\left\{ -\frac{s\epsilon^2}{\max_i 36\left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)} \right\}. \tag{5}$$

See Appendix C for the proof.

To ensure uniformly ϵ-accurate estimates with probability $1 - \delta$, it suffices for the right-hand-side of (5) to be less than or equal to δ. Therefore, with a simple rearrangement of terms, we find that s experiments are sufficient if s satisfies

$$s \geq \frac{\max_i 36\left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}{\epsilon^2} \log\left(\frac{2n^2}{\delta}\right).$$

The above bound tells us that if there is more noise (larger c) or if we desire more accurate estimates (smaller ϵ and δ), then more experiments may be required, which agrees with intuition. However, the term $\sum_{\ell=1}^{n} a_{i\ell}^2$ may be quite large and, as it is a sum of n quantities, may also scale with n. In that case, our estimation procedure may in fact require $O(n \log n)$ experiments in order to achieve uniform ϵ-accuracy, which can be prohibitively large.

3.6 Introducing Structure

The previous result allows for the possibility that with general $\mathbf{A}$ matrices, many experiments may be required to estimate the underlying parameters. Fortunately, we recognize that our problem may have an important inherent structure that allows us to learn the $\mathbf{A}$ matrix much faster than we would otherwise expect.

We consider three different types of structure on the matrix $\mathbf{A}$. In the following sections, we motivate these assumptions, state the number of experiments needed to learn $\mathbf{A}$ in each case, and interpret our results.

3.6.1 Bounded Pairwise Effects

Motivation: Our first assumption is based on the idea that a product can affect the demand for itself or for any other product only by some bounded amount. In other words, varying the price of a product cannot cause the demand for itself or any other product to grow or diminish without limit. In terms of our model, we can state the assumption precisely as follows.

Assumption 2 (Bounded pairwise effects)**.** *There exists a constant b such that for any n, any $n \times n$ matrix $\mathbf{A}$, and any pair (i, j), $|a_{ij}| \leq b$.*

This is our weakest assumption as we do not place any other restrictions on $\mathbf{A}$. In particular, we allow every product to have an effect on every other product. By not imposing any additional assumptions, we can use this variation of the problem as a benchmark to which we can compare our two subsequent variations. Since all elements of $\mathbf{A}$ may be nonzero, we refer to this as the case of "dense" $\mathbf{A}$ matrices.

Result: With this additional assumption, we show that our estimation procedure as described in Section 3.5 can learn all elements of $\mathbf{A}$ to uniform ϵ-accuracy with $O(n \log n)$ experiments.

Corollary 1.1 (Sufficient condition for uniformly ϵ-accurate estimation of dense $\mathbf{A}$)**.** *Under Assumptions 1 and 2, for any $\epsilon \geq 0$,*

$$\mathbf{P}\left(\max_{i,j} |\hat{a}_{ij} - a_{ij}| \geq \epsilon\right) \leq 2n^2 \exp\left\{-\frac{s\epsilon^2}{36\left(nb^2 + c^2/\rho^2\right)}\right\}.$$

Therefore, to ensure uniformly ϵ-accurate estimates with probability $1 - \delta$, it suffices for the number of experiments to be $O(n \log n)$.

This result also gives an upper bound on the number of experiments needed to learn the entries of $\mathbf{A}$, in the sense that with the best estimation method, the asymptotic scaling of the number of experiments needed to achieve uniform ϵ-accuracy will be no worse than $O(n \log n)$. However, this upper bound is again not practical as it suggests that in the worst case, the number of experiments needed may scale linearly with the number of products. Because we would like to keep the number of experiments small, we hope to achieve a sublinear rate of growth with respect to the number of products. Fortunately, this is possible if the $\mathbf{A}$ matrix is "sparse," as we discuss in the next section.

3.6.2 Sparsity

Motivation: Although a category may include many items, not all items will have relationships with one another. For example, varying the price of a nighttime cold remedy may not affect the demand for a daytime cold remedy.

Under our model of demand and cross-product elasticities, a pair of items having no interaction corresponds to the respective entry being zero in the $\mathbf{A}$ matrix. If many pairs of items have no relationship, then our $\mathbf{A}$ matrix will have many zero entries, which is referred to as a "sparse" matrix. In terms of our model, we express the assumption of sparsity as follows.

Assumption 3 (Sparsity)**.** *There exists an integer k such that for any n, any $n \times n$ matrix $\mathbf{A}$, and any i, $|\{j : a_{ij} \neq 0\}| \leq k$.*

For each row of $\mathbf{A}$, we bound the number of entries that are nonzero to be no more than k. Interpreting this in terms of products, for each product, we assume that there are at most k products (including itself) that can affect its demand. Note that we do not assume any knowledge of how these nonzero entries are distributed within the matrix. This is important as it means we do not need to know *a priori* which products have a demand relationship with one another and which do not.

Result: As long as the underlying matrix $\mathbf{A}$ exhibits this sparsity structure, we have the following result on the number of experiments needed to estimate $\mathbf{A}$ with uniform ϵ-accuracy using our estimation method.

Corollary 1.2 (Sufficient condition for uniformly ϵ-accurate estimation of sparse $\mathbf{A}$)**.** *Under Assumptions 1, 2, and 3, for any $\epsilon \geq 0$,*

$$\mathbf{P}\left(\max_{i,j} |\hat{a}_{ij} - a_{ij}| \geq \epsilon\right) \leq 2n^2 \exp\left\{-\frac{s\epsilon^2}{36\left(kb^2 + c^2/\rho^2\right)}\right\}.$$

Therefore, to ensure uniformly ϵ-accurate estimates with probability $1 - \delta$, it suffices for the number of experiments to be $O(k \log n)$.

This result shows that if the $\mathbf{A}$ matrix is sparse, the number of experiments needed scales on the order of $O(k \log n)$, instead of $O(n \log n)$ as for the case of dense $\mathbf{A}$ matrices. Thus, the number of experiments needed grows logarithmically (hence, sublinearly) in the number of products n and linearly in the sparsity index k. As long as k does not increase too quickly with n, this may be a significant improvement over $O(n \log n)$. As anticipated in the introduction, sparsity can yield much faster learning. The gap between a theoretical requirement of $O(k \log n)$ and a theoretical requirement of $O(n \log n)$ experiments could be dramatic for practical purposes in settings with a large number of products, and therefore in estimation problems with a large number of parameters. Of course this requires that k does not grow too quickly with n. We will investigate this possibility in Section 4.

By thinking about the amount of abstract "information" contained in a sparse matrix as opposed to in a dense matrix, we can gain some intuition as to why a sparse matrix is easier

to estimate. When trying to learn a model, if we know that the true model lies in a restricted class of possible models, then we expect to be able to learn the true model faster than if no such restrictions were known. Our assumptions of sparsity effectively reduce the universe of possible $\mathbf{A}$ matrices in this manner. If $\mathbf{A}$ could be any $n \times n$ matrix, then for each row of $\mathbf{A}$, there would be on the order of n bits of unknown information (i.e., a constant number of bits for the value of each entry in the row). On the other hand, if we knew that the row has only k nonzero entries, there would instead be on the order of k bits of unknown information (i.e., a constant number of bits for the value of each *nonzero* entry in the row). There would also be uncertainty in the location of the nonzero entries. There are $\binom{n}{k}$ ways of choosing k entries out of n to be the nonzero ones, and therefore there are $\binom{n}{k}$ possible locations of the nonzero entries within the row, which can be encoded as an additional $\log_2 \binom{n}{k}$ bits of unknown information, which is approximately of order $O(k \log n)$ bits. Based on these rough calculations, we can see that knowing that a matrix is sparse with only k nonzero entries reduces the degrees of freedom and amount of uncertainty and therefore allows for faster estimation.

3.6.3 Bounded Influence (Weak Sparsity)

Motivation: Corollaries 1.1 and 1.2 are both based on the intuition that the substitution and complementarity effects between products are bounded. This was done through placing hard bounds on the magnitude of *each* pairwise effect (i.e., the magnitude of each element of $\mathbf{A}$) and by limiting the number of possible relationships a product can have (i.e., the number of nonzero elements in each row of $\mathbf{A}$).

An alternative approach, in the same spirit, is instead to bound the aggregate effect on each product's demand due to all price variations. The intuition here is that although there may be many products, the demand for any individual product cannot be swayed too much, no matter how many other products there are or which products' prices are varied. This can be thought of as a "weak" sparsity assumption: we do not assume that many elements of $\mathbf{A}$ are zero; instead we assume that the overall sum across any row of $\mathbf{A}$ stays bounded. We express this assumption in terms of our model as follows.

Assumption 4 (Bounded influence)**.** *There exists a constant d such that for any n and any $n \times n$ $\mathbf{A}$ matrix, the following inequality is satisfied for every i:*

$$\sum_{j=1}^{n} |a_{ij}| \leq d.$$

As another interpretation, Assumption 3 can be thought of as bounding the ℓ_0 "norm" of the rows of $\mathbf{A}$: $\|\mathbf{a}_i\|_0 \leq k$. Assumption 4 above can be thought of as a relaxation that instead bounds the ℓ_1 norm of the rows of $\mathbf{A}$: $\|\mathbf{a}_i\|_1 \leq d$.

Result: Using similar analysis, we show that the number of experiments needed to achieve uniform ϵ-accurate estimation under the assumption of bounded influence is on the order of $O(d^2 \log n)$.

Corollary 1.3 (Sufficient condition for uniformly ϵ-accurate estimation under bounded influence)**.** *Under Assumptions 1 and 4, for any $\epsilon \geq 0$,*

$$\mathbf{P}\left(\max_{i,j} |\hat{a}_{ij} - a_{ij}| \geq \epsilon\right) \leq 2n^2 \exp\left\{-\frac{s\epsilon^2}{36\left(d^2 + c^2/\rho^2\right)}\right\}.$$

Therefore, to ensure uniformly ϵ-accurate estimates with probability $1-\delta$, it suffices for the number of experiments to be $O(d^2 \log n)$.

The above result shows that even with a weaker sparsity condition, where we allow all parameters to be nonzero, we are still able to achieve an order of growth that is logarithmic in the number of products. Note that if Assumptions 2 and 3 are satisfied with constants k and b, respectively, then Assumption 4 will also be satisfied with $d \triangleq kb$, and so the bounded influence assumption can subsume the combination of bounded pairwise effects and sparsity assumptions. However, using the more general bounded influence assumption to capture sparsity leads to a weaker result because it does not leverage all of the structural details of the sparsity assumption. Specifically, with $d = kb$, Corollary 1.3 would give a scaling of $O(k^2 \log n)$ for learning a k-sparse $\mathbf{A}$ matrix (where the dependence on b has been suppressed), which is slower than the scaling of $O(k \log n)$ given by invoking Corollary 1.2.

3.7 Lower Bound

The previous results provide upper bounds on the number of experiments needed for accurate estimates. For example, in the case of sparsity, using our estimation method, no more than $O(k \log n)$ experiments are needed to achieve uniform ϵ-accuracy. However, these results do not tell us whether or not there exists another estimation method which requires even fewer experiments. Given our demand model, the bounds on the allowable price variations, and the noise in the data, information theory tells us the maximum amount of information about the a_{ij}'s that can be learned from a single experiment. This fundamental limit in the "value" of each experiment in estimating the $\mathbf{A}$ matrix then allows us to calculate a lower bound on the number of experiments required. We do not actually need to develop a specific estimator that achieves this lower bound, but we know that no estimator can do better than this lower bound.

For the special case of i.i.d. Gaussian noise, we now present such a lower bound on the number of experiments needed, which shows that no matter what estimation procedure we use, there is a minimum number of experiments needed to achieve uniform ϵ-accuracy. The only requirement we impose on the estimation procedure is that it relies on experiments with bounded percentage price changes. The bounds we impose on the percentage price changes can be justified by practical considerations: the natural lower bound on the price changes comes from the fact that the prices cannot be negative, while the upper bound on the percentage changes captures that the manager of a store is likely to be opposed to dramatic price increases for the purposes of experimentation.

Theorem 2 (Necessary condition for uniform ϵ-accurate estimation under sparsity with Gaussian noise)**.** *For $\lambda > 0$, let*

$$\mathcal{A}_{n,k}(\lambda) \triangleq \left\{ \mathbf{A} \in \mathbb{R}^{n \times n} : |\{j : a_{ij} \neq 0\}| = k, \forall i = 1, \ldots, n; \min_{i,j:a_{ij} \neq 0} |a_{ij}| \geq \lambda \right\}$$

be the class of $n \times n$ $\mathbf{A}$ matrices whose rows are k-sparse and whose nonzero entries are at least λ in magnitude. Let the noise terms be i.i.d. $\mathcal{N}(0, c^2)$ for some $c > 0$. Suppose that for some $\epsilon \in (0, \lambda/2)$ and $\delta \in (0, 1/2)$, we have an estimator that

(a) experiments with percentage price changes $x \in [-1, \tilde{\rho}]$, for some $\tilde{\rho} \geq 1$ (i.e., the price for each product cannot fall below 0 and cannot increase by more than $100 \cdot \tilde{\rho}\%$), and

(b) for any $\mathbf{A}$ matrix in $\mathcal{A}_{n,k}(\lambda)$ achieves uniformly ϵ-accurate estimates with probability $1 - \delta$.

Then, the number of experiments used by the estimator must be at least

$$s \geq \frac{k \log(n/k) - 2}{\log(1 + k^2 \lambda^2 \tilde{\rho}^2 / c^2)}.$$

The proof is given in Appendix D.

As the number of products grows, the asymptotically dominant scaling terms are

$$s \geq \Omega\left(\frac{k \log(n/k)}{\log k} \right).$$

Since $\log k$ is small compared to k and $\log n$, we have an essentially matching lower bound to the $O(k \log n)$ upper bound in Corollary 1.2, which shows that our estimation procedure achieves close to the best possible asymptotic performance.

3.8 Discussion

The previous results demonstrate the power of sparsity in multiple flavors. Without any assumptions on the structure of the problem, the number of experiments needed may grow linearly with the number of products. For our target regime of large numbers of products, this leads to a solution that appears to be practically infeasible. However, by recognizing the inherent properties of the problem, we show that even by using randomly designed experiments we are able to learn $\mathbf{A}$ in a number of experiments that scales only logarithmically with the number of products. With a large number of products, the difference between linear and logarithmic is tremendous: for $n = 100$, $\log(100) \approx 4.6$. This gives hope that we can indeed learn the $\mathbf{A}$ matrix in a practically feasible number of experiments.

While our findings help reveal how many experiments are required, it is also helpful to ask how many experiments are feasible. When firms are using field experiments to set policy (rather than academics using them to test theories) we have found they are often willing to run a rather large number of experiments. The answer will clearly depend upon the nature

of the firm's actions and the particular setting. Varying advertising or pricing decisions in online or direct marketing settings can often be implemented at low cost, making it feasible to implement hundreds or even thousands of experiments. For example, Capital One reportedly implements tens of thousands of randomized field experiments each year. In traditional retail formats, the cost of making in-store changes is generally higher, and randomization must often occur at the store level rather than the individual customer level (introducing an additional source of measurement error). However, even in traditional retail settings, firms with multiple locations can implement a large number of experiments in different samples of stores to test pricing, product placement, and other merchandising decisions. For example, one of the authors has worked with a large bricks and mortar retailer who was quickly able to run 200 between-store pricing experiments to decide how to price private label items when national brands are promoted. Documented examples of high-volume experimentation in traditional retail settings include Bank of America varying actions between bank branches and Harrah's varying a wide range of practices across its casinos. In other settings, implementing field experiments is more challenging. For example, when deciding how to manage a distribution network, a firm may be limited to only a handful of experiments every few years, as these experiments will tend to disrupt existing relationships and require extended periods to observe the outcome.

3.9 Other Marketing Decisions

Besides setting prices, firms make many other types of marketing decisions, including which products to advertise or promote. Although our model and analysis have focused on pricing decisions, the model can easily be adapted to advertising or promotion decisions. As with setting prices, promoting a product will (for most products) increase its demand. The substitution and complementarity effects between products will also carry over to promotion decisions. Therefore, we can again use a matrix $\mathbf{A}$ to represent the own- and cross-product elasticities and a vector $\Delta\mathbf{q}$ to represent the percentage change in demand for each product. However, some modifications are required to extend the model to promotion applications.

If we interpret the decision to advertise or promote a product as a binary decision, then the decision variables become

$$\tilde{x}_j = \begin{cases} 1, & \text{if } j \text{ is promoted,} \\ 0, & \text{if } j \text{ is not promoted.} \end{cases}$$

For ease of exposition (and without loss of generality), we will assume that there are no promotions in the control condition. We can then model the percentage change in demand in response to the promotion decisions as

$$\Delta\mathbf{q} = \frac{\mathbf{q}^t - \mathbf{q}^b}{\mathbf{q}^b} = \mathbf{A}\tilde{\mathbf{x}} + \mathbf{w}.$$

In this model, we capture in the $\mathbf{A}$ matrix own- and cross-product promotion responses. This model retains the same form as in (4), where $\tilde{\mathbf{x}}$ takes the place of $\mathbf{x}$.

Given that the model under promotion decisions has the same form as the model under pricing decisions, we can apply a modified form of our estimation procedure to obtain similar results. Specifically, instead of making continuous pricing decisions, we instead make 0/1 Bernoulli decisions for each $\tilde{x}_i$ in the promotion setting. This is essentially the same setup and we can again find estimators for each a_{ij} such that Theorem 1 holds (with slightly different constants). Therefore, we would still be able to achieve uniformly ϵ-accurate estimation with $O(k \log n)$ experiments under sparsity and $O(d^2 \log n)$ experiments under bounded influence.

4 Estimating Sparsity

In order for a retailer to evaluate whether it is feasible to make pricing decisions using field experiments, the retailer needs an estimate of the sparsity parameter (k or d). In this section, we describe two approaches for estimating these parameters: 1) from a "pilot" set of experiments and 2) from historical data. Under both approaches, we use what is essentially a model selection approach. We divide the data into calibration and validation sub-samples. We then repeatedly estimate the $\mathbf{A}$ matrix using the calibration sub-sample for different values of the sparsity parameter, and we choose the sparsity parameter for which the estimated $\mathbf{A}$ matrix has the best fit with the validation sub-sample.

Different variants of this general approach are available, including different measures of "goodness-of-fit" of the validation sub-sample. We can also use different approaches to cross-validate, including m-fold cross validation where we randomly split the data into m buckets and rotate which of the buckets we treat as the validation sample. In the discussion below, we describe the two approaches more formally and present results of both simulations and empirical analysis to illustrate their performance.

In addition to describing how to estimate k and d, our analysis in this section also has a second purpose. Although we have shown that sparsity and weak sparsity ensure that the number of experiments required to obtain accurate estimates grows at a logarithmic rate with n, we must also consider how the sparsity parameters (k and d) grow with n. If k and d grow quickly with n, then the $O(k \log n)$ and $O(d^2 \log n)$ growth rates will again mean that it may be infeasible to use experiments to set prices in large categories.

4.1 Methodology

Let $\mathbf{a}_i$ be the (unknown) $1 \times n$ row vector of elasticities for the ith product. Suppose we have s data points: $\Delta\mathbf{q}_i$ is a $1 \times s$ vector of changes in demand for the ith product, and $\mathbf{X}$ is an $n \times s$ matrix of pricing decisions. For some value τ, we solve the following optimization problem (the "Lasso"), which looks for the $\mathbf{a}_i$ that best fits the data but is still constrained to be "sparse":

$$\begin{array}{rl} \min_{\mathbf{a}_i} & \|\Delta\mathbf{q}_i - \mathbf{a}_i'\mathbf{X}\|_2^2 \\ \text{s.t.} & \|\mathbf{a}_i\|_1 \leq \tau. \end{array}$$

Alternatively, we can express the problem as the following:

$$\min_{\mathbf{a}_i} \ \|\Delta\mathbf{q}_i - \mathbf{a}_i'\mathbf{X}\|_2^2 + \lambda\|\mathbf{a}_i\|_1. \tag{6}$$

Here, τ and λ are tuning parameters that control the level of sparsity of $\mathbf{a}_i$. For each choice of the tuning parameters, we obtain one solution, $\hat{\mathbf{a}}_i$, to the optimization problem. To assess the quality of each solution, we cross-validate it using the given data and select the one that gives the lowest cross-validation error as the best solution. From this best solution, we recover its "sparsity" and propose that measure as an estimate of the true level of sparsity.

Although this methodology focuses on a single product/row i, the same procedure can be performed on each row independently, with the same number of experiments, to obtain estimates of k or d for each row. This procedure then gives us even finer-grained estimates, not just a single k or d bound for the entire $\mathbf{A}$ matrix. Our model calls for a k or d that bounds the sparsity of the entire matrix. Therefore, to arrive at an overall estimate of k or d for the entire matrix, we take the maximum over the individual row estimates. Note that this approach is valid for either hard sparsity (k) or bounded influence (d). We will test the methodology on both cases.

4.2 Pilot Experiments

In order to perform the procedure described in the previous subsection, we first require some data. One possible source of data is a set of "pilot" experiments: a relatively small sequence of pricing experiments and corresponding observed demand.

4.2.1 Simulation

In practice, managers can conduct actual pilot experiments and collect the necessary data. In this subsection, we first consider generating synthetic experimental data using distributions seeded from prior field experiments and test our methodology using a simulation, which proceeds as follows:

1. Choose fixed values of n and d (or k) and generate the true $\mathbf{A}$ matrix randomly from the seed distributions. Choose a fixed value of σ, the standard deviation of the normal error term $\mathbf{w}$. These parameters are not used in the estimation.

2. For any given s:

   (a) Randomly generate $\mathbf{x}$ and $\mathbf{w}$ for s experiments and calculate $\Delta\mathbf{q}$.

   (b) For a range of λ's, find the optimal solutions to (6).

   (c) Perform five-fold cross-validation[8] on the solutions to identify the one with the lowest cross-validation error; call this $\mathbf{a}_i^*$. (Figure 1 illustrates the cross-validation process.)

[8]Split the data set into five buckets. Estimate $\mathbf{a}_i$ on data from four buckets and cross-validate on the fifth. Rotate and do this for all five buckets and calculate the average error.

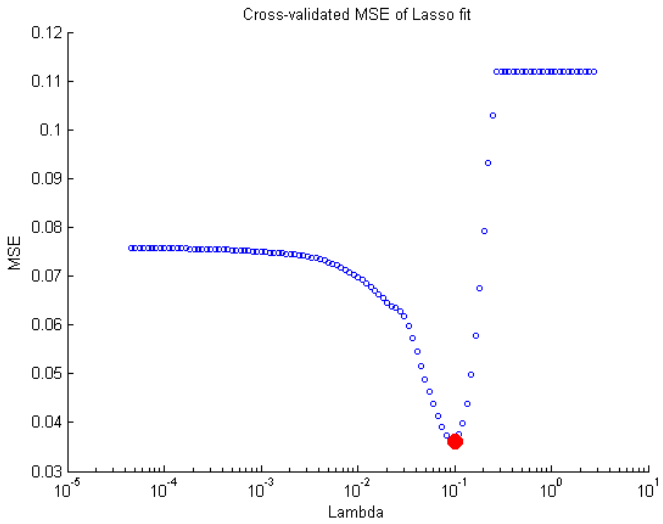


Figure 1: An example of the result of five-fold cross-validation. The value of λ highlighted in red gives the lowest cross-validation error. Large values of λ (to the right) heavily penalize nonzero entries, resulting in the zero vector as the solution, which does not fit the data well. As λ is lowered, we begin to get some nonzero entries in the solution, which provide a better fit of the data. However, as λ becomes even smaller, past the value marked in red, we obtain dense solutions that tend to overfit, resulting in a higher cross-validation error.

(d) Calculate $\|\mathbf{a}_i^*\|_1$ and $\|\mathbf{a}_i^*\|_0$. For the latter, we count only those entries that are above a certain threshold (set at 0.01) in magnitude.

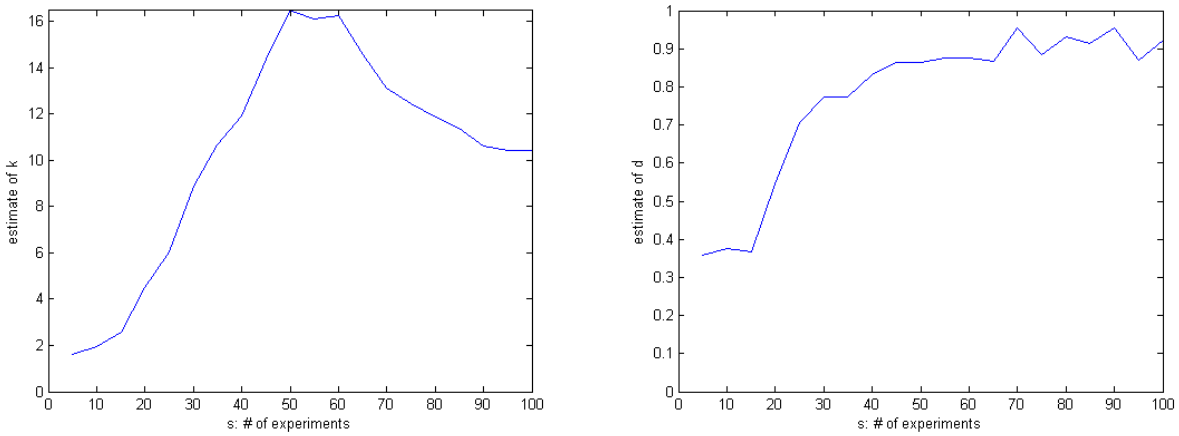
(e) For each s, replicate this 100 times and average the results. Propose the averaged values of $\|\mathbf{a}_i^*\|_1$ and $\|\mathbf{a}_i^*\|_0$ as estimates of d and k, respectively.

3. Plot the estimates of d and k versus a range of values of s, giving a sense of how many experiments are needed to obtain an accurate estimate of the level of sparsity.

Simulations were performed using the following parameters:

$n = 1000$ $s = 5, 10, 15, \ldots, 100$ $\lambda = e^{-10}, e^{-9.9}, \ldots, e$	
$k = 10, b = 1$	$d = 1$
$\sigma = 0.1*$(mean row sum)	$\sigma = 0.1d$

As Figure 2 illustrates, our methodology provides reasonable estimates of k and d with relatively few experiments. These results suggest that using pilot experiments can indeed provide initial estimates of k and d. Knowing these sparsity parameters, we then have a sense of the feasibility of using our main methodology to estimate $\mathbf{A}$.



(a) Estimating k

(b) Estimating d

Figure 2: Plot of the estimates of k and d versus the number of experiments, s. The estimates are near the true values of $k = 10$ and $d = 1$, even with relatively few experiments.

4.3 Empirical Analysis

Running 80 to 100 pilot experiments is not without cost, and so ideally a firm would like to be able to estimate k and d using its existing data. One possibility is to use historical variation in prices to estimate these parameters. Our proposed cross-validation method can be easily adapted to do so.

A limitation of using historical variation in control variables is that this past variation is often not random. This has raised concerns that the resulting elasticity estimates may be biased (Villas-Boas and Winer 1999). However, these limitations are less relevant in this setting, where we are unconcerned about bias in elasticity estimates and instead merely seek a preliminary estimate of k and d.

We use 195 weeks of historical data from a chain of 102 convenience stores, describing prices and unit sales of products in the cold remedies category. The number of products sold in each story varies, due primarily to differences in the square footage size of each store (larger stores offer wider product ranges). We will exploit this variation to illustrate how our estimates of k and d vary with the number of items in the category (n).

4.3.1 Setup

We begin with the 195 weeks of sales data from 102 stores, which we then group into 48 four-week periods in order to reduce the amount of noise in the data. We focus on a specific category (cold remedies) and perform the following procedure for each store independently:

1. If a product is not sold in a given period, no data is available for that product during that period, which means that we do not know the retail price for that product during that period. We fill in this price data by linearly interpolating between the prices for

that product during the two most adjacent periods for which we do have data.

2. However, we know that if no data is available, the quantity sold during that period is zero.

3. After this processing, we have a complete set of sales and price data for each product, for each of the 48 four-week periods.

4. For each product i, we compute the average quantity sold per period and the average price, over the 48 periods. These will serve as the baseline demand (q_i^b) and price levels (x_i^b), respectively.

5. To further reduce noise, we consider only those products which (i) sold over a certain threshold of units per period on average, (ii) sold at least one unit during the first four periods and last four periods (to ensure they were not introduced or discontinued during the middle of the 195 weeks), and (iii) had variations in prices above a certain threshold over the course of the 48 periods.

6. We collect all products that do not pass through the above filter and combine them into a single aggregated "product", which is included together with all other products in the analysis that follows.

7. We calculate category-level seasonality factors for each period, which are used to deseasonalize the raw demand quantities.

8. Using the price data and the (deseasonalized) sales data for each period, we then calculate their percentage change from the previously established baseline levels, as indicated by our model.

9. Equipped with $\Delta\mathbf{q}$ and $\mathbf{x}$, we then use these as input to the Lasso optimization program (6):

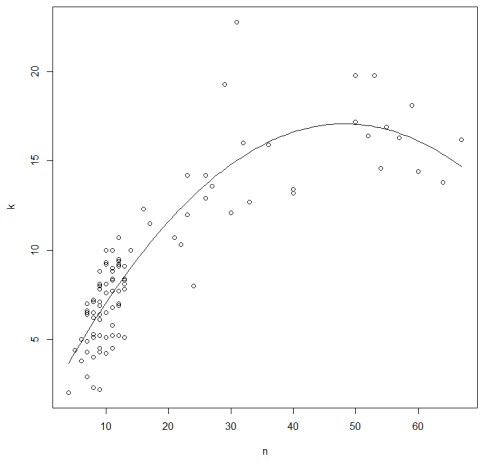
   (a) Lasso estimates vectors, so we estimate $\mathbf{A}$ row-by-row.

   (b) For each row i, we try a sequence of λ parameters and perform five-fold cross-validation in order to identify the value of λ that gives the lowest cross-validation error; call this estimate $\hat{\mathbf{a}}_i^*$. Calculate $\|\hat{\mathbf{a}}_i^*\|_1$ and $\|\hat{\mathbf{a}}_i^*\|_0$ as estimates of k_i and d_i for row i.

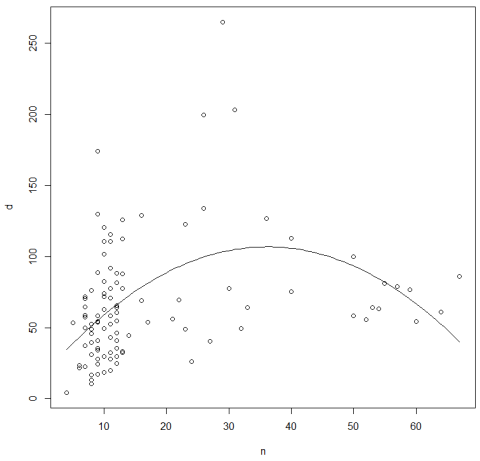
   (c) Because k and d are sparsity parameters for the complete $\mathbf{A}$ matrix, we take the maximum over all of the rows' k_i and d_i to obtain the overall estimate of k and d.

   (d) For robustness, we repeat this entire procedure ten times and average the results.

10. By performing this analysis for each store, we obtain 102 pairs of (n, k) and (n, d) data points, which give us a relationship between the number of products and the sparsity index.



(a) Estimating k



(b) Estimating d

Figure 3: Plots of n versus estimated k and d, including the quadratic fit. Sales threshold: one unit per period on average, standard deviation of price variations threshold: 0.08.

	Coefficient	Estimate	Std. Error	t value
Estimating k	(Intercept)	1.118	0.607	1.843
	1st-order term	0.661	0.059	11.169
	2nd-order term	-0.007	0.001	-7.461
Estimating d	(Intercept)	15.280	11.696	1.306
	1st-order term	5.068	1.142	4.438
	2nd-order term	-0.070	0.018	-3.959

Table 2: Summary of quadratic fit models for four-week periods with a sales threshold of one unit sold per period on average and a minimum standard deviation of price variations of 0.08. The second-order coefficients are negative and significant for both k and d.

11. We fit a quadratic model and verify whether the second-order coefficient is negative and significant, indicating that the sparsity index does not increase linearly with the number of products.

4.3.2 Results

Figure 3 presents the estimates of k and d across all of the stores (each point represents the estimates for a single store). Recall that the number of items in each category varies across the stores, which allows us to investigate the relationship between the sparsity parameters and n. The figures also show the fitted quadratic relationships between the data points, which allow us to evaluate whether the growth in the sparsity parameters is slower than linear. In Table 2, we report the results of these quadratic fit models.

The estimates of k reveal a relatively distinct pattern: the estimates grow with n but the growth rate is slower than linear. In the fitted quadratic equation the quadratic term is negative and highly significant. We can speculate on the reasons for this. It is possible that customers eliminate products from their consideration sets that do not share certain attributes. For example, on a specific trip customers may focus only on nighttime cold remedies or daytime cold remedies. If this is the case, then introducing a new daytime product may not increase k (which is an upper bound on the number of interdependent products) because it only affects demand for the subset of items that share that attribute (i.e., daytime remedies). It was this type of behavior that Tversky (1972) anticipated when proposing that customers eliminate alternatives by aspects.

The estimates of k are relatively small (around fifteen) even in large categories. This suggests that in the cold remedies category, the matrix of cross-price elasticities is sufficiently sparse to make estimation using field experiments feasible. This demonstrates the feasibility of using historical data to obtain initial estimates of k to evaluate when a firm can use experiments to set prices. The data that we have used is readily available to most retailers. Notably, because we obtain estimates of the sparsity parameters for each category in each store, it does not require that retailers have a large number of stores (although having many stores obviously makes experimentation easier).

Notice that for many of the stores we observe only approximately ten items in the cold remedies category. This reflects two things: both the relatively small size of these stores, and the screening of products based on their sales volumes and the level of price variation. Because of this screening, we estimate the findings using the items that have the highest sales volumes and the largest price variation (the filtered items are combined into a single "other" item). To evaluate the robustness of our findings, we have repeated the analysis for different minimum sales and price variation thresholds. We also replicated the findings when grouping the data into ten-week periods. The findings replicate the pattern of results reported in Figure 3 and Table 2. In all of these combinations, the quadratic coefficient regressing k on n is negative and highly significant.

We also report the estimates of d. The fitted quadratic function indicates that the growth of d with n is also sublinear.[9] However, the findings reveal a much less distinct pattern. Notably some of the estimates of d are very large (exceeding 100). Moreover, while our estimates of k are relatively robust, the estimates of d are much less robust and are sensitive to variation in the filtering parameters. One interpretation is that within the cold remedies category, the weak sparsity structure is not sufficient to make it feasible to use experiments to set prices. A second interpretation is that our estimation procedure is not accurate enough to provide reliable estimates of d. Further investigation reveals why it is more challenging to estimate d than k. Recall that k is a count of the number of binary relationships between products, while d is a measure of the aggregate effect size. If there is

[9]In the case of d, sublinear growth could simply reflect customer loyalty or state dependence (see for example Dubé et al. 2008, Erdem 1996, Keane 1997, Seetharaman et al. 1999, Anderson and Simester 2013). If even just a subset of customers is loyal to an existing product (or brand), then the introduction of additional products will have a bounded impact on sales of the existing products. The more customers who are loyal, the less growth we expect in d as n grows.

relatively little variation in the price measure, but considerable noise in the sales measure, then the estimated effect sizes can be large. This makes the estimates of d much more sensitive to noise in the sales measure than the estimates of k.

Summary: We have described how to estimate the sparsity parameters k and d either from a pilot set of experiments or from historical data. Using a sample of actual data from the cold remedies category, we were able to obtain reliable estimates of k. These estimates revealed that k increases with n but the growth is sublinear. Changing the price of an item within the cold remedies category appears to affect demand of no more than fifteen other items, and so the $\mathbf{A}$ matrix (of sales responses) is sparse. The findings illustrate a practical method that managers can use to evaluate whether a product category has a favorable structure to make it feasible to set category prices using field experiments.

5 Simulations

The theoretical results presented so far have focused on the speed of learning. In this section, we present the results of simulations that confirm the relevance of the theoretical asymptotic bounds.

5.1 Simulation Setup

To ensure that our simulations use realistic parameters, we initialize them using data from a large-scale pricing experiment that was conducted for another purpose (Anderson et al. 2010). The experiment was implemented at a large chain of stores that sells products in the grocery, health and beauty, and general merchandise categories. Eighteen of the chain's stores participated in the study, in which prices were experimentally manipulated on 192 products for seventeen weeks, with the treatments randomly rotated across the eighteen stores (see Anderson et al. 2010 for additional details). From this study, we obtained distributions for the diagonal and off-diagonal entries of the $\mathbf{A}$ matrix.

We also specify a collection of parameters that define the simulation: the number of products (n), structural parameters for the $\mathbf{A}$ matrix (b, k, and d), the noise distribution parameter (c), and the error criteria (ϵ and δ). We refer to these parameters together as the simulation definition. In order to compare the dense and sparse cases, we first generate a full matrix using these two distributions for the dense case and then randomly set all but k entries in each row to zero for the associated sparse case. Instead of selecting an arbitrary value for k, we use the empirical results from Section 4.3.2: for any given n, we use the quadratic fit (plus some additive noise) to calculate the associated value of k.

5.2 Estimation of A

Given an $n \times n$ matrix $\mathbf{A}$ generated using the distributions described above, along with a definition of parameters, we can then use the procedure described in Section 3.5 to estimate

A.

To simulate one experiment, we generate a random vector $\mathbf{x}$ with values sampled from the uniform distribution on $[-\rho, \rho]$, and random noise variables w_i uniformly in the interval $[-c, c]$. Using the true underlying $\mathbf{A}$ matrix, we then calculate the vector of percentage changes in demand $\Delta\mathbf{q} = \mathbf{A}\mathbf{x} + \mathbf{w}$ and the statistics y_{ij}, which are unbiased estimators of the a_{ij}'s. As we perform more experiments, we keep a running sum of the y_{ij}'s and compute the sample mean to obtain our estimate $\hat{a}_{ij}$. By comparing these estimates to the true $\mathbf{A}$ matrix, we can calculate the maximum absolute error across all entries: $\max_{i,j} |\hat{a}_{ij} - a_{ij}|$.

Since our criterion of uniform ϵ-accuracy requires the probability that the maximum absolute error is less than ϵ to be at least $1 - \delta$, we run 100 parallel sequences of experiments. Each sequence is essentially an independent instance of the estimation procedure. We incrementally generate more experiments for each sequence, compute updated estimates, and calculate maximum absolute errors. After any number of experiments, each sequence therefore has its own set of estimates and corresponding maximum absolute error. We say that we have achieved uniform ϵ-accuracy when at least a $1 - \delta$ fraction of the sequences have maximum absolute errors that are less than or equal to ϵ.

5.3 Estimation Performance

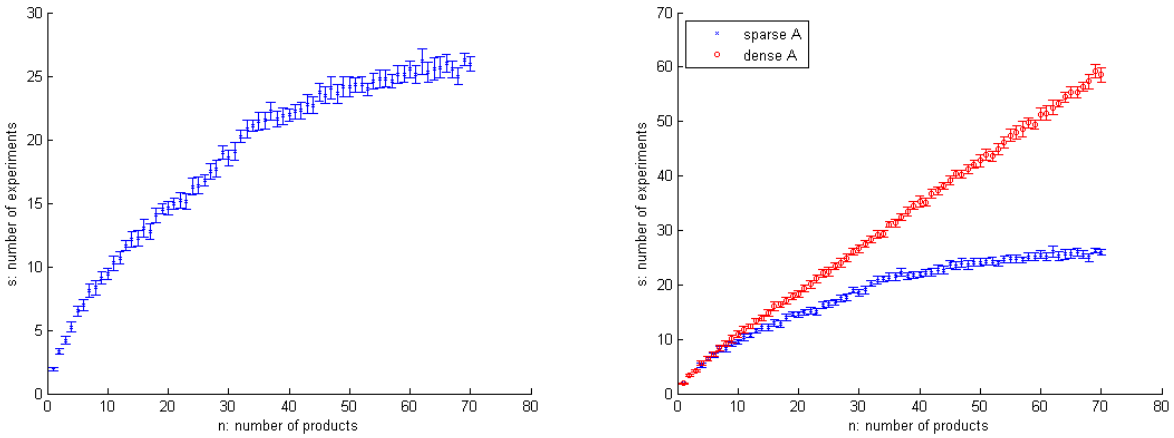
Using the above procedure, we can simulate the number of experiments needed to achieve uniform ϵ-accuracy for any given simulation definition. Because we are interested in how the number of experiments needed scales with the number of products, we fix a particular definition of parameters (except for n) and generate a sequence of matrices $\{\mathbf{A}_n\}$ that increase in size. For each matrix $\mathbf{A}_n$, we determine the number of experiments needed to achieve uniform ϵ-accuracy. For robustness, we replicate the entire simulation 20 times and, for each n, calculate 95% confidence intervals for the number of experiments needed.

In the case of sparse matrices, the resulting plot (Figure 4a) exhibits the logarithmic scaling predicted by our theoretical results. As the number of products grows, the number of experiments required grows much more slowly than the linear benchmark. Additional products require fewer and fewer additional experiments to achieve accurate estimates. On the other hand, Figure 4b shows that the dense case requires many more experiments than the sparse case to achieve the same level of estimation accuracy.[10]

6 Conclusion

While many firms lack the capabilities to estimate sophisticated econometric models, almost any firm can compare the results between experimental treatment and control groups. We have investigated whether conducting these simple comparisons can help firms improve their profits even as the complexity of the problem grows. In particular, we consider settings where actions taken to impact the sales of one product tend to spill over and also affect

[10]The results for the "sparse" case in Figure 4b are identical to the results in Figure 4a (the only difference is the change in the scale of the y-axis).



(a) Sparse **A** matrix (b) Comparison of sparse and dense **A** matrices

Figure 4: When the **A** matrix is sparse, the number of experiments needed to achieve uniform ϵ-accuracy grows only logarithmically with the number of products. When the **A** matrix is dense, the number of experiments needed to achieve uniform ϵ-accuracy grows at least linearly with the number of products. Comparing the cases of sparse and dense **A** shows that learning is much faster in the sparse case. The bars represent 95% confidence intervals. Parameters used for this plot: $\rho = 0.5$, $c = 0.5$, $b = 5$, $\epsilon = 1.5$, $\delta = 0.1$.

sales of other products. As the number of products n grows, the number of parameters to estimate grows as $O(n^2)$. This suggests that the number of experiments required to estimate these parameters will quickly grow beyond what is feasible.

However, we show that if the category exhibits a favorable structure, then firms can learn these parameters accurately using a relatively small number of experiments. We investigate two such structures. The first is sparsity, in which any one product can be affected by at most k products. An important point is that we do not need to know which specific products affect that one product's demand, only that there is a limit to how many such products there are. Given this restriction, the number of experiments required to estimate the matrix of parameters drops from $O(n \log n)$ to $O(k \log n)$.

We also describe a second restriction that yields similar results. Rather than limiting the number of products that can affect any one product, it may be more appropriate to restrict how much the total percentage change in sales of one product can be affected by actions on all of the products. As long as there is a limit to the aggregate magnitude of these interactions, then we can again achieve relatively quick improvements in parameter estimates with a feasible number of experiments.

Our findings provide guarantees about the rate of learning from experiments. These guarantees are obtained using randomized experiments and simple comparisons of outcomes between treatment and control conditions. Firms may increase the rate of learning by optimizing the experimental designs and/or using more sophisticated analyses to estimate the parameters. While our guarantees will continue to hold under these alternative approaches,

future research may investigate the extent to which the bounds can be improved in these circumstances.

We have framed our findings by focusing on the category pricing decisions. However, the results can be easily extended to other marketing decisions in which actions targeted at an individual product spill over to affect other products as well. In the context of learning demand elasticities, we have extended our findings to selecting which products to promote. Other applications could include the allocation of sales force resources across products or the focus of future investments in product development. It may also be possible to extend the results to settings in which marketing actions targeted at one customer (or group of customers) also impact the decisions of other customers. Spillovers between customers may arise when customers can observe the decisions of other customers, or when their decisions depend on the recommendations of other customers. Extending our results to these forms of externalities may present fertile opportunities for future research.

A Multiplicative Model

Consider an alternative model of the following form:

$$\Delta q_i = x_1^{a_{i1}} x_2^{a_{i2}} \cdots x_n^{a_{in}} w_i.$$

Taking logs, we obtain

$$\begin{aligned}\log(\Delta q_i) &= a_{i1}\log(x_1) + a_{i2}\log(x_2) + \cdots + a_{in}\log(x_n) + \log(w_i) \\ &= \sum_{\ell=1}^{n} a_{i\ell}\log(x_\ell) + \log(w_i).\end{aligned}$$

By defining $\Delta\tilde{q}_i \triangleq \log(\Delta q_i), \tilde{x}_\ell \triangleq \log(x_\ell), \tilde{w}_i \triangleq \log(w_i)$, we can rewrite the above as

$$\Delta\tilde{q}_i = \sum_{\ell=1}^{n} a_{i\ell}\tilde{x}_\ell + \tilde{w}_i,$$

which is of the same form as our standard linear additive model.

Suppose that the noise term w_i is log-normally distributed and hence $\tilde{w}_i \sim N(0, c^2)$.[11] We are free to choose the decisions x_ℓ, and so let us choose each one randomly by first choosing u_ℓ uniformly from the interval $[-\rho, \rho]$ and then assigning $x_\ell = e^{u_\ell}$. Thus, $\tilde{x}_\ell \sim U[-\rho, \rho]$. We continue to assume independence among the x's and w's, which translates into independence among the $\tilde{x}$'s and $\tilde{w}$'s. Therefore, we can apply the same estimation method to learn the $\mathbf{A}$ matrix in this multiplicative model. In particular, the statistic defined in Section 3.5 becomes $\tilde{y}_{ij} \triangleq \beta \cdot \Delta\tilde{q}_i \cdot \tilde{x}_j$, which would again be an unbiased estimator of a_{ij}. In addition, our methodology for estimating k and d from empirical data and our simulation procedure, presented in Sections 4 and 5, can be similarly adapted to fit the multiplicative model.

[11] More generally, we can relax the assumption – we require only that $\log(w_i)$ is sub-Gaussian with parameter c and has zero mean.

B Asymptotic Notation

Let $\mathbf{n}$ be a vector of variables; then we say:

(i) $f(\mathbf{n}) \in O(g(\mathbf{n}))$ if there exist constants N and $C > 0$ such that $|f(\mathbf{n})| \leq C|g(\mathbf{n})|$ for all $\mathbf{n}$ such that $n_i > N,\ \forall i$;

(ii) $f(\mathbf{n}) \in \Omega(g(\mathbf{n}))$ if there exist constants N and $C > 0$ such that $|f(\mathbf{n})| \geq C|g(\mathbf{n})|$ for all $\mathbf{n}$ such that $n_i > N,\ \forall i$;

(iii) $f(\mathbf{n}) \in \Theta(g(\mathbf{n}))$ if $f(\mathbf{n}) \in O(g(\mathbf{n}))$ and $f(\mathbf{n}) \in \Omega(g(\mathbf{n}))$.

In the first case, $f(n) \in O(g(n))$ essentially means that $f(n)$ grows *no faster* than $g(n)$ as n becomes large. In this sense, $g(n)$ can be thought of as an "upper bound" on the rate of growth of $f(n)$. An example is $f(n) = 100n$ and $g(n) = n^2$.

In the second case, $f(n) \in \Omega(g(n))$ essentially means that $f(n)$ grows *at least as fast* as $g(n)$ as n becomes large. And so in this case, $g(n)$ can be thought of as a "lower bound" on the rate of growth of $f(n)$. An example is $f(n) = n$ and $g(n) = \log n + 100\sqrt{n}$.

In the last case, $f(n) \in \Theta(g(n))$ means that $f(n)$ and $g(n)$ grow at essentially the same rate as n becomes large. An example is $f(n) = n + \sqrt{n}$ and $g(n) = 2n - 1$, as both grow linearly with n. We say that $f(n) \in \Theta(n)$ and $g(n) \in \Theta(n)$.

As illustrated above, asymptotic notation focuses on the order of growth and ignores constants. To justify the importance of focusing on the order of growth in the regime of a large number of products, let us consider the following example.

Example 1 (Impact of linear vs. logarithmic growth)**.** Suppose that there are two estimation methods, requiring $s_1(n) = n$ and $s_2(n) = 10 \log n$ experiments, respectively, in order to estimate an $\mathbf{A}$ matrix for n products. For a small number of products, such as $n = 10$, the first method requires just 10 experiments, whereas the second method requires $10 \log(10) \approx 23$ experiments. However, with a large number of products, such as $n = 100$, the first method now requires 100 experiments, whereas the second method requires $10 \log(100) \approx 46$ experiments, a much smaller number. As the number of products increases further, the difference between the two methods becomes more and more pronounced.

The purpose of asymptotic notation is to focus on the dominant scaling factor and ignore constants, such as 10 in method 2 of the example above. Although these constants have a relatively larger impact when n is small, they become insignificant as n becomes large. Specifically, we say that for method 1, $s_1(n) \in \Theta(n)$, and for method 2, $s_2(n) \in \Theta(\log n)$.

C Proof of Theorem 1

Proof. Let our decisions be i.i.d. continuous random variables x distributed uniformly on $[-\rho, \rho]$ so that $\mathbb{E}[x] = 0$ and $\text{var}(x) = \mathbb{E}[x^2] = \rho^2/3$. We perform an experiment using a vector of decisions $\mathbf{x}$. Let Δq_i be the observed percentage change in demand for product i, and let x_j be the pricing decision for product j.

Having defined $\beta \triangleq 3/\rho^2$, consider the statistic

$$y_{ij} = \beta(\Delta q_i x_j) = \beta \left(\sum_{\ell=1}^{n} a_{i\ell} x_\ell + w_i \right) x_j,$$

which satisfies $\mathbb{E}[y_{ij}] = a_{ij}$. Therefore, y_{ij} is an unbiased estimator of a_{ij}. Let $y_{ij}(t)$ be the statistic calculated from the tth experiment. By Assumption 1, for each (i, j), the statistics $y_{ij}(t)$ are independent and identically distributed across different experiments t. By the law of large numbers, the sample mean $\hat{a}_{ij} \triangleq \frac{1}{s} \sum_{t=1}^{s} y_{ij}(t)$ converges to a_{ij} as we take many samples from many experiments. We wish to bound the concentration of $\hat{a}_{ij}$ around its mean, a_{ij}.

To do so, we show that $\hat{a}_{ij}$ is sub-Gaussian. A random variable X is sub-Gaussian with parameter $\sigma > 0$ if

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp(\sigma^2 \lambda^2 / 2) \tag{7}$$

for all $\lambda \in \mathbb{R}$. We make use of the following well-known properties:

1. If X is sub-Gaussian with parameter σ, then $aX + b$ is sub-Gaussian with parameter $|a|\sigma$.

2. If X is bounded a.s. in an interval $[a, b]$, then X is sub-Gaussian with parameter at most $(b - a)/2$.

3. If X_1 and X_2 are sub-Gaussian with parameters σ_1 and σ_2, respectively,

   (a) and if X_1 and X_2 are independent, then $X_1 + X_2$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$.

   (b) and if X_1 and X_2 are not independent, then $X_1 + X_2$ is sub-Gaussian with parameter at most $\sqrt{2(\sigma_1^2 + \sigma_2^2)}$.

4. If X is sub-Gaussian with parameter σ, then it satisfies the following concentration bound:

$$\mathbf{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq 2 \exp\left(-\frac{\epsilon^2}{2\sigma^2} \right), \quad \forall \epsilon \geq 0. \tag{8}$$

We first consider the random variable y_{ij}:

$$\begin{aligned} y_{ij} &= \beta \left(\sum_{\ell=1}^{n} a_{i\ell} x_\ell + w_i \right) x_j \\ &= \beta \left\{ \left(\sum_{\ell \neq j} a_{i\ell} x_\ell + w_i \right) x_j + a_{ij} x_j^2 \right\} \\ &= \beta \left\{ V x_j + a_{ij} x_j^2 \right\}, \end{aligned}$$

where we have defined

$$V \triangleq \sum_{\ell \neq j} a_{i\ell} x_\ell + w_i.$$

We now show that V is sub-Gaussian. For each ℓ, x_ℓ is bounded on $[-\rho, \rho]$ and therefore sub-Gaussian with parameter ρ. Hence, $a_{i\ell}x_\ell$ is sub-Gaussian with parameter $|a_{i\ell}|\rho$. Also, under Assumption 1, w_i is sub-Gaussian with parameter c. The random variables $a_{i\ell}x_\ell$ and w_i are all independent. Therefore, their sum, V, is also sub-Gaussian with parameter $\sigma_V \triangleq \sqrt{\sum_{\ell \neq j} a_{i\ell}^2 \rho^2 + c^2}$.

Next, we show that Vx_j is sub-Gaussian using the definition. For any $\lambda \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{\lambda(Vx_j - \mathbb{E}[Vx_j])\right\}\right] &= \mathbb{E}\left[\exp\left\{\lambda(Vx_j)\right\}\right] && (9)\\
&= \int_{-\rho}^{\rho} \mathbb{E}[\exp\{\lambda(Vx)\}]\frac{1}{2\rho}dx && (10)\\
&\leq \int_{-\rho}^{\rho} \exp\left\{(|x|\sigma_V)^2\lambda^2/2\right\}\frac{1}{2\rho}dx && (11)\\
&\leq \int_{-\rho}^{\rho} \exp\left\{(\rho\sigma_V)^2\lambda^2/2\right\}\frac{1}{2\rho}dx\\
&= \exp\left\{(\rho\sigma_V)^2\lambda^2/2\right\},
\end{aligned}$$

where (9) is because Vx_j has zero mean; (10) is obtained by conditioning on the values of x_j; and (11) follows from (7) and the fact that for any $x \in [-\rho, \rho]$, Vx is zero-mean and sub-Gaussian with parameter $|x|\sigma_V$. Therefore, Vx_j is also sub-Gaussian with parameter $\rho\sigma_V$.

Next, we show that $a_{ij}x_j^2$ is sub-Gaussian. Since x_j^2 is bounded in $[0, \rho^2]$, it is sub-Gaussian with parameter $\rho^2/2$. Therefore, $a_{ij}x_j^2$ is sub-Gaussian with parameter $\rho^2|a_{ij}|/2$.

Finally, y_{ij} is a sum of two (dependent) sub-Gaussian random variables: $\beta V x_j$ with parameter $\beta\rho\sigma_V$, and $\beta a_{ij}x_j^2$ with parameter $\beta\rho^2|a_{ij}|/2$. Therefore, y_{ij} is also sub-Gaussian with parameter

$$\begin{aligned}
\sigma_Y \triangleq \sqrt{2(\beta^2\rho^2\sigma_V^2 + \beta^2\rho^4 a_{ij}^2/4)} &= \sqrt{2\left\{\beta^2\rho^2\left(\sum_{\ell \neq j} a_{i\ell}^2\rho^2 + c^2\right) + \beta^2\rho^4 a_{ij}^2/4\right\}}\\
&\leq \sqrt{2\beta^2\rho^4\left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}\\
&= \sqrt{18\left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}.
\end{aligned}$$

Since $\hat{a}_{ij} = \frac{1}{s}\sum_{t=1}^{s} y_{ij}(t)$ is a sample mean of s independent y_{ij}'s, $\hat{a}_{ij}$ is sub-Gaussian

with parameter

$$\sigma_{ij} \triangleq \frac{1}{s}\sqrt{s\sigma_Y^2} \leq \sqrt{\frac{18}{s} \cdot \left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}.$$

We can then bound the concentration of our estimate $\hat{a}_{ij}$ around the true parameter a_{ij} using (8):

$$\begin{aligned}\mathbf{P}(|\hat{a}_{ij} - a_{ij}| \geq \epsilon) &\leq 2\exp\left\{-\frac{\epsilon^2}{2\sigma_{ij}^2}\right\} \\ &\leq 2\exp\left\{-\frac{\epsilon^2}{36\left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}\right\}.\end{aligned}$$

This gives a concentration bound for the error of a particular (i, j) pair. To arrive at the final result, which bounds the maximum error over all (i, j) pairs, we apply the union bound and conclude that

$$\mathbf{P}\left(\max_{i,j} |\hat{a}_{ij} - a_{ij}| \geq \epsilon\right) \leq 2n^2 \exp\left\{-\frac{s\epsilon^2}{\max_i 36\left(\sum_{\ell=1}^{n} a_{i\ell}^2 + c^2/\rho^2\right)}\right\}.$$

□

D Proof of Theorem 2

Proof. For any $\lambda > 0$, consider the class $\mathcal{A}_{n,k}(\lambda)$. Fix some $\epsilon \in (0, \lambda/2)$ and $\delta \in (0, 1/2)$. In what follows, all estimators use the results of s experiments, for some arbitrary s.

Define the sub-class $\mathcal{A}_{n,k}^{\text{const}}(\lambda) \triangleq \{\mathbf{A} \in \mathbb{R}^{n\times n} : |\{j : a_{ij} \neq 0\}| = k, \forall i = 1, \ldots, n; a_{ij} = \lambda, \forall i, j \text{ s.t. } a_{ij} \neq 0\} \subset \mathcal{A}_{n,k}(\lambda)$, which is the class of all $n \times n$ $\mathbf{A}$ matrices whose rows are k-sparse and whose nonzero entries are all exactly equal to λ.

The desired specification is an estimator that for any $\mathbf{A}$ matrix in $\mathcal{A}_{n,k}(\lambda)$ achieves uniformly ϵ-accurate estimates with probability $1-\delta$. In order to obtain a lower bound on the number of experiments needed to meet this specification, it suffices to obtain a lower bound on the number of experiments needed to meet the following looser specification: we let the $\mathbf{A}$ matrix be generated uniformly at random from the sub-class $\mathcal{A}_{n,k}^{\text{const}}(\lambda)$ and require that with probability at least $1-\delta$ the first row of $\mathbf{A}$ is correctly recovered. Because $\mathbf{A} \in \mathcal{A}_{n,k}^{\text{const}}(\lambda)$, all elements of $\mathbf{A}$ are either exactly 0 or λ, and since $\epsilon \in (0, \lambda/2)$, achieving uniform ϵ-accuracy is equivalent to perfectly recovering $\mathbf{A}$, which is also equivalent to perfectly recovering the sparsity pattern of $\mathbf{A}$ (i.e., identifying the locations of all nonzero entries). Let R_1^{const} denote the event of exactly recovering the sparsity pattern of the first row of an $\mathbf{A}$ matrix chosen uniformly at random from $\mathcal{A}_{n,k}^{\text{const}}(\lambda)$.

We now focus on the event R_1^{const} and find an upper bound on its probability. Within the sub-class $\mathcal{A}_{n,k}^{\text{const}}(\lambda)$, there are exactly $N \triangleq \binom{n}{k}$ possible sparsity patterns for the first row of any $\mathbf{A}$ matrix. Moreover, because all nonzero entries are equal to the same value λ, each unique sparsity pattern corresponds to a unique row vector, and vice versa. Suppose that

we randomly choose the first row $\mathbf{a}_1'$ by choosing one of the N possible sparsity patterns uniformly at random. We can then view the sparsity pattern recovery problem as a channel coding problem. The randomly selected sparsity pattern $\theta \in \{1, \ldots, N\}$ is encoded, using a sequence of s experimental decisions $\mathbf{X} \in \mathbb{R}^{n \times s}$, into codewords $\mathbf{r} = \mathbf{a}_1'\mathbf{X} = (r_1, r_2, \ldots, r_s) \in \mathbb{R}^s$. These codewords represent the uncorrupted percentage change in demand for product 1 in each of the s experiments. The codewords are sent over a Gaussian channel subject to noise $\mathbf{w} = (w_1, w_2, \ldots, w_s) \sim \mathcal{N}(0, c^2 I) \in \mathbb{R}^s$, and finally received as noisy measurements $\mathbf{y} = \mathbf{r} + \mathbf{w} = (y_1, y_2, \ldots, y_s) \in \mathbb{R}^s$, which are equal to the observed noisy percentage change in demand $\Delta\mathbf{q}_1$. The goal is to recover the pattern θ from the measurements $\mathbf{y}$.

The power of a Gaussian channel is given by $P = \frac{1}{s}\sum_{t=1}^{s} r_t^2$. Given that $\mathbf{a}_1'$ is k-sparse and any decision x is in $[-1, \tilde{\rho}]$, we have that $|r_t| \leq k\lambda\tilde{\rho}$ for all t, and hence $P \leq k^2\lambda^2\tilde{\rho}^2$. From standard results (Cover and Thomas 1991), the capacity of a Gaussian channel with power P and noise variance c^2 is $\frac{1}{2}\log\left(1 + \frac{P}{c^2}\right)$. Therefore, the capacity of our particular channel is

$$C \leq \frac{1}{2}\log\left(1 + \frac{k^2\lambda^2\tilde{\rho}^2}{c^2}\right).$$

From Fano's inequality (Cover and Thomas 1991), we know that the probability of error, P_e, of a decoder that decodes the sparsity pattern θ from noisy measurements $\mathbf{y} \in \mathbb{R}^s$ is lower bounded as

$$\begin{aligned} P_e &\geq \frac{H(\theta \mid \mathbf{y}) - 1}{\log N} \\ &= \frac{H(\theta) - I(\theta; \mathbf{y}) - 1}{\log N} \\ &= \frac{\log N - I(\theta; \mathbf{y}) - 1}{\log N} \\ &= 1 - \frac{I(\theta; \mathbf{y}) + 1}{\log N}, \end{aligned}$$

where H denotes entropy and I denotes mutual information. The first equality is by the definition of mutual information, and the second equality follows from the fact that θ is chosen uniformly over a set of cardinality N. We can upper bound the mutual information

between θ and $\mathbf{y}$ as

$$\begin{aligned} I(\theta;\mathbf{y}) &\leq I(\mathbf{r};\mathbf{y}) && (12)\\ &= h(\mathbf{y}) - h(\mathbf{y} \mid \mathbf{r}) \\ &= h(\mathbf{y}) - h(\mathbf{w}) \\ &\leq \sum_{t=1}^{s} h(y_t) - \sum_{t=1}^{s} h(w_t) && (13)\\ &= \sum_{t=1}^{s} [h(y_t) - h(y_t \mid r_t)] \\ &= \sum_{t=1}^{s} I(r_t; y_t) \\ &\leq sC, && (14) \end{aligned}$$

where h denotes differential entropy, (12) follows from the data processing inequality, (13) follows from the independence of the w_t's and the fact that the entropy of a collection of random variables $\{y_t\}$ is no more than the sum of their individual entropies, and (14) follows from the definition of channel capacity as the maximal mutual information. And so by Fano's inequality, the probability of error is lower bounded by

$$P_e \geq 1 - \frac{sC+1}{\log N},$$

which immediately gives the following upper bound on the probability of R_1^{const}:

$$\mathbf{P}(R_1^{\text{const}}) = 1 - P_e \leq \frac{sC+1}{\log N}.$$

Therefore, achieving the looser specification of uniform ϵ-accurate estimates of the first row of a random $\mathbf{A} \in \mathcal{A}_{n,k}^{\text{const}}(\lambda)$ with probability $1-\delta$ implies the following condition on the number of experiments s:

$$1 - \delta \leq \frac{sC+1}{\log N} \Longrightarrow s \geq \frac{(1-\delta)\log N - 1}{C}.$$

Consequently, achieving the stricter original specification of an estimator that for all $\mathbf{A}$ matrices in $\mathcal{A}_{n,k}(\lambda)$ achieves uniformly ϵ-accurate estimates with probability $1-\delta$ requires the number of experiments to satisfy the above condition.

With some simple rearrangement, and noting that $\log N = \log \binom{n}{k} \geq k \log(n/k)$ and $\delta \in (0, 1/2)$, we obtain the desired lower bound:

$$s \geq \frac{(1-\delta)\log N - 1}{C} \geq \frac{2(1-\delta)k\log(n/k) - 2}{\log(1 + k^2\lambda^2\tilde{\rho}^2/c^2)} \geq \frac{k\log(n/k) - 2}{\log(1 + k^2\lambda^2\tilde{\rho}^2/c^2)}.$$

□

References

Anderson, E. T., Cho, E., Harlam, B. A., and Simester, D. I. (2010). What affects price and price cue elasticities? Evidence from a field experiment. Working paper.

Anderson, E. T. and Simester, D. I. (2001). Are sale signs less effective when more products have them? *Marketing Science*, 20(2):121–142.

Anderson, E. T. and Simester, D. I. (2013). Advertising in a competitive market: The role of product standards, customer learning, and switching costs. *Journal of Marketing Research*, 50(4):489–504.

Candès, E. J. (2006). Compressive sampling. *Proceedings of the International Congress of Mathematicians*.

Candès, E. J., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509.

Candès, E. J. and Tao, T. (2005). Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley and Sons, New York.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. *Proc. 21st Annual Conf. Learn. Theory (COLT 2008)*.

Dubé, J.-P., Hitch, G. J., Rossi, P. E., and Vitorino, M. A. (2008). Category pricing with state-dependent utility. *Marketing Science*, 27(3):417–429.

Erdem, T. (1996). A dynamic analysis of market structure based on panel data. *Marketing Science*, 15(4):359–378.

Farias, V., Jagabathula, S., and Shah, D. (2013). A nonparametric approach to modeling choice with limited data. *Management Science*, 59(2):305–322.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Hauser, J. R., Toubia, O., Evgeniou, T., Befurt, R., and Dzyabura, D. (2010). Disjunctions of conjunctions, cognitive simplicity, and consideration sets. *Journal of Marketing Research*, 47(3):485–496.

Huber, J. and Zwerina, K. (1996). The importance of utility balance in efficient choice designs. *Journal of Marketing Research*, 33(3):307–317.

Keane, M. P. (1997). Modeling heterogeneity and state dependence in consumer choice behavior. *Journal of Business & Economic Statistics*, 15(3):310–327.

Lewis, R. A. and Rao, J. M. (2012). On the near impossibility of measuring advertising effectiveness. Working paper.

Liu, Q. and Arora, N. (2011). Efficient choice designs for a consider-then-choose model. *Marketing Science*, 30(2):321–338.

Louviere, J. J., Hensher, D. A., and Swait, J. D. (2000). *Stated Choice Methods Analysis and Application*. Cambridge University Press, Cambridge, UK.

Louviere, J. J., Street, D., and Burgess, L. (2004). A 20+ years' retrospective on choice experiments. In Wind, J., editor, *Tribute to Paul Green*, International Series in Quantitative Marketing, chapter 8. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Manchanda, P., Ansari, A., and Gupta, S. (1999). The "shopping basket": A model for multicategory purchase incidence decisions. *Marketing Science*, 18(2):95–114.

Mersereau, A. J., Rusmevichientong, P., and Tsitsiklis, J. N. (2009). A structured multi-armed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control*, 54(12):2787–2802.

Sandor, Z. and Wedel, M. (2001). Designing conjoint choice experiments using managers' prior beliefs. *Journal of Marketing Research*, 38(4):430–444.

Seetharaman, P. B., Ainslie, A., and Chintagunta, P. K. (1999). Investigating household state dependence effects across categories. *Journal of Marketing Research*, 36(4):488–500.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, 58(1):267–288.

Toubia, O., Simester, D. I., Hauser, J. R., and Dahan, E. (2003). Fast polyhedral adaptive conjoint estimation. *Marketing Science*, 22(3):273–303.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, 79(4):281–299.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Villas-Boas, J. M. and Winer, R. S. (1999). Endogeneity in brand choice models. *Management Science*, 45(10):1324–1338.

Wainwright, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741.