DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

# Essays in Industrial Organization and Econometrics

**Permanent link**
http://nrs.harvard.edu/urn-3:HUL.InstRepos:40046463

**Terms of Use**
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

## Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

# Essays in Industrial Organization and Econometrics

A dissertation presented

by

Daniel Pollmann

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Economics

Harvard University

Cambridge, Massachusetts

April 2017

© 2017 Daniel Pollmann

All rights reserved.

*Dissertation Advisor:*
**Professor Ariel Pakes**

*Author:*
**Daniel Pollmann**

**Essays in Industrial Organization and Econometrics**

## Abstract

This thesis includes two essays in Industrial Organization and one in Econometrics. The first two essays study the returns to the scale of data available to firms and its impact on market structure, using actual firm data from an online retailer that optimizes its product display on the basis of revealed preference. While the first chapter presents an approach that is directly empirical looking at the quality of out-of-sample predictions, the second, which is the main chapter of my dissertation, is model-based, allowing for a competitive market analysis and accounting for reoptimization of learning based on firm size. Both papers find moderate to significant returns to data depending on a firm's position on the learning curve. The third essay shows that in settings with interval-censored data, which is common for instance with survey data, information on the marginal distribution of said regressors can substantially tighten identification regions when combined with restrictions on the regression function such as monotonicity and shape.

# Contents

# List of Tables

# List of Figures

# Acknowledgments

I am greatly indebted to my advisors Ariel Pakes, Elie Tamer, and Robin Lee for their invaluable support, guidance, and feedback on my research. I thank my co-author Tilman Dette for the great collaboration on the first essay. I am grateful for the advice and support of Gary Chamberlain, Rebecca Compton, Mike Egesdal, David Laibson, Greg Lewis, Jing Li, Luca Maini, Filippo Mezzanotti, Pepe Montiel-Olea, Brenda Piquet, Mikkel Plagborg-Møller, Michael Pollmann, Guillaume Pouliot, Ann Richards, Elizabeth Santorella, Frank Schilbach, Lukas Schwarz, Andrei Shleifer, Tom Wollmann, Ali Yurukoglu, Tom Zimmermann, and the participants of Harvard's Industrial Organization and Econometrics research seminars. I thank anonymous executives and employees of the retailer that provided the data used in chapters one and two for their openness and time.

# Introduction

The main theme of this dissertation is quantifying the value of data from both an economic and a statistical perspective. The first two chapters constitute, to the best of my knowledge, the first systematic empirical studies on the returns to the scale of data available to firms. This is important in many markets, as the amount of data any one firm possesses is endogenously determined based on its market share. Both chapters draw on data from an online retailer selling consumer durable products, with a more precise description of the setting and data given in the first chapter. The retailer sells a large number of products, which it needs to rank on its catalog pages; crucially, it learns about their popularity through revealed preference. The two chapters are different in that the first chapter presents an approach that is directly empirical looking at the quality of out-of-sample predictions, while the second, which is the main chapter of my dissertation, is model-based, thereby allowing it to answer a wider array of questions while accounting for possible reoptimization. The third paper is an applied econometrics paper, studying a problem in which parameters of interest are only partially identified because regressors are interval-censored. We ask how much identification regions tighten if in addition, data or other statistical information on the marginal distribution of these regressors is available. The three papers are briefly summarized below.

### Returns to Data Scale: Empirical Evidence from Online Retail[1]

For a large online retailer, we estimate the returns to observing the search behavior of additional consumers when this data is used to optimize the ranking of products displayed to

[1] Co-authored with Tilman Dette

all consumers, affecting their search and demand. We estimate hypothetical firm beliefs over product quality on training datasets of different size using a revealed-preference approach, and analyze the prediction quality of these estimates on hold-out data. Given a large and heterogeneous product catalog, there is a substantial size-precision gradient, which translates into quality differences in the predicted optimal ranking that affect consumer search and purchases beyond the initial visit.

### Returns to Data Scale and the Impact on Competition: Evidence from Online Retail

Using actual firm data from an online retailer, this paper addresses two questions relevant for business and regulators: what are the returns to data scale, and when can they act as a barrier to entry? We present and estimate a model of a firm optimizing its catalog display over several thousand products and determine the optimal rate of experimentation resulting in the highest average steady-state payoff using a modification of Thompson Sampling. Even though the statistical return to data becomes smaller as the size of the available data increases, the economic return can remain significant, as the data of the marginal consumer is applied to improve the product ordering for a growing base of inframarginal consumers. For the firm in our data, we estimate an elasticity of the expected number of product clicks per customer with respect to the number of customers of 0.024. This implies that for every additional click provided by an additional consumer, her contribution to the quality of the product ordering displayed to all consumers yields an additional 0.024 clicks. We can calculate the reduction in optimal mark-ups that results from the marginal consumer's contribution to the firm's stock of data; we find a reduction of mark-ups of 3.6% given a price elasticity of -3. We use the estimated learning technology to map out the minimum viable size for entrants to compete in a market in which consumers choose firms based on quality and firms learn about quality through revealed consumer preference.

**Identification and Estimation with an Interval-censored Regressor when its Marginal Distribution is Known**

When regressors of interest are interval-censored, knowledge of their marginal distribution implies additional restrictions, which, combined with shape restrictions, lead to tighter bounds on the regression function. We show how to estimate these bounds, as well as bounds on any linear functional, using a linear programming formulation and establish consistency. Valid Bayesian and frequentist inference can be performed using the computationally attractive Bayesian bootstrap. An application to age-earnings profiles illustrates the usefulness of distribution information.

# Chapter 1

# Returns to Data Scale: Empirical Evidence from Online Retail[1]

## Introduction

As firms collect growing amounts of data about the markets in which they operate, the competitiveness of any one firm can be affected by exactly what and how much data it can draw on for its decision-making. In consumer markets, for instance, a firm with a larger existing base of consumers will, all else equal, likely have more precise estimates of its decision-relevant parameters, such as those describing demand. Therefore, due to endogenous information sets, relative firm competitiveness can be a function of market structure, particularly in settings with large degrees of private information. Furthermore, when future information sets depend on current competitiveness – such as when a firm that offers higher quality at present will attract more consumers in the future – the resulting feedback loop can dynamically reinforce asymmetries between market participants.

We empirically document the magnitude of these effects – how great are precision gains from additional data, how do they translate into static competitiveness, and what is medium-run adoption – in the setting of a large online retailer. This firm sells a large number of

[1] Co-authored with Tilman Dette

products and needs to decide how to rank these when displaying its product catalog to consumers arriving at its website. The attractiveness of products is a set of parameters the retailer estimates based on observed consumer behavior, including how many views and orders a given product received. We have access to the entire browsing, price, and order data of the retailer for a long period of time, meaning that we see exactly that part of the firm's information set which is derived from privately observable consumer behavior. We know that in practice, the firm's decision-making relies heavily on this data, making this an ideal setting in which to study how a firm's ability to display and sell attractive products as well as attract repeat consumer interest depends on the inference on product quality it can draw from its existing customer data.

To isolate the return on additional consumer data, we ask what estimates of product quality hypothetical firms of different sizes would have arrived at, how that would have translated into different decision vectors, and what the resulting changes in static and dynamic outcomes would have been. We construct these counterfactuals by splitting the available data into training datasets of different sizes and analyzing outcomes on hold-out data when solving the firm optimization problem using estimates for the former. More specifically, we run regressions for whether consumers choose to click on and view products from a menu displayed to them, where the parameters of interest are product-level fixed effects. This yields estimates of product quality which approximate those used in practice by the retailer's algorithms. Importantly, we interpret the training data quality estimates not as the true parameter values, but rather the counterfactual beliefs of firms of different sizes, which we use to have each of them reoptimize. We then present several statistics of the economic value associated with the resulting decision vectors calculated on our hold-out dataset. Specifically, we use this data to reestimate the product click regression and obtain statistically independent estimates for the number of product views implied by the model for each of the hypothetical firms. We find that the hold-out estimates imply substantial returns to using larger fractions of the data in identifying the top products, with significantly larger average quality at the top of the first catalog page as well as for average quality of the entire page. We then present

estimates showing that consumers who were displayed a higher quality set of products on their first visit – controlling for granular time trends and other factors – recorded significantly more clicks and purchases on the same day as well as greater engagement in the period after.

We take our reduced-form counterfactual calculations as evidence that in a realistic setting, there are large returns to observing additional consumer data. The main limitation of our approach is that the product rankings present in the data we use to form counterfactual beliefs are chosen by the firm and affect the rate of learning about the set of quality parameters. More specifically, the firm faces an optimization problem precisely because many consumers will only consider the subset of products that is most prominently displayed at the top of the first page of results. By implication, a firm will learn substantially faster about the quality of these products. Firms of different hypothetical sizes could optimize their ranking in light of expected sample sizes; small firms, for instance, may choose to experiment differently or specialize to reduce the number of relevant parameters. In addition, since the actual ranking of products in the observed data is based on the retailer's entire data, it is optimized to efficiently learn about the most relevant elements of the quality vector potentially more so than a smaller firm would know to. One avenue of further research is a more structural analysis of this effect of active learning.

We think that the general mechanism we consider extends beyond the technology and retail sectors to other parts of the economy, in particular consumer-facing firms. In credit and insurance markets, firms typically use predictive models for the risk of potential borrowers or insurees (Bundorf *et al.*, 2012, Einav *et al.*, 2013, 2012). Since rich models will require large sample sizes for training and validation, larger firms can make acceptance and pricing decisions that more accurately reflect underlying risk. Examples of other settings in which firms collect large amounts of consumer data include electronic health records and the utilities and communication sectors.

Even more broadly, this paper connects to a growing literature in economics that highlights new opportunities for research using "big data" and develops appropriate statistical methods for doing so. Here, we look directly at firms as econometricians and analyze what they

can learn from growing amounts of data in high-dimensional settings. We believe that the endogeneity of how much data a given firm observes creates interesting and potentially very important interactions with market structure.

Section 1.1 describes the empirical setting and available data in more detail. Section 1.2 sets up a stylized model of firm optimization and presents our corresponding statistical model which we use for both our own estimates and to model firm beliefs. We then present estimates from this model and the resulting effects on firm optimization in section 1.3 before concluding with a brief discussion in the final section.

## 1.1 Empirical setting and data

The data were obtained from an online retailer selling consumer durable products across a variety of categories. The majority of these categories, taken together, constitute one of the merchandise lines used by the U.S. Census Bureau to subcategorize "Electronic Shopping and Mail-Order Houses" (Bureau, 2015), which includes e-commerce. The retailer sources its products from manufacturers or their distributors and sells them directly to consumers, primarily through one main website, on which it is the only seller. It is among the largest online retailers in its product segment. The product mix varies by category from well-known brands to differentiated niche products, and its main competitors are online and offline retailers selling identical or substitute products.

Within a typical product category, we think of products as substitutes, with each potential customer having unit demand. However, consumers may purchase from several categories. Even before revenue weighting, the median product ordered in our sample exceeds the minimum amount beyond which shipping is free. The percentage of products which is returned by customers is in the single digits, so that we can generally assume that customers order only products they intend to purchase. The products are differentiated vertically, ranging from entry level to more upmarket, as well as horizontally, with taste heterogeneity playing a large role.

The retailer uses direct marketing, primarily via email to registered customers, but also

advertises online on search engines or other websites as well as offline, for example through TV spots. Consumers arrive to the website either directly by entering its URL in their browser, from a search engine, or by clicking on a link or banner in an email or on another website.

We focus on consumers who arrive to the catalog page for a specific product category, where a category is sufficiently narrowly defined that all of its products are at least minimally substitutable. These category pages are the largest channel for orders, and they are the primary means by which consumers can search the differentiated product assortment. Consumers can arrive at these pages either through an external search engine, by navigating through from the homepage, or by entering a search term on the website, which, unless very specific, will refer them to one of the catalog pages, possibly with filters corresponding to their search query applied. The set of products displayed on these pages, in general, does not condition on any consumer-specific information, so the firm problem studied here is that of selecting a default list of products intended to cater to the overall population of consumers.

In the data we use, each of these catalog pages shows a grid of 48 products, and consumers can click through to see additional pages with the same number of products from this category. Alternatively, they can filter the set of products displayed by attributes such as price as well as product characteristics which will vary across categories. For each product, consumers see a photo, the product and brand name, its price, and average reviews. After clicking on a product, a new page loads up that provides additional photos and information for the specific product as well as the option to add this product to the shopping cart and check out to purchase the product.

In our dataset, we observe all of the navigational choices made by the consumer, that is, all category and product pages viewed in their exact order, typically referred to as clickstream. Consumer visits can (sometimes imperfectly) be matched over time to a panel using identifiers based on browser cookies, IP addresses, or log-ins. In addition to product clicks, we also observe whether any of the products were ordered. Crucially, we also observe all products with their position on the catalog page, irrespective of whether they were clicked. Furthermore, we know the price of each product at each point in time.

The retailer sets its own prices, subject to minimum advertised prices set by manufacturers for some of the products. Price variation arises from cost shocks (either to wholesale costs or estimated total costs), explicit experimentation, business logic, and managerial decisions affecting the overall price level, which may translate differentially to the product level. In what follows, we treat the resulting within-product price variation as conditionally exogenous after controlling for the product as well as granular time trends. This stands in contrast to relying on (less common) explicit sales periods; in fact, we discard observations for which the difference between current and median log product price is larger than .25.

In addition, the retailer controls the order in which products appear on the catalog pages. Variation in the sort order arises from product stock-outs, explicit experimentation (so-called A/B tests comparing discrete variations of sort orders as well as giving "exposure" to new or otherwise promising products), and an optimization algorithm that produces a fair amount of variation, which, if not explicitly stochastic, is nonetheless useful (and used in practice) to learn the relative attractiveness of products. Both stock-outs and the optimization algorithm, which is run at regular time intervals to compute a new sort order, generate variation over time, while experiments may lead to variation at any one point in time.

In addition to the above criteria, we restrict the sample to consumers browsing on computers rather than mobile devices, which is true for the majority of visits and an even larger fraction of orders. We also exclude sort orders that were explicitly personalized for a subset of repeat visitors based on which products they had clicked on in the past. This represents a small fraction of traffic, does not apply to new visitors, and is generally less useful for the purposes of our analysis (and this traffic is hence also excluded by the internal optimization algorithm responsible for the sort order displayed to the vast majority of consumers). We additionally exclude visitors flagged as bots by the retailer's internal logic. The final sample includes tens of millions of visits in under two years. In our analysis, we collect the product categories in our sample further into product groups based on an internal taxonomy and estimate the model separately for these groups.

For our analysis, we do not rely exclusively on experimental variation. In this paper,

we study what an actual firm learns about consumer demand, and in practice firms often need to rely on observational data when their optimization and estimation problem is very high-dimensional, such as sorting a very large number of products by attractiveness or pricing using product-specific elasticities with a long tail of products. Explicit experimentation is more commonly used to estimate the treatment effect of discrete variations or different optimization algorithms, each of which produces one set of prices, sort orders, etc., in line with the purpose of the experiment (see Dinerstein *et al.* 2014 for an example). In fact, we specifically aim to show that even for realistic sizes of observational data seen by a large firm, the estimation and economic optimization problem can be sufficiently challenging such that there can be large returns to additional scale, and there simply is not enough traffic on which to run experiments for these to serve as the basis of estimation alone.

With this dataset in hand, we observe exactly what the firm observes about revealed consumer preferences, which is crucial given our empirical interest. It is based on the same underlying data the retailer uses to optimize its sort order and other aspects of its website and business.

## 1.2 Empirical model of product quality

For each category, the firm needs to rank its products $j = 1, \ldots, J$ into sort positions $r = 1, \ldots, R$ for any consumer $i$ that visits the page. We assume that the firm uses a simple separable model to approximate its static optimization problem over the expected profits from any such ranking:

$$\max_{\sigma} \mathbb{E}\left[\sum_{r=1}^{R} Y_{i,j(r;\sigma)} b_j\right] \quad = \quad \max_{\sigma} \sum_{r=1}^{R} \gamma_r \mu_{j(r;\sigma)} b_j, \tag{1.1}$$

where $Y_{i,j(r;\sigma)} \in \{0, 1\}$ is the binary outcome of interest of product $j$ listed in position $r$ given a ranking (or permutation) $\sigma$ for consumer $i$, and $b_j \in \mathbb{R}$ is the associated benefit. The expected value is assumed to be separable in two dimensions: additively across products, and multiplicatively between products and positions, which enter with parameters $\gamma_r$ and $\mu_{j(r;\sigma)}$,

respectively.[2]

When maximizing the expected number of clicks, $Y_{i,j(r;\sigma)}$ is an indicator for whether the product in question is clicked by consumer $i$, and $b_j = 1$ for $j = 1, \ldots, J$. Under the assumption that positions effects are decreasing,[3] $\{\gamma_r\}_{r=1}^R$ forms a decreasing sequence, and it is optimal for the firm to sort products in descending order of product effects $\{\mu_j\}_{j=1}^J$.

The central object of interest to the firm is thus the vector of product effects $\mu$, which translates the question of this paper – how do firms of different size differ in their competitiveness due to the amount of data they observe? – into i) how well firms of different size can estimate and learn the vector $\mu$, and ii) how this impacts their profits. We operationalize this task, derived from the stylized yet empirically relevant model above, by estimating product quality as a fixed effect in a logit regression that also includes a position as well as price effect alongside a rich set of controls:

$$Y_{i,j} \quad = \quad \mathbf{1}\left\{\delta_j - \alpha p_{i,j} + x'_{i,j}\beta + \epsilon_{i,j} \geq 0\right\}, \tag{1.2}$$

where $Y_{i,j}$ is the outcome of product $j$ for consumer $i$, $p_{i,j}$ is the log price, $x_{i,j}$ is a vector including controls such as for position and time, $\delta_j$ is the fixed effect for product $j$, and $\epsilon_{i,j}$ is assumed to follow an iid $EV\,(1)$ distribution conditional on all the right-hand side variables.

Under this assumption, the estimated coefficients have a causal interpretation as certain average elasticities with respect to price and position. However, a truly structural model of the data-generating process should also account for dependence between products in both search and purchase decisions, which is ignored here. We nonetheless find the estimates on price and position (presented in section 1.3.1) useful both on their own and as a guide for the magnitude of effects that should be captured in a structural model.

The main purpose of this regression, however, is to deliver estimates of product quality

[2]Lahaie and McAfee (2011) use the same model to argue that in constructing an efficient ranking, some degree of shrinkage should optimally be applied to estimates of advertiser effects when these are uncertain. See Jeziorski and Segal (2015) and Jeziorski and Moorthy (2015) for empirical models weakening these separability assumptions in the sponsored-search context.

[3]See Ursu (2016) for evidence of large effects in a field experiment run by a travel intermediary.

that are qualitatively similar to those a firm would have arrived at. A Bayesian firm would use the data to update its prior, leading to posterior beliefs on product qualities, while a frequentist firm would also likely shrink its estimates if it cares about precision rather than unbiasedness. Put more practically, for a large number of products, a firm seems unlikely to favor a product that was successful in its only observation over one that consistently outperformed for many observations, at least in the static problem analyzed here. For this reason, we impose a ridge L2 penalty on the fixed-effect coefficients, leading to a penalized likelihood function which adds to the unpenalized likelihood based on model (1.2) the sum of squared fixed-effect coefficients weighted by a constant $\lambda$. The maximand of this expression is a maximum a posteriori estimate and equals the mode of the posterior distribution from using a Normal prior $\delta \sim \mathcal{N}\left(0, \frac{1}{2\lambda}\right)$.[4]

In the logit model, for small probabilities,

$$\begin{aligned} \Pr\left(Y_{i,j} = 1 \mid p_{i,j}, x_{i,j}\right) &= \frac{\exp\left(\delta_j - \alpha p_{i,j} + x'_{i,j}\beta\right)}{1 + \exp\left(\delta_j - \alpha p_{i,j} + x'_{i,j}\beta\right)} \\ &\approx \exp\left(\delta_j - \alpha p_{i,j} + x'_{i,j}\beta\right), \end{aligned}$$

which factors into a product effect and a position effect, yielding an approximate correspondence to the relevant parameters of the stylized model (1.2) above, for which the optimal policy is simply assortative. Taking the fixed-effect vector $\delta$ to be our empirical analogue to $\mu$, we thus assume that the firm sorts its products according to its vector of estimates $\hat{\delta}$, which, as we discuss above, need not be the maximum likelihood estimate. In addition to economic interpretability, the decision for the logit model is motivated by computational considerations, which loom large for a dataset of the size considered and will play an even larger role for any firm that needs to regularly update these estimates.

Finally, we note that for the parameters of the logit model (1.2) to be consistently estimated, we require the number of observations per product to grow large to avoid incidental parameter

[4]We therefore choose $\lambda$ based on the variance of fixed-effect estimates in a typical product category.

bias.[5] Here, we are specifically interested in an empirical setting in which fixed effects cannot all be estimated with arbitrary precision. We find it plausible, however, that the bias in the estimates of the "common" parameters, chiefly price and position, can vanish sufficiently fast to be of second order to our empirical question, while at the same time, the gain in precision in the fixed effect estimates that would result from sampling additional observations of the respective units would still be economically meaningful. We are unaware of theoretical or simulation results on incidental parameter bias in a setting such as the present in which different units of the panel are potentially sampled at very different rates. Bias considerations may furthermore be of less than usual importance since the estimation problem at hand is closer to a prediction problem, specifically for which products have the highest quality as perceived by consumers.

## 1.3 Results

We now present estimates for how firm competitiveness varies with firm size as a result of how much data a firm has available to estimate product quality. We consider competitiveness to be the true quality the firm is able to serve consumers after solving optimization problem (1.1) when using its own (noisy) estimates. We look at both short-run click outcomes, which correspond exactly to this measure of competitiveness, and long-run (orders and future visits) outcomes, which may be directly or indirectly affected as well.

Throughout this section, we will work with different non-overlapping subsamples of the original dataset, using a hold-out sample to evaluate the prediction quality a firm would have achieved from a given training sample when running the logit regression (1.2).[6] This

[5] An alternative approach (Chamberlain, 1980), based on a conditional likelihood, does not suffer from this type of bias but has at least two other deficiencies in our context: i) it does not yield straightforward estimates of the fixed-effect parameters (which are considered nuisance parameters in many panel data models), ii) the computation of the conditional likelihood becomes extremely computationally burdensome and numerically unstable (underflow problems) once the panel dimension is moderate rather than small.

[6] This relies on numbering browser cookies based on order of arrival and assigning them to groups based on this number. For instance, to get two groups of the same size, we would divide the sample into one subsample with all even arrival numbers and another with all odd. Due to the large number of new cookies on any given day, this yields approximately random, temporally stratified assignments.

ensures that the quality estimates we assume the firm to use are statistically independent of our realized competitiveness measure, which is important since true competitiveness (1.1) is unobserved. Instead, we rely on an empirical analogue obtained by re-running regression (1.2) on the hold-out data.

By working with subsamples, we are naturally limited to counterfactual experiments in which the firm is shrunk relative to its actual size. We think that these are nonetheless interesting since the firm in question is relatively large, and because the effects in question can be expected to change continuously as a function of firm size, allowing reasonable extrapolation beyond the actual firm size. A more principled though also more model-dependent alternative approach would rely on a structural model, which could also allow for active learning through experimentation on the firm side.

We start this section by providing evidence that in our setting, changes in the sample size translate into noticeable changes in the precision of a firm's quality estimates. Then, we present the core results of this paper: a firm's quality level can be substantially affected by the amount of available data, and higher quality translates into more clicks and purchases the day of, and though measured with less precision, likely in subsequent days and weeks.

### 1.3.1 Quality model

The main parameters of interest in our empirical model (1.2) are the product qualities measured by fixed effects, position effects, and click price elasticities. Across product groups, the estimated click price elasticity varies from -.68 to -2.29 with a median of -1.56. While purchase elasticities tend to be a multiple of these numbers, we nonetheless take them as evidence that consumer search on the catalog pages we consider is directed and meaningfully reflects underlying preferences. Figure 1.1 shows the estimated position effects on the first catalog page as an average over all categories, with the shaded intervals around the mean representing one standard deviation of the effect over product categories in each direction. The drop from the first position to the bottom of the page is substantial, reducing click interest by around 70%.

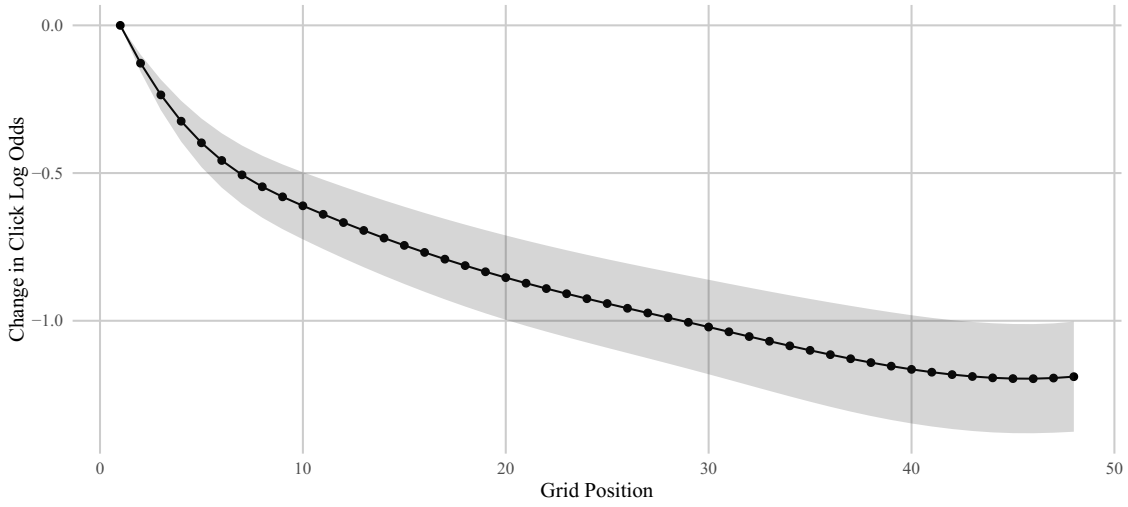


**Figure 1.1:** *Position effect (first page) relative to first position*

Our fixed-effect estimates suggest substantial heterogeneity in product quality, as illustrated by a histogram of the estimated product fixed effects for an example category in figure 1.2. It is important to remember, however, that these do not necessarily give the true distribution of quality parameters. The distribution of the maximum likelihood estimates would be substantially more dispersed than the distribution of the underlying parameters due to estimation error. We shrink these estimates to reflect the quality inferences a firm would have drawn, and the resulting estimates form a distribution that can have greater or lower variance than the true quality distribution. For interpretation, products to the left of the distribution are not necessarily of low quality to all consumers; the estimates are particular average levels, and with heterogeneous preferences, products may appeal to consumer types of different mass. Also, it is perhaps interesting that estimates in the right tail are shrunk less than those in the left tail because they are estimated on a greater number of impressions due to the retailer's own optimization. This is reflected in the shift of the quality distribution once reweighted by impressions.

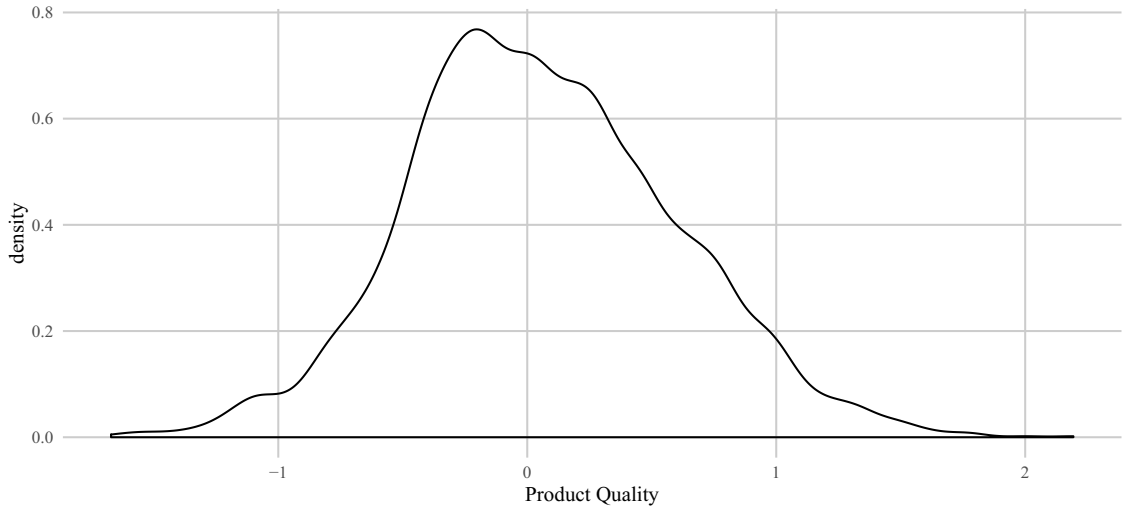


**Figure 1.2:** *Quality histogram for example category*

### 1.3.2 Effects on precision

We first present evidence of sample variability as a function of sample size by plotting the fixed-effect estimates from regression (1.2) run on two non-overlapping subsamples, where we vary the size of each of the subsamples as a fraction of the original full dataset. This illustrates the extent to which two identical firms, represented by the two subsamples, would have agreed in their quality assessment of different products.

We compare sample variability along two dimensions. First, as stated, we vary the fraction of the original data used. Second, we perform a comparison across product categories, which vary by how many consumers visited the corresponding catalog pages in the original data as well as by how product impressions break down over products; an impression is recorded for a product whenever it appears in a menu on a page visited by a consumer. To this end, we construct a measure of the extent to which a particular product category is dominated by its head or tail products as well as of the traffic a particular category saw. Let $N_c$ be the total number of impressions for a particular product category, and let $s_j$ be the fraction of product impressions that were received by product $j$. Then, $\phi_c = N_c \cdot \sum_j s_j^2$, the Herfindahl index of product impressions in category $c$ multiplied by the total number of product impressions

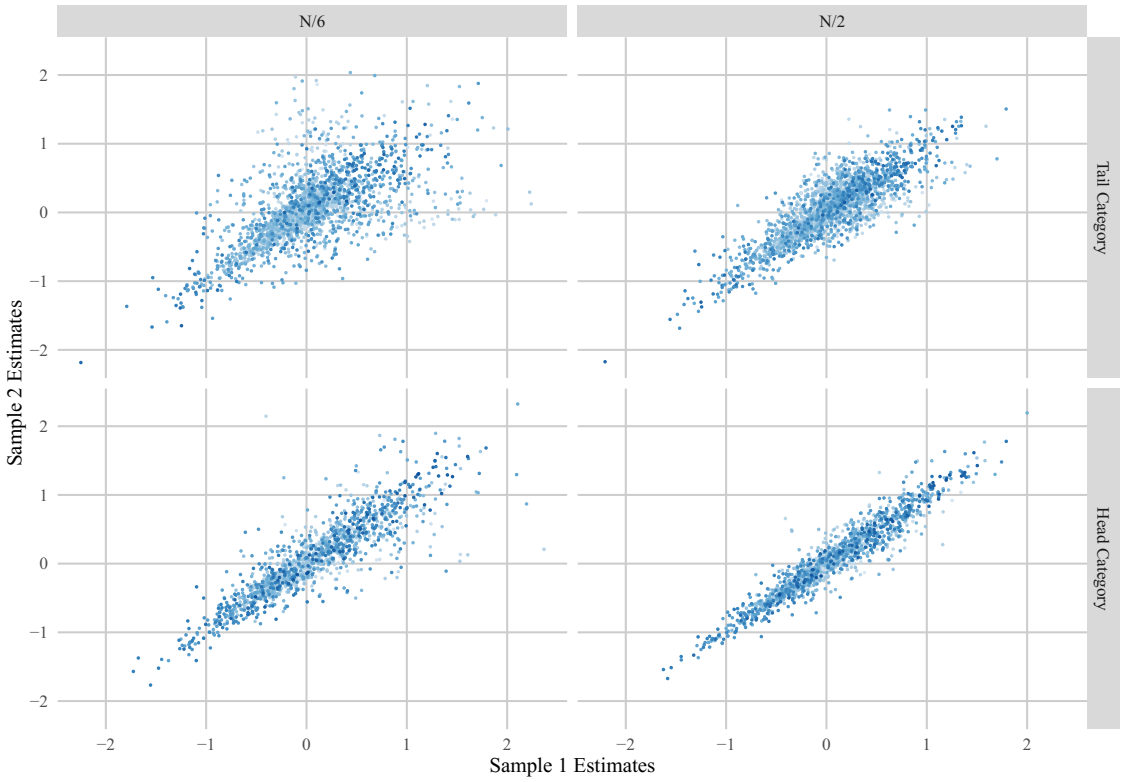


**Figure 1.3:** *Variability in quality estimates for example categories and different fractions of training data*

in that category, is exactly the expected number of impressions we would get if we drew a product impression at random from its empirical distribution over products and looked at the total number of impressions the associated product received in our sample. This measure is proportional to the number of consumers in a particular product category and the concentration statistic $\sum_j s_j^2$, meaning that it will be higher for product categories which are either dominated by its head products or see more traffic. We think of the measure as useful because it indicates how many observations the parameter estimate of the average product being displayed is based on.

Figure 1.3 provides such an assessment of sample variability in the form of a 2-by-2 plot with a less concentrated (tail) category on the top contrasted with a highly concentrated (head) category on the bottom (for roughly equal $N$), using one sixth of the original data on

the left and one half on the right. In each of the four displays, a dot represents a product, and its first and second coordinate are equal to the estimates from two different samples. We observe in the top left display that for a tail-dominated category, the dispersion in quality estimates is substantial when using only one sixth of the data, with many of the observations far off the 45 degree line along which there would be perfect agreement between the two samples. For the optimization problem (1.1), it is of greatest relevance to identify and properly order the highest-quality products. While the estimates overall roughly align along the 45 degree line, the signal-to-noise ratio substantially deteriorates when one zooms in on the right tail of each of the four displays, where the highest-quality products are located. As we triple the sample size for the tail category to one half of the original data and consider the top right display, we see a substantial improvement in precision, as evidenced by much lower dispersion from the 45 degree line. We see a similar improvement as we move from tail to head category for each of the two data fractions in the bottom half of the plot, though some dispersion still remains.

Next, we plot the view-weighted correlation between samples as well as average standard error in product quality estimates for each category against our statistic $\phi_c$ (figure 1.4), again for the two data sizes of one sixth and one half of the original data. In addition, we show how the standard error in the estimates compares to the standard deviation in the quality estimates at the class level. The latter variation pools true underlying product heterogeneity and estimation error, although the estimates have been shrunk in order to reduce the effect of the latter. While its variance is increasing in $\phi_c$, the interpretation is thus not entirely straightforward. We note, however, that the categories with the highest precision are actually the ones with the largest variance in the distribution of estimates, which implies that differences in the latter cannot be driven by estimation error. Rather, it may reflect either differences in true underlying heterogeneity or the fact that estimates based on more observations are shrunk less.

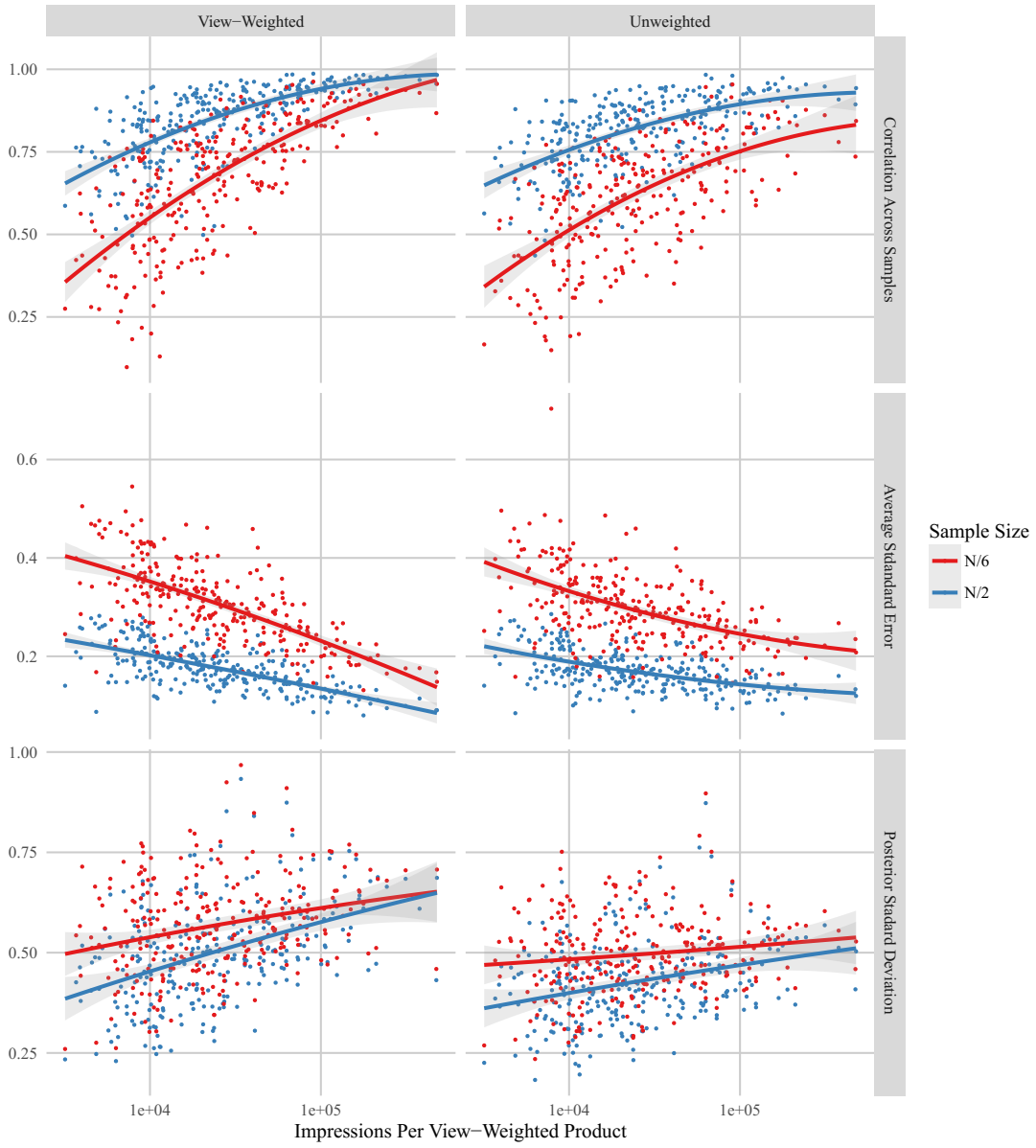


**Figure 1.4:** *Variability in estimates by average number of impressions of products displayed ($\phi_c$)*

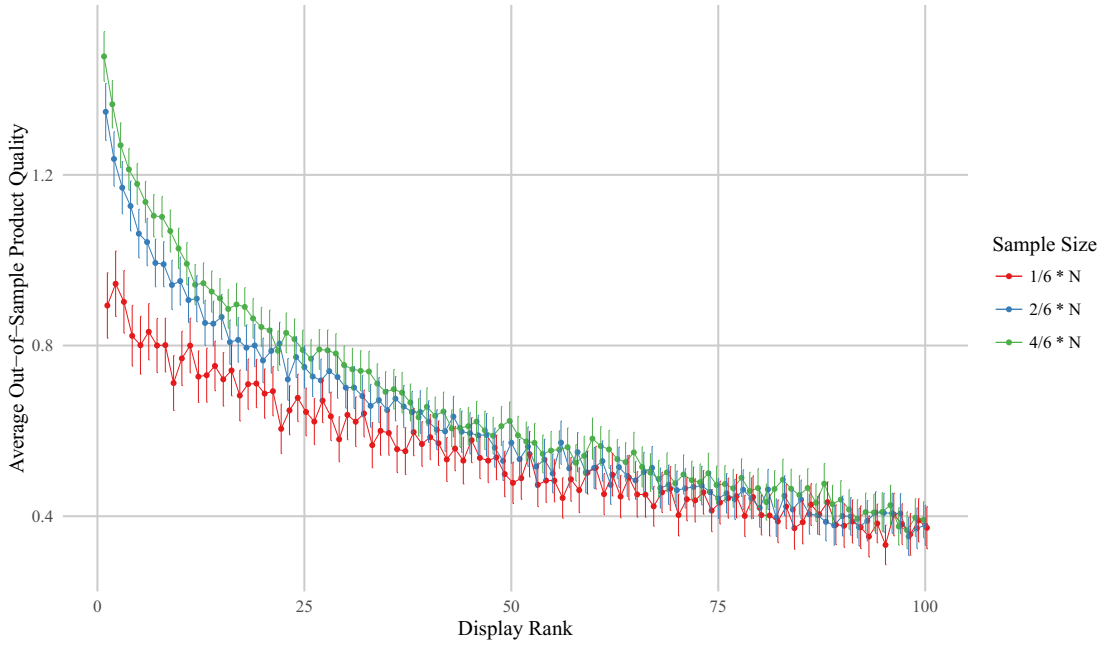


**Figure 1.5:** *Average out-of-sample quality by position for sort order based on different training data sizes*

### 1.3.3 Effects on clicks, purchases, and repeat visits

We now analyze how firm competitiveness varies as a function of the available data. To do so, we compare outcomes in hold-out data for firm decisions taken based on estimates from training datasets of different size.

We begin by calculating the quality of a counterfactual ranking chosen based on estimates from samples of different size. To this end, we estimate logit regression (1.2) on training data consisting of one sixth, one third, and two thirds of the original data as well as on hold-out data containing one sixth of the original observations. We then construct a ranking of the top 100 products according to the different training data estimates to model the solution hypothetical firms of different sizes would have used for optimization problem (1.1); this corresponds to the first two pages of products for a given category. For each product chosen, there is now a hold-out quality estimate which we can use to analyze the quality of the ranking. In figure 1.5, we plot an average across product categories of said quality, weighted by the

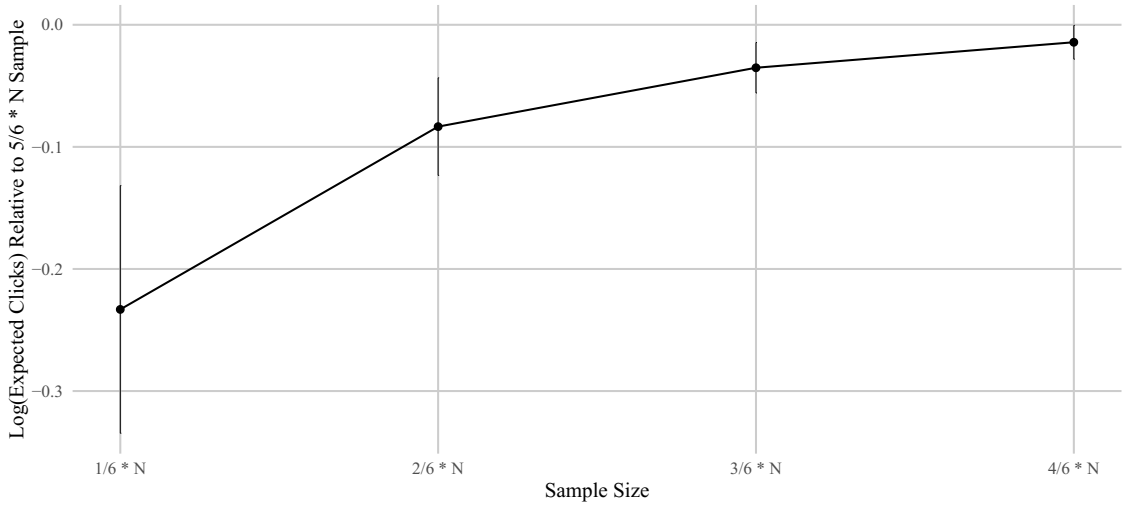


**Figure 1.6:** *Gain in expected number of clicks from larger training data*

respective number of consumers, for each of the 100 positions. The values on the y-axis are our estimates for the fixed effects that enter the logit link function in (1.2) relative to a mean of zero. It is evident that larger samples are particularly powerful in identifying the top products. We see that the effect decreases by position, and that the gap between one sixth and one third of the sample is much larger than that between one third and one sixth, though it still remains economically significant.

Figure 1.6 shows the average implied loss in the number of expected clicks relative to using five sixths of the data. Going from just one sixth of the data to one third brings an improvement of roughly 14 log points, while doubling the data size again to two thirds yields an additional gain of 8 log points.

Having documented a relationship between the size of training data used and quality supplied, we now turn to the effect of quality on different outcomes. Figure 1.7 plots changes in the realized number of clicks in hold-out data against changes in the expected number of clicks according to model estimates while controlling for granular time and other effects. The slope is large for both new and existing customers and very close to linear as seen by how well the averages of 20 equal-sized bins align. At roughly .4, though, it is quite different

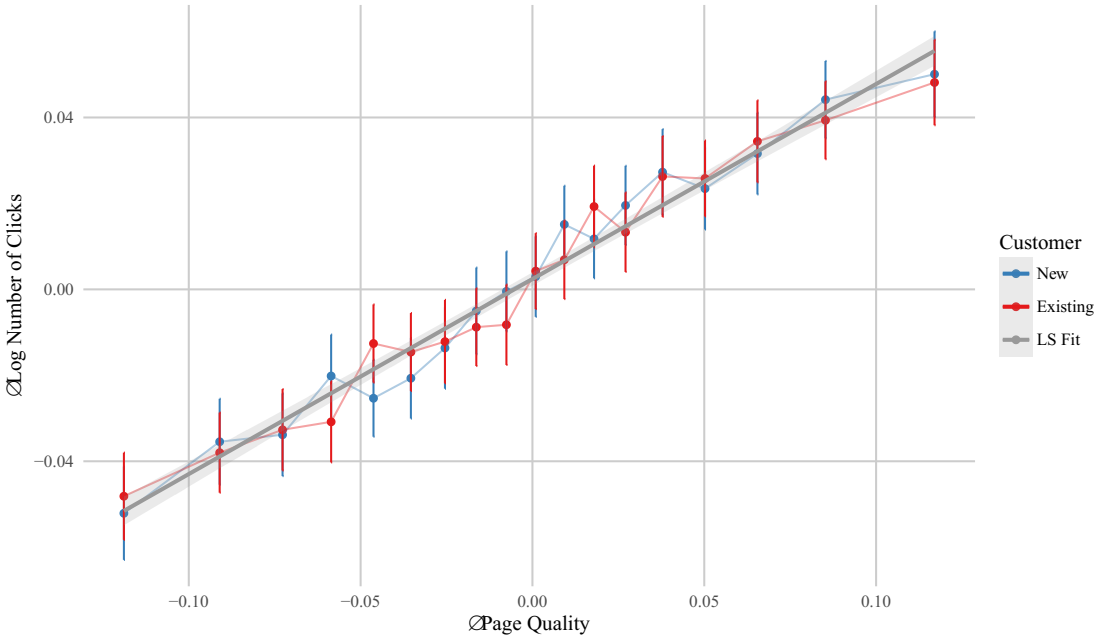


**Figure 1.7:** *Actual number of clicks against page quality/predicted number*

from the perfect prediction scenario of a unit slope. Possible reasons include misspecification, estimation error, and consumers finding other ways to view products. Overall, however, changes in the predicted quality strongly correlate with the actual number of clicks received.

Finally, we consider the effect of changes in the quality of a catalog page on purchases and repeat impressions. For 20 equal-sized bins, figure 1.8 shows the effect of this quality for new and existing customers on the probability of placing an order (top) or looking at an additional page (both in any product category). We see mostly increasing relationships, with larger slopes for new customers for whom we only consider the very first catalog page they saw in our data, which suggests that consumer beliefs are most sensitive to initial experience, and that continuing customers are either generally more loyal or have already accumulated a stock of positive experiences, making them less sensitive. The effect appears to persist over time, even beyond one week, though at that point, admittedly, the corresponding regression estimates, presented in table 1.1, become relatively noisy.

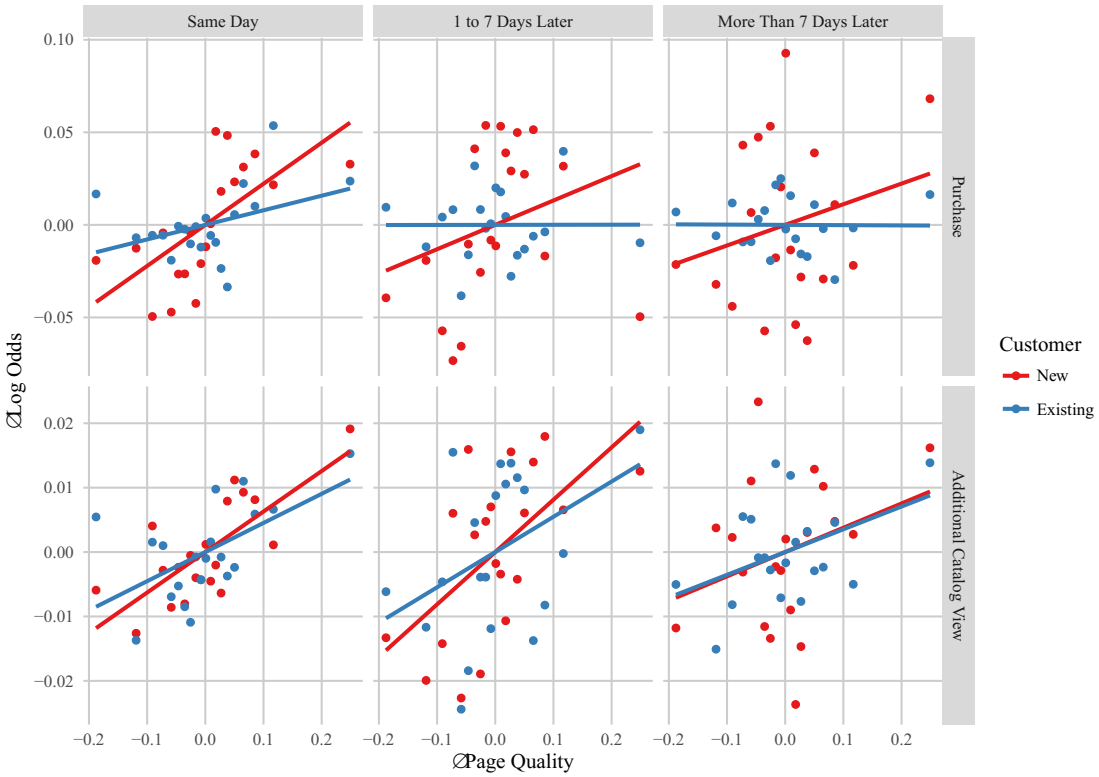


**Figure 1.8:** *Purchases and repeat impressions against quality*

| Customer Type | Same day | 1-7 days after | >7 days after |
|---|---|---|---|
| | | Purchase | |
| New | 0.210*** | 0.079 | 0.090 |
| | (0.050) | (0.078) | (0.084) |
| Existing | 0.045 | 0.012 | -0.009 |
| | (0.042) | (0.047) | (0.044) |
| | | Additional search | |
| New | 0.064*** | 0.054 | 0.022 |
| | (0.014) | (0.031) | (0.029) |
| Existing | 0.043** | 0.057** | 0.042** |
| | (0.014) | (0.018) | (0.015) |
| | | Number of clicked products | |
| New | 0.393*** | | |
| | (0.012) | | |
| Existing | 0.411*** | | |
| | (0.011) | | |

*Notes:* This table reports coefficient estimates on the effect of an out-of-sample estimate of the quality/expected number of clicks of a catalog page by customer type. All regressions control for category and customer type specific time trends via splines with 5 and 10 degrees of freedom, respectively. Standard errors – clustered at category-day level – are reported in parentheses. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

**Table 1.1:** *Effect of out-of-sample quality prediction on economic outcomes*

## 1.4 Discussion

Our approach to modelling product quality is more closely related to empirical models describing differentiated product demand in product space (e.g., Hausman, 1994) rather than characteristic space (e.g., Berry, 1994, Berry *et al.*, 1995), which is perhaps the more common choice in industrial organization and related fields today. In our setting, however, the primary objects of interest are measures of product quality, which motivates estimating these explicitly. In addition, the sort order algorithm used by the firm in practice also treats product quality levels as individual parameters rather than combinations of preferences over characteristics. Part of the problem is likely the difficulty of modelling quality as a function of covariates for millions of products across a wide range of different product categories, but more importantly, many of the products are very strongly differentiated based on tastes, and these are difficult to capture on the basis of covariates alone. It would nonetheless be interesting to consider an approach that shrinks product quality levels towards a function of covariates; while this should improve precision for any size of dataset, it is not entirely clear whether it would be more beneficial for firms with more or less data. Furthermore, a covariate-based approach would run the risk of reducing the variety of top-ranked products in terms of characteristics.

In competitive settings, the importance of data may also depend on whether these data and the resulting optimization decisions are privately or publicly observable. For instance, in the setting at hand, firms are able to observe the sort orders chosen by competitors, which allows for social learning and may reduce informational asymmetries between firms. The usefulness of this information will then depend on the overlap in product catalogs between retailers as well as on the extent to which they specialize and cater to different consumer preferences. Firms may also experiment to learn qualities resulting in statically suboptimal actions, which complicates social learning for their competitors in this setting because outcomes are privately observed and inference can only be drawn from actions played. Observability will also be affected when firms tailor or personalize their sort order by conditioning on covariates or past behavior (e.g., Fradkin, 2015). Depending on the structure of demand, this will further increase the importance of having large amounts of data, since quality needs to now be

estimated conditionally, and individual cells of data can be quite small.

Finally, we note that we have only considered returns to data while keeping the available technology fixed. Many technology firms in particular have hired economists, computer scientists, and data scientists to solve complex optimization and estimation problems. In many cases, these solutions will scale relatively well, so that the associated fixed costs will amortize significantly faster for larger firms. As a result, these firms may command better infrastructure and human capital and therefore be able to use their data resources more efficiently.

# Chapter 2

# Returns to Data Scale and the Impact on Competition: Evidence from Online Retail

## Introduction

In many markets, the competitiveness of firms is directly related to the amount of information they have, in particular as firms collect more and more data. An otherwise identical firm with more customers will likely have more precise information about the demand for its products. Search engines and online retailers employ sophisticated algorithms that use logged visitor data to determine which search results and products to display. Due to the often vast number of options available, there are potentially non-trivial returns to additional data even at large sample sizes.

Using actual firm data from a relevant setting, this paper presents and estimates a model to address two questions important for business and regulators: what are the returns to data scale, and when can they act as a strong barrier to entry? We present a model of a firm that needs to rank the products it sells to consumers online and learn about their quality through revealed consumer preference. Our model includes position effects, the distribution of product

quality, and a set of dynamic parameters, and we estimate these using firm data, including the clickstream data firms typically use to optimize their product ranking. This provides us with both the relevant parameters of the firm's objective function and its learning technology. We propose an adapted version of Thompson Sampling in which we can control the degree of experimentation to allow firms to dynamically optimize their product orderings. We find that myopic play is optimal in the sense that it maximizes the average payoff the firm receives under a steady-state distribution. We provide evidence that this is not due to a general lack of performance of the proposed method, but rather, that the reason lies in the nature of the economic problem and our parameter estimates.

Our model allows us to understand the impact of data on a firm's business and optimal strategies. Even though the statistical return to additional data becomes smaller as the size of the available data increases, the economic return can remain significant, as the data of the marginal consumer is applied to improve the product ordering for a larger and larger base of inframarginal consumers. For the firm in our data, we estimate an elasticity of the expected number of product clicks per customer with respect to the number of customers of 0.024. This implies that for every additional click provided by an additional consumer, her contribution to the quality of the product ordering displayed to all consumers yields an additional 0.024 clicks. Under assumptions on the firm profit function, we can calculate the reduction in optimal mark-ups that results from the marginal consumer's contribution to the firm's stock of data; under one parametrization, we find a reduction of mark-ups of 3.6% for an example price elasticity of -3.

We use the estimated learning technology to map out the minimum viable size for entrants to compete in a market in which consumers choose firms based on quality and firms learn about quality through revealed consumer preference. In particular, we model the effect of consumer quality sensitivity and market size. Our estimates predict thresholds for an entrant's initial market share, which we verify through simulation. The threshold is higher the more sensitive consumers are, and for a given sensitivity, it is lower the larger the market, not just as a market share, but in absolute terms as well.

There are three main reasons to use a structural model alongside the descriptive analysis presented in chapter 1. First and foremost, we are interested in understanding the role of competition under different configurations of firm asymmetries. Since we only observe data for one firm, this is outside the scope of a purely descriptive analysis. Second, firms of different size may optimize differently, in particular in their trade-off of exploration and exploitation, and the data we observe is sampled from a firm of one given size. Third, and this is perhaps the motivation most peculiar to the learning setting, the information set of a firm is endogenously determined, where the new information obtained in any period reflects the information from previous periods and thus the size of the firm.

More specifically, even absent any demand-side dynamics that result in the feedback loop mentioned earlier, firms with more existing data may not only incur higher period profits because they can provide a better ranking, but this ranking also affects which products' quality they learn about the fastest. As a result, a larger firm with more existing data may also more efficiently search the space of product qualities. The researcher, on the other hand, only observes outcomes given the ranking, or, design matrix, chosen by the firm in the data. Note that the issue is not that a smaller firm – which we simulate by using a smaller fraction of the available data – would draw any direct inference from the ranking of products in the data, but that it will obtain a lot of information about relevant high-quality products that it would otherwise not have known to be as relevant to learn about.

In several instances, competition authorities have considered data and scale effects in their evaluation of conduct or proposed deals. The Department of Justice concluded in 2010 that the Search Agreement between Microsoft and Yahoo would increase Microsoft's performance and thereby competitive pressure in the market by providing it with greater scale and larger amounts of data, which are likely to matter particularly for tail queries.[1] On the same

[1] "The search and paid search advertising industry is characterized by an unusual relationship between scale and competitive performance. The transaction will enhance Microsoft's competitive performance because it will have access to a larger set of queries, which should accelerate the automated learning of Microsoft's search and paid search algorithms and enhance Microsoft's ability to serve more relevant search results and paid search listings, particularly with respect to rare or 'tail' queries. The increased queries received by the combined operation will further provide Microsoft with a much larger pool of data than it currently has or is likely to obtain without this transaction. This larger data pool may enable more effective testing and

matter, the Europe Commission finds, in its evaluation of studies submitted by Microsoft and Yahoo, that while "for the most frequent queries, the overall gap between engines is very small," Google appears to perform better in terms of relevance for some queries (European Commission, 2010). However, this "does not provide evidence that scale leads to higher relevance for users since the above studies do not take into account the technology of the different search engines which are not related to scale." Our model allows us to do precisely that: we can hold fixed the technology, both statistical and in terms of data collection, between firms and single out the effect that arises from size.

An FTC staff report (Wall Street Journal, 2015)[2] discusses the role of data and scale in the search engine market, focusing on the market leader Google and Microsoft with its search engine Bing. Both parties agree on the existence of a "virtuous cycle" that also includes publishers and advertisers. Disagreement arises over the extent to which the scale of data is a factor at the current size of both firms, given that there are "substantially 'diminishing returns' ": "[t]he main bone of contention between Google and Microsoft is where on this scale curve Microsoft currently operates. [...] neither party can identify a fixed number of queries or ads that constitutes the 'minimum efficient' point of operation."[3] We use the estimates of our model, which of course come from a different setting, to identify this point as a function of the market environment.

Disagreements persist over whether to call the phenomenon we analyze a "network effect" or whether it is simply an instance of learning by doing. Stucke and Grunes (2016) refer to it as a "network effect, which arises from the scale of data: the more people [...] contribute data, the more the company can improve the quality of its product, the more attractive the

---

thus more rapid innovation of potential new search-related products, changes in the presentation of search results and paid search listings, other changes in the user interface, and changes in the search or paid search algorithms. This enhanced performance, if realized, should exert correspondingly greater competitive pressure in the marketplace." (Department of Justice, 2010)

[2]The FTC inadvertently sent an unredacted partial version of the report, containing only the even-numbered pages, to The Wall Street Journal in response to an unrelated Freedom of Information Act request.

[3]"According to Microsoft chief economist [...] Susan Athey, Microsoft's search quality team is greatly hampered by having insufficient search volume to conduct experiments. With improved search quality, particularly for 'tail' queries, Bing asserts that it will be better positioned to compete with Google for users. [...] Bing is at the lower part of the scale curve where 'each percentage point is critical.' "

product is to other users, the more data the company has to further improve its product [...];" the authors use search engines as a specific example. Hal Varian (Bruegel, 2016), at the time chief economist at Google, argues that this is a pure supply-side phenomenon of learning by doing, and therefore not a true network effect, neither direct nor indirect. It may be difficult to decide where to draw the line: it is perhaps less controversial that the business of Waze, a navigation app that bases recommendations on users' driving data and real-time traffic updates, is subject to network effects, and that it needs to achieve sufficient scale in a market in order to be viable from a product quality perspective. The algorithms of search engines and online retailers similarly base their recommendations on the accumulated experience from user data.

The online sector is perhaps the most obvious but certainly not the only part of the economy in which data is likely to matter for market outcomes. For instance, insurers and lenders rely on rich models predicting expected losses when underwriting policies, and larger firms with more historical data on which to train these models may be able to set prices that more accurately reflect underlying risk. Small firms with less historical data on which to train their models may end up with a less desirable population of risks because of their less accurate models, a form of winner's curse (General Insurance Research Organisation, 2009). Firms with more accurate models may therefore be able operate at a lower level of mark-ups for the "good" pool of consumers to cover their fixed costs and as a result further increase their market share and thereby share of data in the market.

Tambe (2014) studies the returns on investment in big data technology and skilled labor, a complement to the data assets at the heart of this paper. He finds that between 2006 and 2011, investments in Hadoop – the type of system clickstream data such as used in this paper are typically stored in – were associated with a 3% faster productivity growth. Importantly, this relationship only holds for firms with significant data assets.

Outside economics, there have been several studies on the effect of sample size on prediction accuracy. On a collection of publicly available datasets created from online and offline human

behavior,[4] de Fortuny *et al.* (2013) find substantial improvements from increased sample sizes, with marginal increases even to massive scale, leading them to "the observation that firms (or other entities) with massive data assets may indeed have a considerable competitive advantage over firms with smaller data assets." In a study comparing classic logistic regression models to decision trees, Perlich *et al.* (2003) find greater returns to sample size for the latter, suggesting that the importance of having large amounts of data may hinge on and be fueled by the development and adoption of modern statistical techniques. In another learning curve analysis, Cho *et al.* (2016) train convolutional neural networks on CT images and find that the improvement in classification accuracy from using a larger fraction of the data available to them varies by body part but is generally large.

The empirical relevance of our analysis is predicated on the existence of ranking or position effects, whereby a product will receive more clicks if it is displayed near the top of the ranking. This is intuitive and almost mechanical: all consumers will see the set of products at the top of the first page, whereas they need to scroll down or click through to another page in order to see products that are ranked further down. Economically significant position effects have been documented in a variety of settings: Ursu (2016) analyzes a field experiment on hotel searches run by a travel intermediary that randomized product positioning; Narayanan and Kalyanam (2015) and Goldman and Rao (2014) find position effects in search advertising using a regression discontinuity design and repurposing existing business experimentation, respectively; Athey and Imbens (2015) find heterogeneous effects in an experiment that moves Bing search results from the first to the third position; Feenberg *et al.* (2017) find that NBER Working Papers listed first in its weekly announcement emails on new papers receive 25% more cites, even though the order is plausibly exogenous.

In our setting, the retailer needs to rank potentially several thousand products. Our empirical problem is thus different from the studies mentioned above, which tend to consider position effects only over a relatively limited range of positions.[5] Instead, we need to know

[4]Examples are classifying the gender of a user in the Yahoo Movies dataset based on which movies they rated and predicting which Flickr pictures users rate highly.

[5]The actual number of options available to consumers there may still be very large, such as in the studies

effects such as moving a product from position 1,000 to position 100, since this will inform how the overall quality of a sort changes as products of different qualities swap positions as well as how much more the retailer learns about the more highly ranked products' quality through revealed consumer preference. In light of the studies above, it is of little surprise that we find position effects not just among the top positions, but especially as one moves further down the ranking.

While for a small number of options, position effects are sometimes interpreted as the outcome of a signalling game or behavioral biases (e.g., Feenberg *et al.*, 2017), a search cost interpretation is plausible when the set of options is so large that it is unrealistic for a consumer to review all the available options, as in our setting. Since our focus is on the supply-side problem, it is less relevant to know the precise micro foundations behind the position effects, so long as we can estimate how they affect the firm optimization problem, which consists providing and learning about quality.

The paper is organized as follows. Section 2.1 introduces an estimable model of the firm's ranking problem. Section 2.2 examines firms' optimal learning strategies in a dynamic environment for the model estimated in section 2.1. Sections 2.3 and 2.4 use the same estimated model and the conclusion from section 2.2 to compute firms' returns to data and determine under which conditions data creates prohibitive barriers to entry. We conclude with a brief discussion.

## 2.1 Estimating a model of the ranking problem

### 2.1.1 Overview

The goal of this section is to develop a model that allows us to compute the returns to data for firms and simulate counterfactual dynamics to understand conditions under which data creates barriers to entry. We start with the per-period profit function of the firm, which is simplified but resembles proxies used by firms in practice. This already allows us to discuss

on hotel searches and Bing.

the static profit and dynamic learning incentives that are central to our economic question. We specifically turn to firms' dynamic strategies in section 2.2.

To recover a set of parameters that can be plugged into our economic model, we estimate a closely related econometric model that allows us to control for some of the intricacies of our empirical setting. This allows us to make useful simplifying assumptions in our economic model without having to impose them in our empirical estimation. We discuss how the economic and econometric model fit together when we show how to operationalize the economic model by parametrizing it on the basis of the empirical estimates. To provide some reassurance, we show that our economic model, when configured in such way, fits a number of data patterns quite well.

### 2.1.2 Empirical setting and data

Our empirical setting and data are the same as in chapter 1, which provides a more detailed description of both. In this paper, we discuss the different types of data and variation used in the respective sections on estimation.

An online retailer sells a large number of products and needs to decide how to rank these on its catalog pages. The firm does so on the basis of clickstream data containing the logged website interactions of its visitors. Specifically, the firm looks at which products were historically displayed to consumers and which ones received the most clicks relative to their level of exposure. This is a regular process which updates daily. Because of the large number of diverse products, of which most consumers only see a very limited subset, this is a problem in which having a large amount of historical data available is vital for statistical precision and economic outcomes.

### 2.1.3 Static firm problem

We assume that a firm's per-period objective function is the expected number of product clicks made by a consumer, which depend on the ordering $\rho$ that these products are displayed in. We suppress all unnecessary subscripts until they become relevant in later sections and

write the objective function as

$$\pi(\rho) = \mathbb{E}\left[\sum_{j=1}^{J} Y_j\right], \tag{2.1}$$

where $Y_j$ is an indicator for whether product $j$ is clicked.

We write $Y_j = Y_j^{(1)} \wedge Y_j^{(2)}$, with $Y_j^{(1)}$ equal to one if product $j$ appears on a catalog page viewed by the consumer in a position that she pays attention to and zero otherwise, while $Y_j^{(2)}$ is equal to one if the consumer likes the product enough to click on it provided she has taken notice and zero otherwise. For the product to be clicked, the consumer needs to both find and like the product, so that $Y_j = 1$ if and only if both $Y_j^{(1)} = 1$ and $Y_j^{(2)} = 1$. Any ranking $\rho$ is a permutation of product indices such that $\rho(j)$ gives the position that product $j$ is placed in. We assume that $Y_j^{(1)}$ and $Y_j^{(2)}$ are independent, and that $\mathbb{E}\left[Y_j^{(1)}\right] = \gamma_{\rho(j)}$ and $\mathbb{E}\left[Y_j^{(2)}\right] = \mu_j$, so that the probability that a product is found depends only on the position it is placed in but not directly on its identity, whereas the probability that a consumer likes the product depends only on its identity. This implies that $\mathbb{E}[Y_j] = \gamma_{\rho(j)}\mu_j$, meaning that the probability of a consumer clicking the product in question is multiplicatively separable in the two underlying probabilities.

Crucially, the firm does not know $\mu$, which we term the vector of product qualities, but has beliefs $\mu \sim p(\cdot)$ for a non-degenerate distribution $p(\cdot)$. In contrast, we assume that the firm knows the vector of position effects $\gamma$ with certainty. Thus,

$$\pi(\rho) = \sum_{j=1}^{J} \gamma_{\rho(j)} \mathbb{E}_\mu[\mu_j], \tag{2.2}$$

where $\mathbb{E}_\mu[\cdot]$ is the expectation obtained by integrating with respect to the belief distribution $p(\cdot)$ of $\mu$.

In practical terms, the firm decides on an ordering in which to display its products to consumers. A consumer navigates to the section of the website for the category of products she is interested in. Provided she does not filter her search by product attributes or the like, she is displayed the set of products that are at the top of the firm's ranking. If she scrolls down on the page or navigates to the next of potentially many pages of products from this

category, she will see additional products. Just like for all products previously seen, she has the option of clicking on these to obtain more information and potentially purchase them, a decision we do not model here. Alternatively, she can filter the products displayed by attributes, availability options, or price, which restricts the set of products to a subset of the full catalog. While we describe product search at an online retailer, this process is quite similar to that of a consumer using a search engine, which typically returns multiple pages of results in list form and allows consumers to refine their search sequentially.

We have in mind a large number of products $J$, potentially several thousand. It is unlikely that the consumer will actually see all of them by navigating through all the (unfiltered) catalog pages for this category, nor is she likely to seriously consider all the products displayed, be it because of inattention, a bias against products further down on the page, or perhaps fully rational search with search costs. Whatever the reason, this implies position effects by which a given product is typically less likely to be clicked if it is ranked further down. In our formulation as well as in many practical settings, it does not matter for the firm why a product that was ranked in a particular position was not clicked, and so we do not explicitly model the consumer search and decision process, which is not the primary focus of this paper. Indeed, on the supply side, it is probably less crucial whether the position effects reflect economic primitives but rather whether these parameters can assumed to be structural in the sense that they remain constant when the ordering chosen by the firm or any other aspects of the environment change.

In the following, we assume position effects to be weakly decreasing, meaning that if a product is ranked further down, it is less likely to be clicked by a consumer. This assumption is supported by ample evidence in the literature as well as our own estimates in section 2.1.6. It implies that the per-period objective function is trivially maximized by ranking products in decreasing order of their expected qualities under the beliefs held by the firm. In fact, it can easily be shown that another commonly used metric for the quality of a ranking of products or search results, click-through rate,[6] is statically maximized by the same solution (see section

[6]The click-through rate refers to the probability that at least one product is clicked.

A.1.1 for a short proof), although this relies more crucially on the simplifying assumption that there is no systematic heterogeneity in consumer preferences over products, discussed in more detail below. Placing the most popular products at the top appears furthermore as an intuitive and therefore hopefully empirically robust assumption.[7]

To focus on the supply-side learning questions at the core of our paper, we have made several simplifying assumptions on the consumer side. There is no product variety motive, both because we do not model systematic heterogeneity in consumer preferences and because the objective function is additively separable across products. However, from a supply side perspective, a model with preference heterogeneity collapses into our simple multiplicatively separable model if position effects and product qualities are uncorrelated in the distribution of consumer types. In this case, preference heterogeneity is irrelevant for the supply-side problem, and we can simply assume that the firm works with the simpler model presented above, which is not the true model but sufficient for the firm's purposes.[8] If we want to cast the model in terms of a structural consumer choice model, we can think of $Y_j^{(2)}$ as an indicator that the utility of clicking on product $j$ upon having found it is positive, and under straightforward assumptions, we can map beliefs $p(\cdot)$ into beliefs on product utility levels.

The dynamic trade-off faced by the firm is obvious even in the static model: it can exploit by prioritizing products with high expected quality or explore products with high variance in beliefs by placing them in high ranks, such that the firm receives signals on their quality relatively quickly. Before we discuss firms' dynamic strategies in detail in section 2.2, we present our estimation of the relevant static parameters. Readers who wish to skip the details of our estimation can proceed directly to section 2.1.7.

[7]Going beyond clicks, if expected profit at the product level is constant across positions, the model can be trivially extended by multiplying product quality with expected profit. This results in an objective function that is directly interpretable as gross profit, and provided that expected profit per click is constant and known, the firm's beliefs result from a trivial transformation of the beliefs discussed above, leaving the learning problem essentially unchanged. We proceed with our simpler model (2.2) since it is used in practice and allows us to focus both analysis and presentation on the central theme of this paper.

[8]In the presence of product heterogeneity, the firm could learn more efficiently by accounting for the resulting covariance structure in its observed data.

### 2.1.4 Empirical model

The essence of the static version of the firm problem is that products receive more attention the higher they are ranked in their category, and that firms should therefore prioritize higher-quality products. Hence, the two central empirical objects in the static problem are the position effects and the distribution of product qualities.

In our estimation, we use logit models at the consumer-product level for the binary click decision. We include fixed effects at the product level to capture product qualities, an additive position effect, as well as a set of controls to control for price, time, and other effects. These logit models imply a slightly different functional form than our economic model (2.2), but for small probabilities, they are very similar; in our setting, for any product, the probability that a random consumer clicks on it is at most a low single-digit percentage. In our economic model, if we write $\tilde{\gamma}_\cdot = \log \gamma_\cdot$ and $\tilde{\mu}_\cdot = \log \mu_\cdot$ for the logarithmized versions of our parameters,

$$\begin{aligned} \mathbb{E}\left[Y_j\right] &= \gamma_{\rho(j)}\mu_j \\ &= \exp\left(\tilde{\gamma}_{\rho(j)} + \tilde{\mu}_j\right). \end{aligned} \tag{2.3}$$

Our econometric logit model instead specifies, for consumer $i$ and product $j$, a product fixed effect $\delta_j$, a flexibly specified set of position variables $z_{i,j}$, and controls $x_{i,j}$ that include, inter alia, time effects and log price,

$$\mathbb{E}\left[Y_{i,j}\right] = \frac{\exp\left(\delta_j + z_{i,j}'\lambda + x_{i,j}'\beta\right)}{1 + \exp\left(\delta_j + z_{i,j}'\lambda + x_{i,j}'\beta\right)} \tag{2.4}$$

$$\approx \exp\left(z_{i,j}'\lambda + \delta_j + x_{i,j}'\beta\right), \tag{2.5}$$

where the approximation results from the fact that the numerator is consistently quite small. The close correspondence between (2.3) and (2.5) implies that we can map an estimated distribution of product effects $\delta_j$ into one for the product qualities $\mu_j$ in (2.2) and use estimated position parameters $\hat{\lambda}$ to form the vector of position effects $\gamma$.

We run two different versions of model (2.4) to estimate the distribution of product qualities and position effects, respectively, using data from the same large exemplary product

category. We leverage the same set of regressions run in chapter 1 to estimate the distribution of qualities, described in more detail in section 2.1.5 below, and run a separate regression in section 2.1.6 using slightly different data and experimental variation to obtain unbiased estimates of position effects. We verify that this strategy yields reasonable estimates by comparing its predictions to moments in the data in section 2.1.7.

To obtain consistent estimates of position effects, we use data from an experiment described in section 2.1.6 that randomly displayed different product sort orders to different consumers. Since the experiment is not extremely long, we use a conditional likelihood approach that remains consistent because it circumvents estimating the fixed-effect parameters, which could otherwise lead to incidental parameter bias. Importantly, we use as position variable the overall ordering of all products in the category, which is exactly the decision variable $\rho$ of the firm in (2.2), and which we have separate data on. The products that are actually displayed to a consumer are a function of this ordering, the filters applied by the consumer, the number of catalog pages the consumer navigates through to, and any potential logic that is overlaid on the front end; we observe that information separately in the clickstream data for every consumer. As discussed further in section 2.1.6, the variable used here is more appropriate for the estimation of position effects from a supply-side perspective, and neither variable can be recovered from the other.

For the distribution of product qualities, however, we use actual clickstream data, which is consistent with model (2.2), since the product qualities correspond to the probability that a product is clicked given that it has been displayed to and found by the consumer. This leads to greater precision by eliminating the variance that comes from whether a product was actually displayed to a consumer. In addition, it makes the estimates somewhat more robust to deviations from the assumptions of homogeneous position effects across products and independence between navigation/sort and click decisions, since we can control for what was actually displayed to a consumer. We might otherwise falsely include a type of sort or other heterogeneity that we abstract from in our model in the relevant variance of product qualities that controls the outcomes of our simulations. This clickstream data is also precisely

the data used by the firm's algorithms when constructing its product ordering.

### 2.1.5 Estimating the product quality distribution

Given infinite data, not only the distribution of product qualities but also the quality level of each individual product pointwise could be estimated with arbitrary precision. However, in light of the very spirit of our exercise, this would be unreasonable to assume. If we were to use the density of estimated product fixed effects in specification (2.4) as our estimate of the product quality distribution, this would conflate true heterogeneity in product quality with estimation error, which would likely lead to an overestimate of the former. Instead, our estimation approach needs to take into account that product qualities are measured with error.

A fully Bayesian approach to this problem would typically specify a parametric model for the likelihood of observing the given data along with a prior on its parameters as well as a hyperprior on the parameters of the product quality distribution. For computational reasons, we instead compute a regularized likelihood estimate using a penalized log-likelihood that places a scaled L2-penalty on the product fixed effects. The resulting estimates are the maximum a posteriori for a Normal prior population distribution from which the product qualities are drawn.

We estimate the variance of the product qualities using an approximate method described in Stiratelli *et al.* (1984). The authors propose an expectation-maximization (EM) approach that treats the product quality levels as missing data and alternates between updating the variance of product qualities (M-step) and calculating the expected value of this variance based on the posterior given the previous value of the variance (E-step). The E-step and the M-step are very closely intertwined because rather than computing the expected value of the variance using either numerical integration or simulation based on the posterior, the authors directly calculate the expected value of the quality variance by approximating the posterior of each product quality parameter using the posterior mode. More specifically, each posterior is approximated by a Normal distribution around the posterior mode using the negative

inverse Hessian of the marginal (penalized) log-likelihood as variance. This fits neatly into the framework of maximizing the penalized log-likelihood of the model for different values of the penalty parameter utilized in chapter 1. Following Stiratelli *et al.*, pp. 964-65, the expected value of the quality variance is calculated as the sum of the average squared posterior mode (as an approximation to the posterior mean) and the average posterior variance (approximated by the asymptotic distribution of the regularized likelihood). We plug this value back into our marginal (penalized) log-likelihood to iteratively find the fixed point of this mapping.

This EM approach is relatively straightforward and computationally efficient because it amounts to a simple grid search for the scalar variance parameter and we have already estimated a finely spaced regularization path of posterior modes in chapter 1. Each of these estimates requires very few Newton-Raphson iterations, which are cheap even for a large number of product quality parameters when exploiting the sparsity pattern of the Hessian and the partitioned inverse formula (e.g., Greene, 2004). As a by-product, the diagonal of the negative inverse Hessian gives us the required estimate of the asymptotic variance of each product fixed effect.

This computationally convenient procedure is an approximation, though one that has precedence in the literature, where it is commonly referred to as penalized quasi-likelihood. It trivially becomes precise as the number of observations for each product goes to infinity, but its finite-sample accuracy hinges on the quality of the so-called Laplace Normal approximation to the posterior (e.g., Fitzmaurice *et al.*, 2008, p. 18). We have found this method to work relatively well in small-scale simulations.

Figure 2.1 shows the variance calculated as a function of the regularization parameter. The fixed point of the EM algorithm, which is the intersection of this curve and the 45° line, gives an estimate of the variance in product qualities of 0.529, composed of a variance between the posterior modes of 0.482 and an average within variance of the posterior mode estimates of 0.047. This implies that a product one standard deviation above the median in quality is slightly less than twice as likely to receive a click as the median product when placed in the same position. Figure 2.2 plots a histogram of the posterior modes of product qualities for a

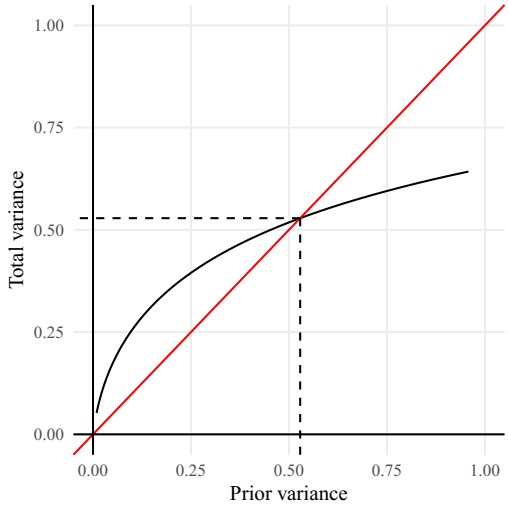


**Figure 2.1:** *EM fixed-point estimation of product quality variance*

prior at the estimated variance. The right half of the histogram is more refined, since each of these estimates is based on a larger amount of data, which goes back to the endogeneity of ranks and the amount of information available for each individual product.

### 2.1.6 Estimating position effects

For many product categories in our data, consumers have several thousand products to choose from. The retailer assigns a ranking to them which allocates products within and across catalog pages. In order to view a product that is further down the ranking, a consumer needs to scroll down, navigate to the next catalog page, or filter results to restrict the number of alternative products.[9] Mainly because of filters, the rank given to a particular product is not necessarily the position in which it is encountered by a consumer.[10] Our estimated position effects are net of the search technology available on the website of the retailer in our data,

[9] In our data, this does not affect the ranking, just the set of eligible products, as for the most part, there is only one global ranking for a category.

[10] Product unavailability is another reason why a product ranking that is different from the one intended by the firm is displayed to the consumer. We use rankings that take into account product availability, which is a (minor) source of plausibly exogenous position variation in our data (see also Conlon and Mortimer, 2013). In our simulations, we abstract from product unavailability.

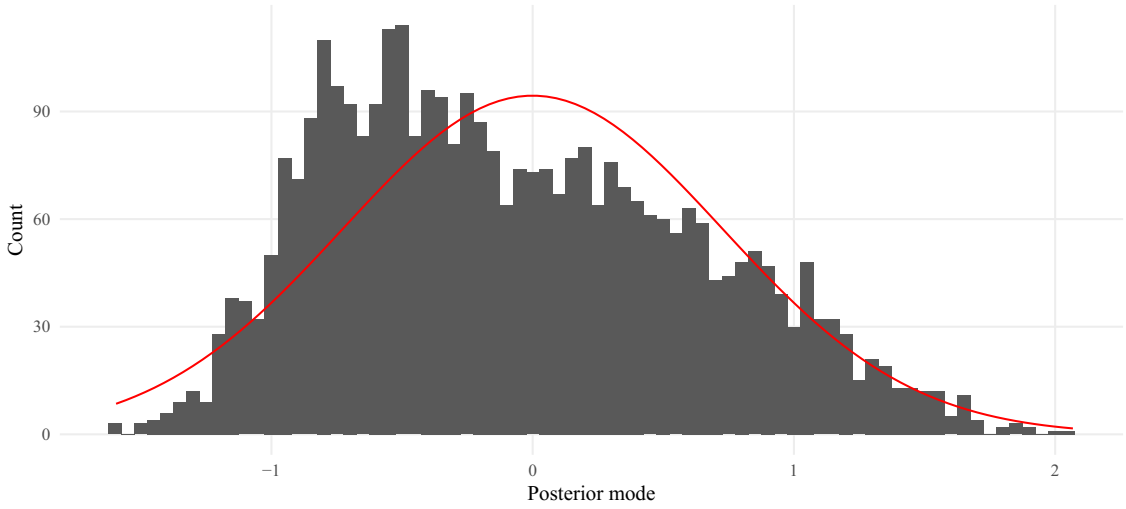


**Figure 2.2:** *Histogram of product quality estimates*

which is representative for the online retail industry. In our model (2.2), the retailer does not care about the position a particular product was displayed in when a consumer clicked it. Instead, it needs to understand the effect of its global ranking, to which consumers may then apply filters as they search for the best product match. Studies that ignore filtering and work with clickstream data on the position that a product was displayed in after consumers filtered may misestimate the effective position effects.

The data required for our regression is different from what is observed in clickstream data. Not only would the clickstream data give only the position a product was displayed in, but we would only know the position of products that were actually displayed to consumers. In contrast, we would for instance not know what would have been available on the next catalog page that the consumer did not choose to navigate to. Whether a product receives such an impression, a prerequisite for a consumer clicking it, is itself a function of the ranking chosen by the firm and an important part of the position effect: a product that is on the 23rd catalog will likely receive few clicks not primarily because consumers do not make many clicks when going to the 23rd page, but rather, because few of them navigate that far in the first place.

Fortunately, we separately have data on the firm's ranking of all products, of which only a

small subset may have been displayed to any one consumer. This allows us to exactly estimate how the decision variable of the firm – the overall ranking – translates into the probability that different products are clicked by consumers, who may or may not use different filters and navigation tools.[11] The regressions run below use this overall ranking set by the firm as the position variable. Of course, the default unfiltered first catalog page displays exactly the ranking that forms the top of the overall ranking.

We exploit variation induced by an experiment the retailer ran. It took all products of a set of categories and randomly split them into sets A and B with equal probability. It then took the product ordering produced by the retailer's algorithm and created two additional orderings, which preserved the original ordering within sets but placed all products in A above those in B and vice versa. Visitors to the website were randomly allocated to one of three treatments arms: the original ordering and the rearranged orderings. We use the variation between a product's position in the original ordering and the position it takes in the arm that moves set A or B to the top, depending on whether the product in question is part of set A or B. Since the experiment is relatively short and the available data on any one product is limited, we use the conditional likelihood approach for the fixed-effects logit model (Andersen, 1970, Chamberlain, 1980) to avoid incidental parameter bias from short panels. We flexibly estimate the position effects using B-splines (e.g., Hastie and Tibshirani, 1990) of order three, with the associated knots chosen as quintiles of the empirical conditional distribution of positions given that a product was clicked.

Figure 2.3 plots the estimated position effects, relative to the first position with a pointwise 95% confidence band, where position is on a log scale, for the same product category as the quality estimates. There is a significant drop right for the first products, which gradually

[11] Our simulations will take the position effects as structural and abstract from the possibility that the ranking influences consumer filtering and search behavior. While this rules out very interesting micro-level consumer behavior, we think that the variation in our data is sufficiently close to what we intend to simulate – in other words, we do not go that far out of sample – that we feel comfortable with this assumption. It is perhaps a bigger concern for the field experimented used by Ursu, which completely randomized product rankings. If this directly affected consumer behavior, which appears difficult to check with the data available there, one might worry about the external validity of the estimated effects (for instance, consumers resorting to filters or even abandoning the site due to lower than expected quality at the top if consumers are used to seeing optimized rankings).

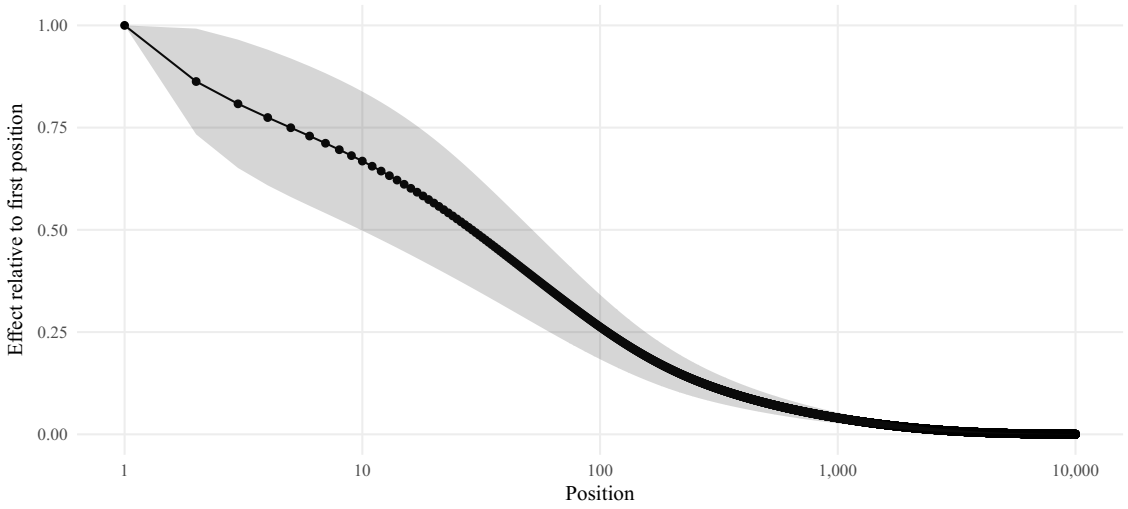


**Figure 2.3:** *Click probability relative to first position*

becomes more pronounced. Relative to being in the first position, a product can expect a fraction of 66.8% of the clicks in position 10, 26.3% in position 100, 4.0% in position 1,000, and 0.03% in position 10,000. We obtain qualitatively similar results for other categories that are of similar size.

### 2.1.7 Operationalizing firm beliefs on product quality

For our supply model, we need a rule for how the firm updates its beliefs $p(\cdot)$ on $\mu$ based on the click data it observes. The conceptually and computationally most convenient approach is to specify a prior that is conjugate to the likelihood of the data. Since the click variable is binary, we work with a Beta-binomial model. The precise form, including how historical data of different vintages is treated, is presented in section 2.1.8. We want the firm to use the true population distribution from which the product qualities are drawn, which means that we need to fit our estimated population distribution to a Beta distribution. We simulate probabilities under the logit model with product qualities drawn from a Normal distribution with the variance estimated in section 2.1.5 and solve for the two parameters of the Beta distribution to fit the simulated mean and variance.

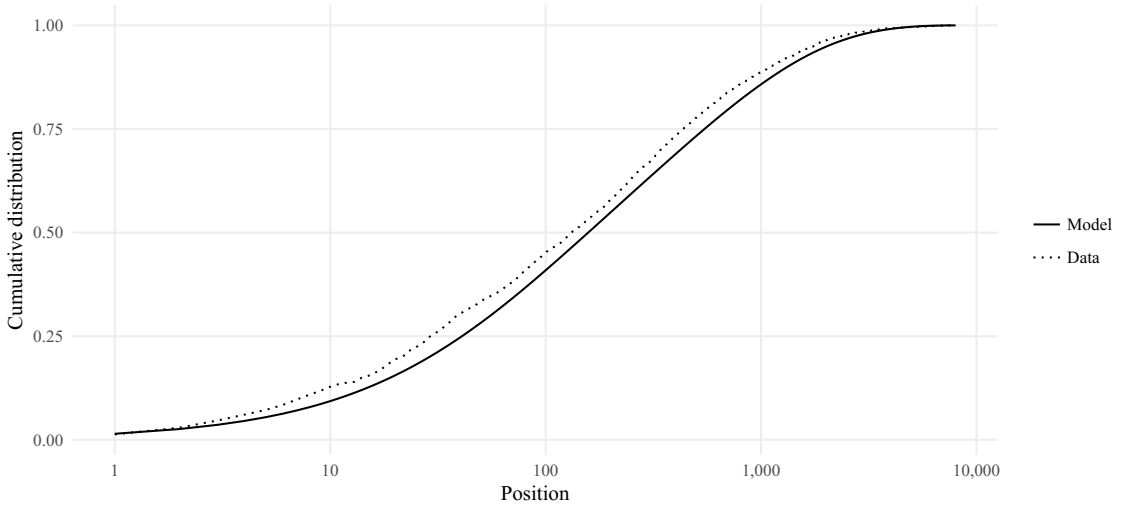


**Figure 2.4:** *Click distribution under model and data*

In order to evaluate our modelling assumptions, we compare the distribution of clicks in the data to that predicted by the model estimated in the previous two sections. We compute the distribution under the model by assuming that products are perfectly sorted.[12] Figure 2.4 plots the resulting cumulative distribution functions for model and data and shows that our model reasonably captures heterogeneity across positions and products.

In the clickstream data, the firm does not necessarily know whether the consumer actually scrolled down and looked at all products that were displayed on a page. Yet, position effects even within a page are very important. Since the Beta-binomial model in its basic convenient form does not allow for any additional controls, we cannot quite match the way that the firm learns from its clickstream data in practice, where it observes clicks for a catalog page of products and controls for position effects within that page when updating its beliefs on product popularity; we do the same in our estimation in section 2.1.6.[13] Instead, we assume

[12]In practice, the click distribution may be more or less concentrated than under perfect sorting. At first thought, the concentration of clicks should be lower under imperfect sorting, since the correlation between product quality and position effects will be smaller. However, the correlation may vary locally across different parts of the quality distribution if it is easy/sufficient data exists to identify the top products but it is difficult to sort the remainder that receives less exposure to consumers. In that case, the model may actually underestimate the concentration of clicks in the top positions.

[13]Updating under the Normal-logit model cannot be done in closed form and requires simulation methods

that the firm observes the variables $\left(Y_j, Y_j^{(1)}\right)$, so that it can deduce $Y_j^{(2)} \in \{0,1\}$ if and only if $Y_j^{(1)} = 1$. Our assumption would be exactly accurate if each catalog page held only one product and the consumer would have to request an additional page to see the next product (which would not be very user-friendly). In our data, however, the products that fit on one page are only a small subset of the overall category – for the example category we use in our estimation less than 1% of the total – so that we consider the assumption in our economic model sensible.

### 2.1.8 Dynamic parameters

There are three main dynamic parameters in our problem, which link together the problem faced by the firm in every period, along with the fact that the firm obviously learns about payoff-relevant parameters from past outcomes. The first two of these parameters pertain to how the firm's stock of information depreciates over time, and these are relevant in both a single-firm and a competitive setting. The third parameter is relevant only in a competitive setting and dictates how consumers choose which firm to patronize based on past quality. We delay introducing this last parameter until section 2.4 because rather than attempt to estimate it, we will determine for which ranges of values we can observe different competitive outcomes.

Empirically, product entry and exit as well as changes in quality of continuing products are important in our data. While the former is directly observed, evidence for the latter consists of changes in the firm's chosen product sort order, fluctuation in the number of clicks received by individual products, and the fact that the retailer uses decaying weights for past observations to optimize the bias-variance trade-off in estimating product attractiveness: the more distant past is less informative for current attractiveness, but using only very recent data would result in large variance in the estimates and therefore adversely impact precision.

Some degree of "depreciation" in information is required for the system to converge to a non-degenerate steady-state distribution or rest point, in which no firm has perfect (or at

such as MCMC, which would be computationally prohibitive in our simulations.

least arbitrarily precise) information about all payoff-relevant parameters in the current and all future periods. This depreciation should not be taken as a literal loss of or diminishing information, but rather that the relevance of the acquired stock of information decreases as time goes on. Entry of new products will introduce new parameters for firms to learn and prevent any firm from having (practically) perfect information forever. To facilitate analysis and interpretation, we prefer the economic problem to be stationary. If the set of products continues to expand, this will not be the case. Instead, we leave the cardinality of the set of products constant by setting the number of new products exactly equal to the (stochastic) number of discontinued products. We specify an identical constant hazard rate $\delta$ for every product in every period, which makes the analysis simpler since exit is a geometric hazard and there is no need to include product age or identity as an additional state variable that matters for firm optimization. The qualities of all new products are drawn from the same population distribution as the original products. We calibrate to patterns of entry, exit, and typical product age in our data by setting $\delta = 0.005$.

In addition to products moving in and out, the qualities of existing utilities can change over time, which will render older data less useful because it no longer comes from the current data generating process. Firms apply less weight to an observation the older it is and set these weights to optimize a trade-off between bias and variance. To operationalize the weights used by the firm in our data, which we know, we modify the Beta-binomial updating so that for an initial prior $\mu_j \overset{iid}{\sim} \text{Beta}\,(\alpha_0, \beta_0)$,

$$\begin{aligned}
\alpha_{j,t} &= \alpha_0 + \sum_{h=1}^{\infty} \theta^{h-1} \sum_{i=1}^{N_{j,t-h}} Y_{i,j,t-h} \\
&= \alpha_0 + \theta\left(\alpha_{j,t-1} - \alpha_0\right) + \sum_{i=1}^{N_{j,t-1}} Y_{i,j,t-1}
\end{aligned}$$

and

$$\begin{aligned}
\beta_{j,t} &= \beta_0 + \sum_{h=1}^{\infty} \theta^{h-1} \sum_{i=1}^{N_{j,t-h}} \left(1 - Y_{i,j,t-h}\right) \\
&= \beta_0 + \theta\left(\beta_{j,t-1} - \beta_0\right) + \sum_{i=1}^{N_{j,t-1}} \left(1 - Y_{i,j,t-1}\right),
\end{aligned}$$

where we set $\theta = 0.9675$, so that weights drop to 50% after three weeks. While this may seem like aggressive discounting in a world where there are returns to data in the form of increased precision, data subtly has a much longer legacy by influencing the sort and thereby the data that is collected in subsequent periods.[14]

Even though we have firms discount older data due to time-varying utilities in practice, we keep qualities constant. We abstain from modelling time-varying qualities for two reasons: first, it is decidedly non-trivial to estimate for us as econometricians from the data observed; second, it would complicate coherent updating of the quality beliefs tremendously. Notice also that $\theta$ is a tuning parameter, and firms with less data would likely choose a relatively larger value, reducing variance at the expense of greater bias. We instead impose it as a global parameter, which is perhaps made more palatable by the fact that since the qualities in our model do not actually change over time, there is no bias. As a result, relative to their individually optimal tuning parameter $\theta$, a smaller firm will experience larger than optimal variance, but, at the same time, it is also spared the greatest level of bias because our model does not include it.

## 2.2 Dynamic learning

In this section, we discuss our model of the firm's dynamic optimization problem. Firms in many industries, particularly online, not only passively learn from historical data, but they engage in active learning by running experiments and choosing actions that allow them to learn which ones are the most profitable. In our setting, firms learn more about a product's quality the more consumers see it and have the option of clicking it. When the firm decides on a product ordering $\rho$ in our model (2.2), it therefore needs to take into account how much information the ordering $\rho$ provides about the value of all potentially optimal orderings.

The dynamic version of our model, in which the firm receives a discounted stream of payoffs as in (2.2), does not have a known optimal solution and is extremely high-dimensional due to

[14]This is precisely the limitation of the descriptive analysis in chapter 1 cited as motivation for a structural model in the introduction to this paper.

the enormous number of possible orderings in every period. Instead, we present a heuristic approach that generalizes a method known as Thompson Sampling. Our method allows the firm to control the degree of experimentation and nests standard Thompson Sampling and Bayesian Myopic play, a form of passive learning, as its two corner cases. Perhaps surprisingly, it turns out that myopic play is optimal in the sense that it maximizes the expected payoff the firm receives under a steady-state distribution. We provide evidence that this is not due to a general lack of performance of the proposed method, but rather, that the reason lies in the nature of the economic problem and our parameter estimates. For clarity, our results also do not imply that firms should never engage in active learning or experimentation, it just appears that for learning about product qualities alone, when position effects are known (perhaps from an earlier experiment), it is better to play the statically optimal solution, which results in a sufficient amount of learning.

### 2.2.1 Set-up of problem and value computation

We consider an environment in which time is discrete and in every period $t$, which we think of as one day, $N_t$ consumers arrive at the firm's website. This formulation allows us to think of $N_t$ as the firm's current size, and we will simply refer to it as firm size in the following. We endogenize the number $N_t$ in section 2.4 but hold it fixed for now. Our leading example will be a firm of size $N_t = 1,000$. The true number of consumers in our data for the product category we focus on is a small multiple of this number, but we set $N_t = 1,000$ for presentation and computation purposes and to preserve the confidentiality of our data. Instead, we adjust the intercept in model (2.4) to match the overall click levels in our data. We can verify that this has no material effect on our analysis. Similarly, we set the number of products in our data to be $J = 10,000$.

We allow firms to display a different product ordering to each consumer, so that $\rho_t$, chosen in every period $t$, is a $J \times N_t$ matrix. While there is no consumer heterogeneity or other static reason for the firm in our model to differentiate orderings between consumers, it is potentially important for learning. In reality, it is common for online firms to direct traffic to different

variants in a randomized fashion.

We simulate outcomes under the data-generating process estimated for our model in section 2.1 and for different strategies that imply orderings $\rho_t$ as a function of firm beliefs $(\alpha_t, \beta_t)$, discussed in more detail below. First, we draw from the distribution of product qualities. These are fixed within each simulation run but vary across simulations. Within each run, in every period $t$, we take Bernoulli draws of $Y_{ijt}^{(1)}$ and $Y_{ijt}^{(2)}$ for all consumers $i$ and all products $j$, where the probability $\Pr\left(Y_{ijt}^{(1)} = 1\right)$ depends only on the rank $\rho_t\left(j, i\right)$ assigned by the ordering $\rho_t$, while $\Pr\left(Y_{ijt}^{(2)} = 1\right)$ is a probability that depends only on the identity $j$ of the product and the initial random draw for its quality.

The resulting chain of period draws converges to a steady-state distribution in the sense that eventually, the realized equivalent of the objective function (2.1), $\sum_{i=1}^{N_t} \sum_{j=1}^{J} Y_{ijt}$, stabilizes and fluctuates only moderately around some level. This level generally shows relatively small variance across chains, at least when $N_t$ is fixed and constant. After removing a sufficient burn-in period, we can take all chains of draws and average the period values to calculate the expected steady-state value of using a particular strategy. We turn this into a fraction of the expected number of clicks at the theoretical optimum and refer to this as the quality of the ordering in the following.

### 2.2.2 Thompson Sampling

Many economic problems involve a repeated choice between options with uncertain payoff distributions; that is, not only is the payoff of a particular option a random variable, but the parameters of its distribution are themselves unknown. The economic agent faces an exploitation-exploration trade-off between optimizing for the current round based on what has already been learned about payoffs (exploitation) and maximizing what is learned for subsequent rounds by picking an option that, in expectation, is inferior to another but still has a non-zero probability of having the highest one-round payoff (exploration).

Typically, as in our setting, a Bayesian set-up in which the agent updates prior beliefs over the set of payoff parameters using observed payoffs is assumed. One approach to play in such

settings is randomized probability matching, the best known variant of which is Thompson Sampling (Thompson, 1933, Scott, 2010). Agents who use Thompson Sampling select between the available actions at random according to a simple heuristic: each option is played with the probability, as implied by posterior beliefs, that its payoff is the highest out of all options. This strategy has intuitive appeal: actions with greater potential are played with higher probability even if they are inferior in expectation, reflecting the option value that they can yield a high reward but need not be played further if they do not; thus, "variance" can beat "mean."

We next state the principle behind Thompson Sampling formally in very general terms and later applied to the product ranking problem. Let the set of possible arms be given by $\mathcal{A}$, and for now, assume that every arm $a \in \mathcal{A}$, if played, yields a random one-dimensional reward $r$. Based on an observed history of play and outcomes $h_t$, beliefs about the vector of payoff-relevant parameters $\mu$ are denoted by $p(\mu \mid h_t)$. For a fixed value of $\mu$ and an arm $a \in \mathcal{A}$, let the expectation of $r$ be denoted by $Q_\mu(a) = \mathbb{E}_{r|\mu;a}[r \mid \mu]$. Then, the probability $\tau(a \mid h_t)$ that the arm $a$ is sampled and played under Thompson Sampling is given by[15]

$$\begin{aligned} \tau(a \mid h_t) &= \Pr\left(Q_\mu(a) > Q_\mu(a') \, \forall a' \in \mathcal{A}, \, a' \neq a \mid h_t\right) \\ &= \int \mathbf{1}\left\{a = \arg\max_{a' \in \mathcal{A}} Q_\mu(a')\right\} p(\mu \mid h_t) \, d\mu. \end{aligned}$$

For the stochastic multi-armed bandit problem most familiar to economists, with geometric discounting and independent prior distributions for each arm, it is known that the optimal strategy is to deterministically play the arm with the highest Dynamic Allocation Index, or more commonly, Gittins index (Gittins and Jones, 1974, Gittins, 1979, Whittle, 1980). The Gittins index equals the value of the optimal stopping problem that maximizes the expected period-average payoff from the respective arm using the discount factors as weights. It can alternatively be interpreted as the fair retirement value if the agent can decide in each period whether to continue playing the arm or take the retirement payout. The indexation result is

[15] We can ignore ties by assuming that the joint belief distribution is absolutely continuous, which will be the case in our application.

significant because it allows the analyst to compute indices separately for each arm, which means that the complexity of the problem is linear rather than exponential in the number of arms.

In the product ranking problem at hand, however, to the best of our knowledge no such results on optimal play are available. We can view the product ranking problem as a multi-armed bandit problem in two different ways, depending on whether we think of an arm as a complete ordering of products or as an individual product. Under the former definition, the set of arms $\mathcal{A}$ is made up by all possible permutations of the product orderings $\rho$. Not only may this approach be computationally prohibitive for Gittins play due to the extremely large number of arms, but the joint prior distribution no longer takes a product form when the payoff of each arm is given by a function of the parameters of all products. When this independence assumption is violated, it is no longer possible to rely on indexation for optimal play. In some cases, the set of arms can be partitioned into independent clusters (Pandey *et al.*, 2007, Dickstein, 2014), but such a partitioning is not generally available in the product ranking problem. Linearly parametrized bandits, in which rewards follow a distribution indexed by the inner product of an unknown parameter vector and a coefficient vector that is specific to each arm, appear to be the closest in the literature to the product ranking problem if an arm is defined as a complete ordering; a number of algorithms have been proposed and studied for this setting, none of which are exactly optimal, however (see Rusmevichientong and Tsitsiklis, 2010, and references therein).

The important difference in the product ranking problem is that rewards are separately observed for each product, which may not matter from a static utility perspective but affects learning. This suggests that one might define an arm as an individual product and perhaps order products by their Gittins index. However, for the problem of selecting a fixed number of arms to play simultaneously (rather than just one at a time), Sundaram (2005) provides a counterexample showing that it is not generally optimal to simply select the arms with the

highest Gittins indices.[16,17]

Thompson Sampling is known to be used in practice at Google for directing traffic to different arms of an experiment (Scott, 2012) and at Microsoft for click-through rate prediction in sponsored search (Graepel *et al.*, 2010); in addition, it has at the very least been analyzed in papers by researchers at Facebook (Eckles and Kaptein, 2014), eBay (Hsieh *et al.*, 2015), Yahoo (Chapelle and Li, 2012), and LinkedIn (Tang *et al.*, 2013). Much of its appeal to practitioners stems from its intuitive nature, flexibility, and low computational requirements. In particular, in the classic multi-armed bandit setting where only one arm is played at a time, it only involves drawing from the posterior distribution of the reward parameters $p(\mu \mid h_t)$, computing the associated expected reward $Q_\mu(a)$ for every arm $a$, and selecting the arm with the highest expected reward under this set of draws. As a result, each arm is sampled according to the posterior probability that its expected payoff is the highest. Motivated by strong performance in simulations (Chapelle and Li, 2012), the theoretical properties of Thompson Sampling have recently received increased attention, and there are now results showing that Thompson Sampling achieves different types of lower bounds on statistical regret (Agrawal and Goyal, 2012, Kaufmann *et al.*, 2012, Agrawal and Goyal, 2013, Russo and Van Roy, 2014).

Furthermore, Thompson Sampling is known to work well with delayed feedback and batch (as opposed to online) updating. Even if updating does not occur in real time, but only at fixed (e.g., daily) intervals, so that the distribution of draws is fixed for one "batch" of observations, it can still explore the reward space relatively efficiently compared to the majority of deterministic approaches (e.g., Chapelle and Li, 2012). This is of great value to practitioners but also the researcher: it allows us to model competition dynamics over discrete

[16]Whittle (1988) shows that this "index policy" is asymptotically optimal where the number of arms grows while the fraction of arms to be played simultaneously remains constant.

[17]It would be interesting to compare the performance of an index policy that ranks products by their Gittins index to that obtained by Thompson Sampling, and in particular, to have a firm that plays an index policy compete with one that uses Thompson Sampling. The formulas provided in Gittins *et al.* (2011, Section 8.4) to approximate the index values may reduce computation time. Computing Gittins indices remains a challenge if play is to occur in batches (more below).

time periods and to interpret the size of one batch as the size of a firm's consumer base in one period.

Since it is explicitly Bayesian, Thompson Sampling can incorporate prior information about the (likely) distribution of payoff parameters, recognizing the uncertainty in payoff estimates based on few observations and "shrinking" quality estimates accordingly. At the same time, arms which have not been explored extensively have a large posterior variance, resulting in relatively many extreme draws for these arms, so that they end up being played quite frequently until sufficient evidence has accumulated to stop doing so. Relative to an approach that selects arms based on frequentist point estimates, this may lead to more balanced and intelligent exploration. There are a number of other heuristics for the multi-armed bandit problem, some of them similar in spirit, such as the well-known Upper Confidence Bound (UCB) approach, which ranks arms according to a value given for each of them by the sum of its mean and its standard deviation multiplied by a factor.[18]

### 2.2.3 Adapted Thompson Sampling

To use classical Thompson Sampling in the product ranking problem, we sort a set of draws from the posterior distribution of the product quality parameters to determine the ranking displayed to the next consumer. This procedure results in draws for the full ranking according to the probability that it has the highest payoff out of all possible permutations. It thus takes all possible orderings as the set of available arms and chooses as arm $a \in \mathcal{A}$ any possible ordering $\rho$. Formally,

$$a \in \mathcal{A} = \left\{ \rho : \ \rho = P\left(1, \ldots, J\right)' \text{ for some } P \in \{0,1\}^{J \times J} \text{ s.t. } \sum_i P_{ij} = 1 \text{ and } \sum_j P_{ij} = 1 \forall i, j \right\}.$$

To see why each ordering is drawn according to the probability that it is statically optimal, observe that since position effects are assumed to be monotone, the optimal ranking orders products by quality. By drawing parameters and then sorting, we draw from the probability distribution that qualities exactly follow the order in question. Since we can draw a different

[18] See Russo and Van Roy (2014) for the connection between Thompson Sampling and UCB algorithms.

ordering for every consumer $i = \{1, \ldots, N_t\}$, it is easy to see why Thompson Sampling is attractive for exploration when updating of beliefs occurs in batches as in our model.[19]

In this basic form, Thompson Sampling does not involve any tuning parameters that control the degree of the experimentation. We introduce a Thompson rate $\kappa \in [0, 1]$ that allows for this by taking draws from the distribution of $\kappa\mu + (1-\kappa)\,\mathbb{E}\,[]$ for $\mu \sim p\,(\cdot \mid h_t)$ as above. In other words, our draws are a linear combination of a draw from the posterior distribution of $\mu$ and its posterior mean. The weight $\kappa$ controls this combination and thereby the degree of experimentation. One attractive property of this approach is that it nests both classical Thompson Sampling ($\kappa = 1$) and Bayesian Myopic play ($\kappa = 0$) at its two extremes. This approach is technically no longer randomized probability matching because the probability that an arm is played is no longer necessarily increasing in its probability of being optimal, so that some of the theoretical performance guarantees do not apply. Below, we look at performance explicitly in our setting.

In our adapted Thompson Sampling method, we take draws from the posterior distribution of $\mu$ as before but form linear combinations of these draws and the posterior mean. Like earlier, sorting these (transformed) draws and using the resulting vector of indices as an ordering $\rho$ amounts to drawing from a distribution over orderings; the support of this distribution is the set $\mathcal{A}$ defined above. The probability this distribution places on an ordering $\rho \in \mathcal{A}$ is equal to the probability that $\rho$ is the ordering which places one product above another if and only if the true difference in their qualities exceeds the negative difference of their posterior mean qualities scaled up by a constant $\frac{\kappa}{1-\kappa}$ controlled by the firm. Thus, for a given posterior mean of product qualities, the firm's beliefs over product qualities induce a probability distribution over orderings; specifically, $\rho^{(\kappa)} : \mathbb{R}^J \to \mathcal{A}$ is a mapping from realizations of the posterior belief distribution into product orderings such that for $\kappa > 0$,

$$\rho^{(\kappa)}\,(\mu) \quad = \quad \left\{\rho :\, \rho\,(i) < \rho\,(j) \Leftrightarrow \mu_i - \mu_j > \frac{1-\kappa}{\kappa}\left(\mathbb{E}_\mu\left[\mu_j\right] - \mathbb{E}_\mu\left[\mu_i\right]\right)\right\},$$

[19] We thus implement Thompson Sampling at the individual consumer level, as is most common. This is different from Thompson Sampling at the batch level – e.g., same ordering for all consumers in a period, which would negate many of its benefits – since it cannot be optimal at the batch level under an absolutely continuous belief distribution to display different product orderings to different consumers.

which is a singleton with probability one for an absolutely continuous belief distribution as in our setting. By construction, the induced distribution over orderings is a proper probability distribution: $p(\cdot \mid h_t)$ is a proper probability distribution, and the mapping $\rho^{(\kappa)}$ is single-valued.

The economic interpretation is that for the firm to play a product higher than another to which it is inferior in expectation, it requires that there is a probability that its actual expected gain from this reordering is greater than some multiple of the associated expected loss under current firm beliefs. This multiple is a problem-dependent parameter chosen by the firm.

### 2.2.4 Empirical optimality of Bayesian Myopic play

In this section, we show that given the model estimated in section 2.1 and the problem and criterion defined in section 2.2.1, the optimal Thompson rate is actually consistently $\kappa = 0$, which corresponds to Bayesian Myopic play. We argue that there are meaningful settings in which our proposed adaptation to Thompson Sampling improves upon both classical Thompson Sampling and Bayesian Myopic play, but that given our parameters and some of the peculiarities of the product ordering problem, there generally is little benefit to experimentation purely for the sake of learning product qualities. While this also gives a preview on the relationship between firm size and the quality of product rankings, we delay a thorough discussion until section 2.3.

Figure 2.5 plots the expected fraction of the optimal number of clicks achieved in steady state as a function of the Thompson rate $\kappa$ for different firm sizes $N_t$. As expected, quality is increasing in firm size. More central to this section, quality is strictly decreasing in the Thompson rate for all firm sizes, and the drop is more pronounced for smaller-sized firms, where it can be quite significant. We thus conclude that Bayesian Myopic play, passive learning, is optimal or near-optimal for all relevant configurations in our setting. In the following sections, we are therefore comfortable assuming passive learning rather than optimizing explicitly in equilibrium.

This is perhaps a surprising result, and we therefore consider a slightly different scenario, both to provide reassurance that our adaptation of Thompson Sampling is not irrelevant as a benchmark and to analyze what peculiar features of the problem at hand render active learning less relevant. We consider a setting as before, with $N_t = 1,000$ and $J = 10,000$, but we limit the number of available positions by setting the elements of $\gamma$ to zero after a certain position. This can be likened to a setting of a physical store that has multiple branches, each with limited shelf space, and for each of these branches, the store needs to decide in every period which products to carry and which of them to place on what shelf. Figure 2.6 plots the steady-state performance achieved as a function of the Thompson rate for different levels of capacity, taking into account the reduced number of positions. The fewer positions available, the more important Thompson Sampling becomes, with a larger optimal rate $\kappa$, but if capacity is large, there is little need to, and gains over Bayesian Myopic play are at best marginal.

One big difference to most analyses of Thompson Sampling in multi-armed bandit problems lies in the correlation between arms in our setting. Unlike there, it is possible for the firm to learn about the quality of an ordering $\rho'$ from playing an ordering $\rho$ because of the separability assumptions in (2.2). This makes it less important to play statically suboptimal solutions to learn and tilts the balance away from exploration to exploitation. The correlation between arms is higher the "gentler" position effects are.

In another sense, the firm already learns more about the quality of all products than it would optimally like to within the confines of our model. Ideally, it would like to just clone its top products (given that our model does not include any substitution), but instead, it can only use each product once and therefore learns about the quality of less relevant products at a relative rate that depends on the vector of position effects $\gamma$, which is gentle or smooth enough for Bayesian Myopic play to be optimal unless we trim it off after some position. We speculate that another reason may be that the degree to which the product in a particular position affects firm quality is directly linked to the rate at which the firm learns about a product placed there, a near-universal feature when firms learn observationally from their

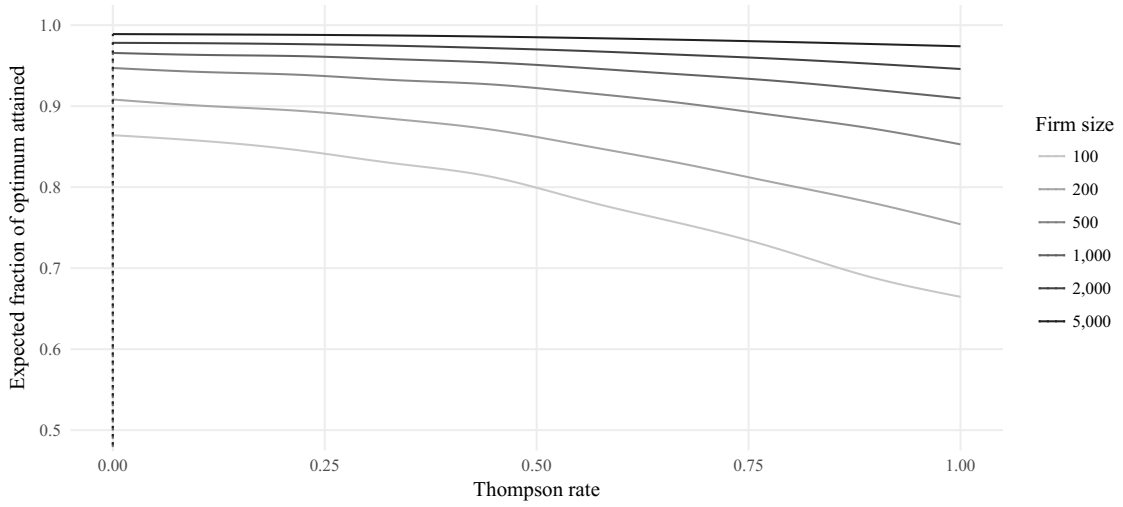


**Figure 2.5:** *Value of experimentation for different firm sizes*

running business.

We also emphasize again that firms in general may have other motivation to engage in active learning and experimentation, as our results only pertain to learning about product quality, where we also assume that position effects are known perfectly.

## 2.3 Rate of learning and returns to data

In economics, the term learning curve typically refers to the relationship between cumulative experience/production and quality, or its dual, cost. In machine learning, it is similarly used to describe the precision of an estimator or prediction algorithm as a function of sample size. In our setting, each chain of play converges over time to some steady-state distribution, and in that steady state, quality is stationary/stable, because effective cumulative experience does not increase over time as new data only offsets the "depreciation" in the existing stock. It therefore makes sense to draw two different learning curves. First, quality as a function of time as cumulative experience increases, eventually levelling off; we can do this for different firm sizes $N$ corresponding to the number of consumers per period. Second, mean steady-state quality as a function of firm size, thus varying cumulative experience in steady state.

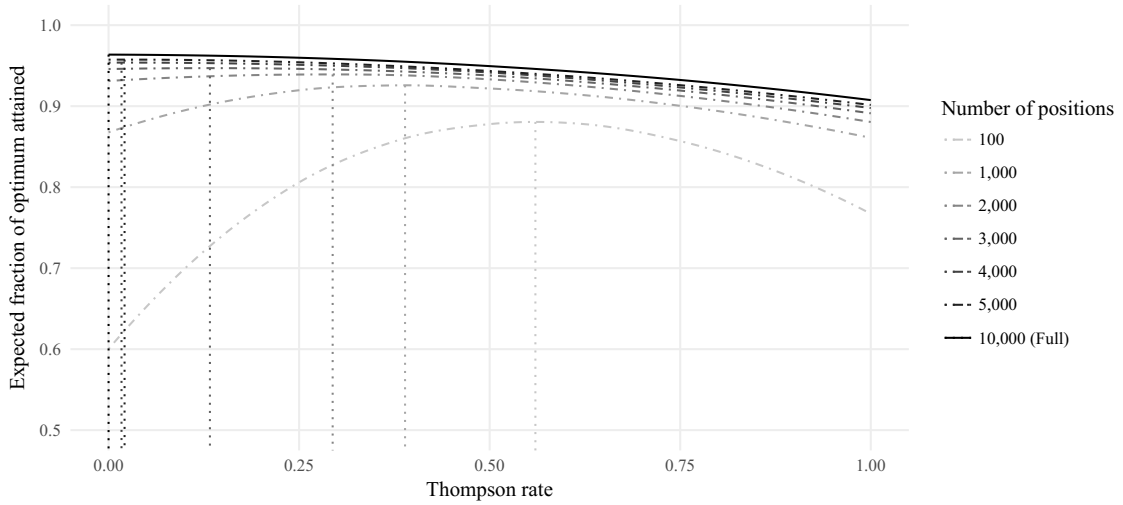


**Figure 2.6:** *Value of experimentation under capacity constraints*

As discussed in section 2.2.4, we assume that firms are Bayesian Myopic learners. In figure 2.7, we plot average paths for the fraction of the optimum number of clicks achieved over time by firms of different sizes, each starting out without any data. Unsurprisingly, we see that the quality of larger firms is strictly higher than that of smaller firms: learning occurs faster and converges to a higher steady-state level. Not only does each of the curves eventually plateau rather than increase indefinitely, but all of them appear to be concave everywhere. While this is obvious for the statistical problem, it is noteworthy that this property carries through to the economic problem.

The ordering $\rho$ played by the firm is a highly non-linear, discontinuous function of its beliefs on product quality $\mu$. While the firm's expected value (2.3) under its own beliefs, when evaluated at the implied ordering, is a continuous function of these beliefs, the true expected value is again discontinuous. However, the quality product distribution is unimodal and smooth, as are the position effects, which may lead to a relatively smooth learning curve. Figure 2.8 plots three sample learning paths for firm size $1,000$ showing that the patterns observed for the average are present at the individual path level as well.

While we have little to compare to, learning appears to happen relatively fast. A firm of

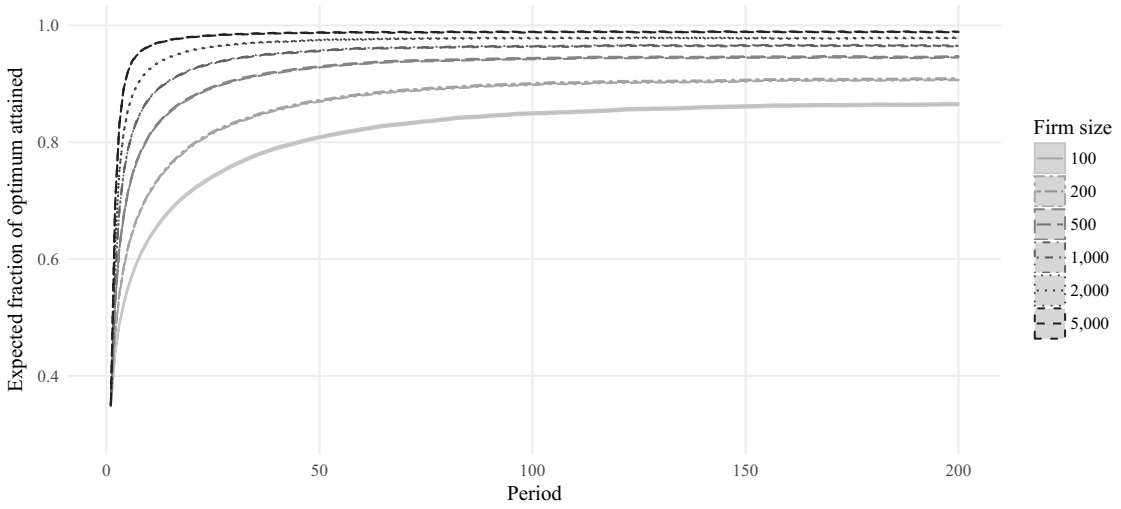


**Figure 2.7:** *Average learning curve over time*

size $N = 1,000$ attains 90.6% of its eventual level in period 10, 95.7% in period 20, and 99.0% in period 50; recall that we interpret periods as days. These numbers are perhaps slightly unrealistic, because they assume that the firm knows all other aspects of the environment, including position effects, from the outset. While it seems plausible that in a steady state, it will know these effects, and that they will be relatively stable over time, it is likely that the firm would initially have to learn about the position effects along with the product qualities.

As described above, experience in this setting is not simply the product of time and size: the same level that a firm of size $1,000$ has in period 100 is attained by firms of size $2,000$ and $5,000$ in periods 25 and 10, respectively. In contrast, the smaller firms in the figure never achieve that level. Larger firms learn at a disproportionally faster rate due to depreciation in firms' information. On the other hand, the fact that we do not allow for intraperiod optimization but have the firms update in daily batches affects this comparison in the opposite direction.

Figure 2.9 shows the average steady-state quality (expected number of clicks) attained as a function of firm size. Again, we see an increasing concave relationship. For small firm sizes, the rate of growth is very high, but eventually, since our quality metric cannot rise above

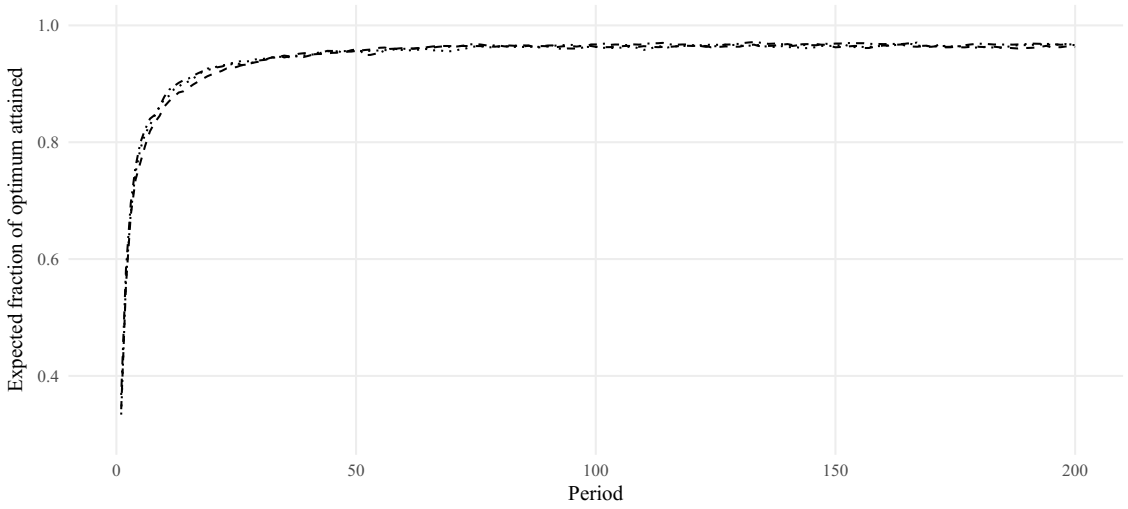


**Figure 2.8:** *Sample learning curves over time*

100%, it becomes very slow. In steady state, we observe that a firm of size 100 attains 86.3% of the optimum quality, 90.7% and 94.6% for sizes 200 and 500, respectively, 96.5% for our focal firm size of 1,000, and 97.8% and 98.9% for firm sizes 2,000 and 5,000, respectively. The rate at which the statistical precision increases thus becomes smaller, but we continue to see moderate gains even at relatively large firm sizes.

Figure 2.10 plots the beliefs on product qualities of two firms of size 1,000 facing an identical environment, with the same product qualities and timing of product entry and exit; each point represents the posterior mean of one product, and beliefs are taken from the same period after the chain converges for both firms. What differs between the two firms is the realizations of consumer demand they observe and the orderings they choose as a result. If both firms knew all product qualities with perfect precision, all points would line up on the 45° line. We see a relatively close agreement for the top products, with only minor differences in ordering. As we go down the belief distribution, the dispersion off the diagonal increases, indicating lower precision, though at least in the mid range of quality, beliefs still generally cluster around it. There are a few cases in the mid range of quality in which one of the firms has picked up on a product's potential while the other has not. This stems from the

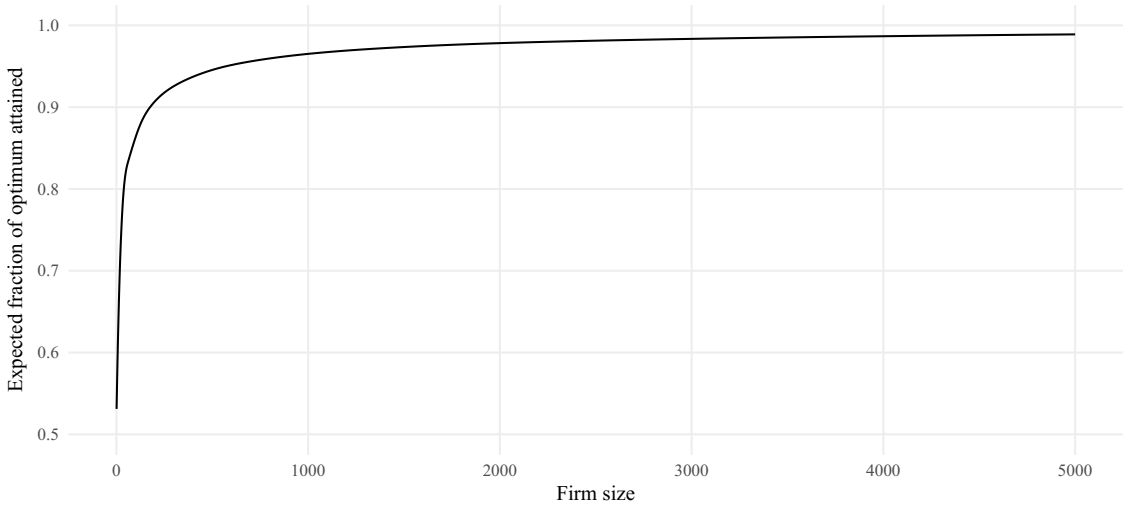


**Figure 2.9:** *Average steady-state quality against firm size*

endogenous ordering of products, where in some cases, the stochastic consumer click process determines whether the product receives sufficient exposure for the firm to learn its true quality. At the lower end of product quality (or popularity), precision is generally lower but of course also not as valuable.

We can quantify the size of the learning externality and information value of an additional consumer by comparing it to her direct contribution to the firm's objective in the form of her (expected) clicks. Denoting the steady-state expected clicks of a consumer as a function of firm size by $\pi(N)$, the direct contribution of the additional consumer is exactly $\pi(N)$, while the indirect contribution, through providing additional data to improve the product ordering displayed to all consumers, is equal to $\pi'(N) \cdot N$. The ratio between the indirect and the direct effect, $\frac{\pi'(N) \cdot N}{\pi(N)}$, can be recognized as the elasticity of expected clicks per consumer with respect to the number of consumers, or, in our terminology, the quality-firm size elasticity.

Figure 2.11 plots this elasticity as a function of firm size. The elasticity drops as firm size increases, with values of 0.080 for size $N = 100$, 0.057 for $N = 200$, 0.037 for $N = 500$, 0.024 for our focal firm size of $N = 1,000$, 0.015 for $N = 2,000$, and 0.009 for $N = 5,000$. Notably, the elasticity does not diminish to zero for the range of firm sizes we simulate and is in fact

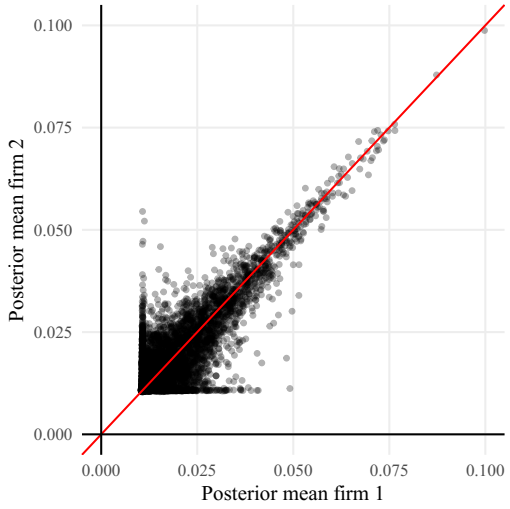


**Figure 2.10:** *Example beliefs of two independent firms in identical environment*

relatively stable past around firm size 3,000, only decreasing very gently thereafter. One of the main take-aways here is that even though the statistical return to additional data, as reflected in the quality of the ordering shown in figure 2.9, may eventually become relatively small, the economic return can remain significant, as the data of the marginal consumer is applied to improve the product ordering for a larger and larger base of inframarginal consumers. At our focal size $N = 1,000$, for every additional click provided by an additional consumer, her contribution to the quality of the product ordering displayed to all consumers yields an additional 0.024 clicks.

It is interesting to compare these results to some existing estimates. According to the FTC staff report on Google (Wall Street Journal, 2015), "Sergey Brin [of Google] testified that a 'rough rule of thumb' might be, as query volume doubles, a search engine might expect to see a one percent increase in quality."[20] While no formal definition of quality is given, it is likely that a metric very similar to ours measuring consumer response is implied. In our

[20]"Preston McAfee, Yahoo's former chief economist, suggested that 'having 2-3 times as many user observations,' particularly for 'tail' queries, would result in substantially more than a one percent increase in quality – indeed, doubling a search engine's queries would be 'an enormous advantage.' McAfee suggested that a 3-to-1 advantage in query volume could result in a 70 percent increase in 'precision' for the search engine's ability to answer unique queries."

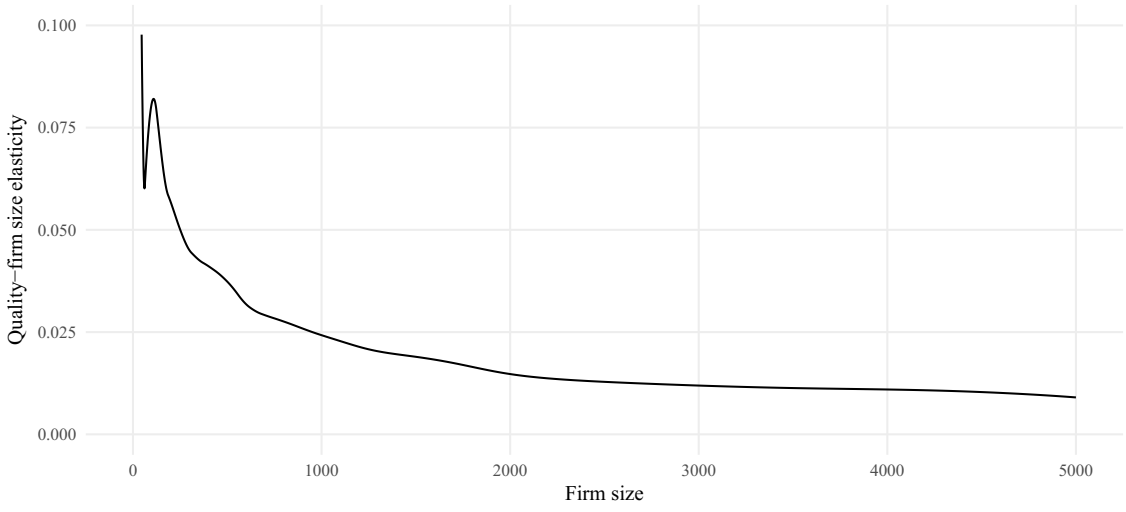


**Figure 2.11:** *Elasticity of quality with respect to firm size*

simulation, doubling the data from a firm size of 1,554 to twice that number results in a 1% in quality, while the elasticity implied by Brin's rule of thumb of 0.014 is observed at 2,060. Both of these values are firmly within the range of firm sizes we analyze.

We can also ask how endogenous quality affects prices and mark-ups. Consider a simple model in which the firm sets a price $p$ and has $N(p)$ consumers, each making an expected number of clicks $\pi(N(p))$ that depends on the number of consumers. For each click, the firm earns an expected profit of (or proportional to) $p - c$, leading to an objective function of $N(p)\pi(N(p))(p-c)$. Assuming $N(p)$ and $\pi(N(p))$ have a (locally) constant elasticity of $\varepsilon_N$ and $\varepsilon_\pi$ with respect to price and size, respectively, the combined price elasticity of $N(p)\pi(N(p))$ is $\varepsilon_N(1+\varepsilon_\pi)$, leading to an optimal price of $p^* = \frac{\varepsilon_N(1+\varepsilon_\pi)}{\varepsilon_N(1+\varepsilon_\pi)+1}c$. The reduction in mark-up relative to a setting in which quality is exogenous and constant is given by $\frac{\varepsilon_N\varepsilon_\pi}{\varepsilon_N(1+\varepsilon_\pi)+1}$. As an example, assuming $\varepsilon_\pi = 0.025$ and a price elasticity $\varepsilon_N$ of $-3$, the reduction is 3.6%.

## 2.4 Minimum viable size of entrants

In this section, we introduce a model of competition to determine conditions under which data prevent entrants from gaining a foothold in the market and to understand the types of competitive configurations that are possible in the steady state of such a market. We use the estimated learning technology to map out the minimum viable size for entrants to compete in a market in which consumers choose firms based on quality and firms learn about quality through revealed consumer preference.

Suppose a market is contested by two firms, one of them, $m = 1$, a large incumbent and the other, $m = 2$, a smaller entrant. We are interested in conditions that allow the entrant to establish itself in the market rather than being forced out because it cannot compete on quality. Assume that in each period, there are $N$ consumers, and that 10% of them actively choose which of the two firms to patronize, while the remaining 90% stay with the firm they patronized in the previous period. For each consumer $i$ actively choosing, the utility of firm $m$ in period $t$ is given by

$$u_{itm} \quad = \quad \alpha_{0,m} + \alpha_1 \log(q_{tm}) + \epsilon_{im}, \tag{2.6}$$

where $q_{tm}$ is the current quality of firm $m$ and $\epsilon_{im} \overset{iid}{\sim} EV(1)$. We assume that the two firms have symmetric utility, so that $\alpha_{0,1} = \alpha_{0,2}$.

The interesting feature of this model is that quality $q_{tm}$ is the outcome of what the firm learned using its data up to and including period $t-1$. We assume that both firms learn using the technology described and estimated in sections 2.1 and 2.2.

We assume that firm 2, the entrant, makes some initial investment into advertising and marketing such that it enters the market in period 1 with a share $\mathfrak{s}_{2,1}$. It also makes a proportional initial investment into consumer research that allows it to enter with quality beliefs drawn from the steady-state distribution of a firm of constant size $\mathfrak{s}_{2,1} \cdot N$. The incumbent, firm 1, owns the rest of the initial market with steady-state beliefs from proportional data that it acquired by being in the market for a sufficiently long period of time.

We are interested in how large $\mathfrak{s}_{2,1}$, the entrant's initial share, needs to be for it to sustain

its market presence in the long run. In addition, we want to know what kind of steady state the market will eventually evolve to. Both an outcome in which firm 2 is pushed out of the market because it cannot increase quality sufficiently fast[21] and a symmetric one in which firms 1 and 2 are equals in terms of quality and market share are obvious steady states. Is there also an asymmetric steady state in which both firms are active in the market?[22]

We can use the estimates of section 2.3 for guidance on these two questions. While quality is not a deterministic function of size, it can nonetheless shed light on the possible dynamics in this market, and we can test the accuracy of its predictions using simulation from the exact model presented. For a given market share of firm 2 and the resulting estimated average qualities of the two firms, figure 2.12 plots the market shares implied by (2.6) for a market size of $N = 1,000$ and a quality sensitivity parameter of $\alpha_1 = 20$.[23] As the utility structure is symmetric, we only plot the range of 0 to 50% market share. There are three types of steady states: one in which firm 2 has essentially zero share, another in which it has a small share (9.7%), and finally the symmetric outcome in which both firms split the market evenly. Importantly, only the extreme outcomes in which only one firm is present in the market or both have identical shares are stable points of the dynamic system, while the intermediate outcome is not. The figure illustrates that starting either just below or just above this value causes the system to diverge from this steady state and instead move towards towards one of the two stable outcomes.

The interior fixed point is thus more appropriately interpreted as a threshold above which an entrant's initial share likely allows it to be competitive and eventually catch up with the incumbent, while for initial values below this threshold, the entrant is likely to eventually lose its footing in the market because asymmetries in quality are too large and exacerbate over time. We confirm this prediction by taking random draws from a market with the structure

[21] This requires that a sort not based on data is sufficiently bad relative to the quality sensitivity parameter $\alpha_1$ that a firm without any data gets (close to) zero market share.

[22] One of the main contributions of Besanko *et al.* (2010), whose model of learning by doing also features forgetting, is that it can explain such asymmetric equilibria.

[23] This corresponds to an elasticity of 10 at a symmetric market share division.

described above for varying levels of the entrant's initial share, represented by the shade of their line in figure 2.13. The dashed red line represents the threshold identified by the steady-state analysis, and we can see that firms entering below this level eventually vanish, while those with a larger initial share converge to the symmetric steady state over time. There is one exception: a firm narrowly below the predicted threshold eventually claims half of the market after hovering around the knife-edge steady state at the threshold for over 100 periods.

The two main comparative statics of interest for the threshold are with respect to the market size $N$ and the consumer quality sensitivity $\alpha_1$. Figure 2.14 plots the market share threshold for entry to be viable as a function of the quality sensitivity parameter $\alpha_1$ for three levels of market size, 1,000, 2,000, and 5,000. The threshold is increasing in the sensitivity parameter: for low values, entry is always feasible, in an intermediate range, it depends on the initial market share, and if consumers are highly sensitive to firm quality, successful market entry is impossible in all but the largest markets (unless the entrant comes in with an equal share of the market right away). In general, larger markets support entry at a lower threshold all else equal.

Perhaps surprisingly, the threshold is smaller not only as a market share but also as an absolute number: at $\alpha_1 = 20$, it is 98 for a market with 1,000 consumers, 42 for 2,000, and 29 for 5,000. This result is most easily understood when thinking of the dynamics underlying figure 2.12. Consider doubling the size of the overall market while keeping the size of the entrant constant. The quality of the entrant clearly stays constant, and provided the incumbent is on a relatively flat part of the learning curve, the quality differential between entrant and incumbent increases only very moderately. Then, as the overall market is now twice as large, there is almost twice as much demand for the entrant. Thus, starting at the same absolute firm size as at the threshold of the smaller market must now result in a vertical jump from red to black line, which is only possible if the threshold is lower (in absolute terms) for the larger market. Of course, in all of this, we assume that the set of available products does not change with the size of the market – if it did, then in our thought experiment, the incumbent might get moved into a steeper portion of the learning curve again, and this could

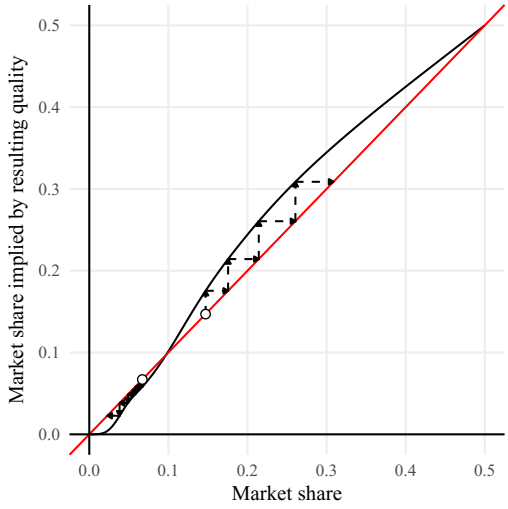


**Figure 2.12:** *Competitive learning dynamics*

undo the result.

## Conclusion

Using actual firm data from a large online retailer, this paper has presented and estimated a model to address two questions important for business and regulators: what are the returns to data scale, and when can it act as a strong barrier to entry? We conclude that there are moderate returns to data for large firms and that even as statistical returns from additional data vanish as the available stock grows, economic returns can remain significant, as the data from an incremental customer can be used to improve the quality served to a larger and larger existing base of customers. We use the model of learning and quality as a function of the customer base to map out the minimum viable size for entrants to compete in a market in which consumers choose firms based on quality and firms learn about quality through revealed consumer preference. Clearly, more research is needed for estimates of consumers' quality sensitivity, a central parameter; this is likely not to be easy.

The firm sizes in our simulation have been set up to allow comparison to the firm in our data. One limitation of our simulations is that they assume that the firm's data are

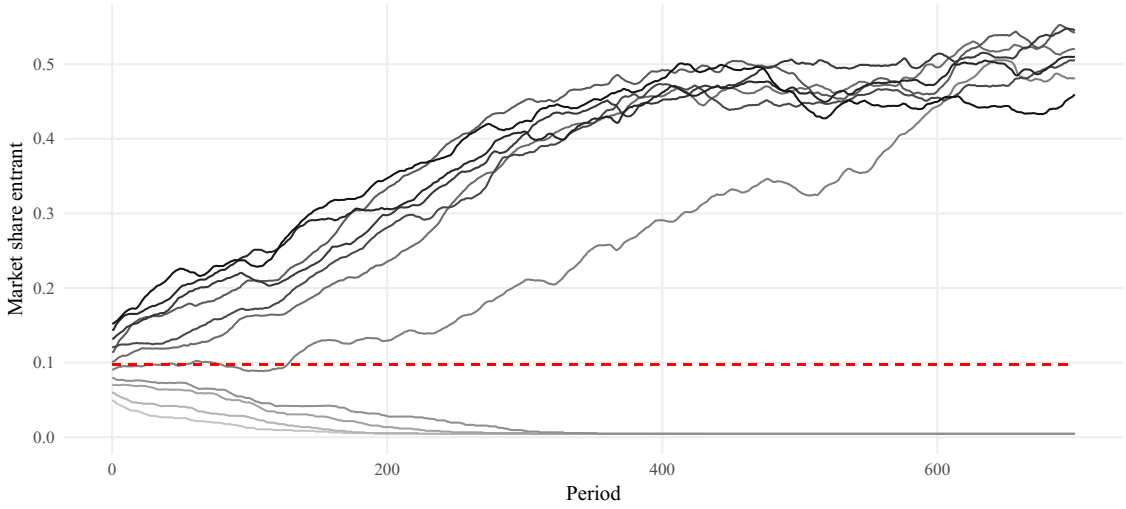


**Figure 2.13:** *Market share evolution for entrants with different initial share*

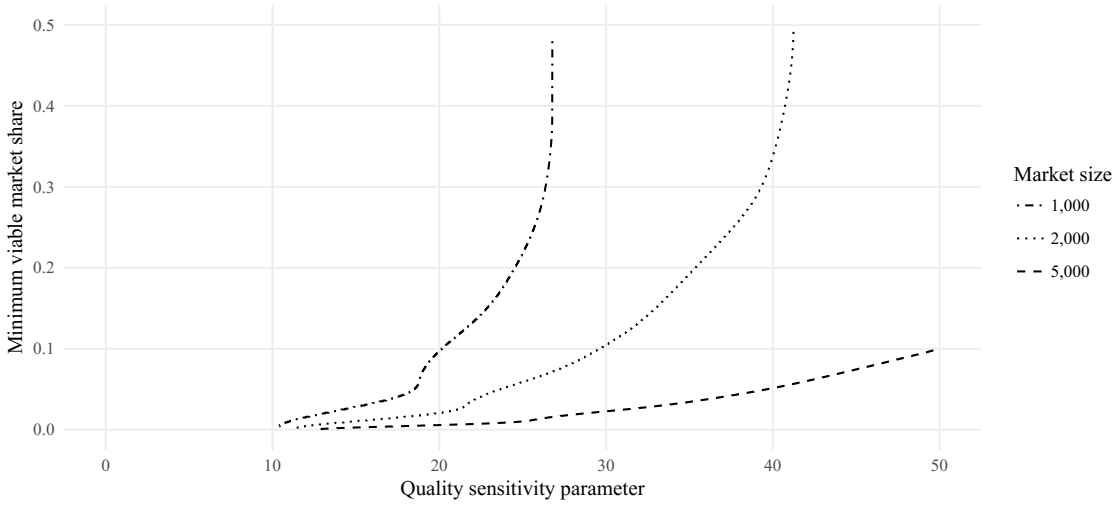


**Figure 2.14:** *Minimum market share required to sustain position*

independent draws from the distribution we have estimated. In practice, there are important factors that could lead to correlation and reduce the effective sample size (for a given firm size) and rate of learning; these include correlated shocks to advertising and promotion as well as model misspecification.

# Chapter 3

# Identification and Estimation with an Interval-censored Regressor when its Marginal Distribution is Known

## Introduction

We are interested in a regression function $\mathbb{E}\left(Y_i|V_i = \cdot\right)$, but the regressor $V_i$ is latent; instead, in addition to an outcome $Y_i$, we observe an interval $I_k$ in which $V_i$ is contained. Furthermore, we know the marginal distribution of $V_i$, possibly from another dataset. We ask the question how knowing this distribution informs bounds on the regression function or any linear functional of it.

Empirical researchers indeed often have access to only interval-censored micro data, but may know the marginal distribution. In the Health and Retirement Study, for instance, individuals unwilling to report their exact level of wealth are given the option to choose from a number of brackets. At the same time, the researcher may have access to the marginal distribution of wealth from, for instance, the Survey of Consumer Finance. Another typical

interval-censored variable is income. In the Current Population Survey, respondents' total annual household income is recorded in bins of varying width, the Consumer Expenditure Survey allows households to provide annual household income as contained in one of 13 different-sized bins, and the German microcensus reports personal/household monthly net income in 24 different-sized bins.

In the above examples, the precise value of $V_i$ was not recorded for at least a subset of the interview participants. Another cause of interval-censored variables are privacy concerns due to which data cannot be released in a version that matches outcomes with exact values of covariates. In many location problems, for instance, the address of a household is not completely reported in the data, but rather just the zip code, and sometimes only its first three or four digits. This means that the distance to a hospital or store is only known to belong to an interval. Access to records providing the exact population density within the zip code region allows the researcher to calculate the exact distribution of distance among households.[1] It is indeed common for applied research to rely on multiple sources of data, particularly in industrial organization, but also in labor economics.

On the other hand, in the case of customer, administrative, and other sensitive data, the data provider often corrupts the data for privacy reasons, but could potentially reveal the sample distribution of the interval-censored regressor. Administrative data in particular is often available in an easily accessible public use version with some variables given only in intervals, and in a restricted-access version with the precise values of all variables. The public version of the UK Labour Force Survey for instance reports age in buckets, while the restricted version provides exact birth date.

Our empirical application considers estimation of age-earnings profiles given earnings and interval-censored age based on the Current Population Survey (CPS). It informs two questions of practical interest: how much does the researcher gain from obtaining (an estimate of) the marginal distribution of an interval-censored variable? If the data was censored after

[1]For the United States, population density at the census block level is available at `http://server.arcgisonline.com/ArcGIS/rest/services/Demographics/USA_Population_Density/MapServer`.

collection for confidentiality reasons, how useful would it be to researchers if data providers shared its sample distribution? In our application, gains appear substantial, in particular if additional restrictions are placed on the shape of the regression function.

Interval-censored data is one of the canonical examples in the recent but now extensive literature on partial identification. For interval-censored regressors, Manski and Tamer (2002) derive sharp bounds on the regression function assuming the latter is monotone. This paper shows how to make use of information on the marginal distribution of the regressor to derive tighter bounds. In order to derive the sharp identified set, we apply a result from the ecological inference problem studied in Cross and Manski (2002). For binary choice models, Magnac and Maurin (2008) show that marginal information can be very useful and potentially yield point identification. Our inference procedure draws on a Bayesian approach to partially identified models with valid frequentist interpretation developed in Kline and Tamer (2016).

We assume that the regressor either follows a discrete distribution or that its distribution can be approximated in such way. This allows us to establish the sharp identified set based on the results in Cross and Manski. We give precise conditions under which a linear programming analogue estimator is consistent for the true identified set. To perform valid Bayesian and frequentist inference, we establish the convergence of an estimable data statistic and the posterior distribution. While Kline and Tamer require convergence in total variation of the data statistic, which is not satisfied here, we provide a result by which the weaker convergence in distribution can be sufficient, extending their results and possibly allowing for additional applications. In cases in which the data provider censors a variable after collection but provides its marginal distribution, so that the latter is estimated on the same sample as the outcome distribution, exact inference is no longer possible. Instead, we provide a conservative approximation that appears to perform well in our application.

The organization of the paper is as follows. Section 3.1 specifies the problem and introduces the notation used in the remainder of the paper. Section 3.2 presents the sharp identification region and discusses shape constraints on the parameter space. Section 3.3.1 shows how to formulate the problem in terms of equality and inequality constraints to set up a linear

programming estimator, section 3.3.2 establishes its consistency, and section 3.3.3 shows how to perform inference. Section 3.4 illustrates these methods with an age-earnings profile application, and finally, we conclude. Proofs are collected in section B.1 of the Appendix.

## 3.1 Set-up and notation

Suppose that for a sample $i = 1, \ldots, N_1$, we observe an outcome $Y_i \in \text{Supp}(Y) \subset \mathbb{R}$. Instead of the exact value of a latent discrete scalar regressor $V_i \in \text{Supp}(V) = \{\gamma_l\}_{l=1}^{L} \subset \mathbb{R}$ (where $L$ is finite), we observe what we will refer to as "interval" indicators, $D_{k,i} = \mathbf{1}\{V_i \in I_k\}$ for sets $I_k$, $k = 1, \ldots, K$. These latter sets are chosen to be ascending, $\min I_{k'} > \max I_k$ for $k' > k$, and exhaustive, $\text{Supp}(V) = \cup_{k=1}^{K} I_k$. We assume that $V_i \overset{iid}{\sim} \pi > 0$, where $\pi$ is an $L$-vector with $\sum_{l=1}^{L} \pi_l = 1$. Defining $\delta(\cdot)$ as an index function which, for $\gamma_l \in I_k$, returns as $\delta(l) = \mid \{l' : \gamma_{l'} \in I_k, \gamma_{l'} \leq \gamma_l\} \mid$ its position in $I_k$, we can write $V_i \mid V_i \in I_k \overset{iid}{\sim} \bar{\pi}^{(k)} > 0$ for a normalized $\mid I_k \mid$-vector $\bar{\pi}^{(k)}$ with $\sum_{l:\gamma_l \in I_k} \bar{\pi}^{(k)}_{\delta(l)} = 1$. We let $F_{Y|V}$ denote the distribution of $Y_i$ conditional on $V_i$.

Our application is to an age-earnings profile, with wage earnings as outcome $Y_i$, and age in integer years as regressor $V_i$. However, age is only observed in five-year "intervals" such as $\{20, 21, \ldots, 24\}$, $\{25, 26, \ldots, 29\}$, etc. Since the regressor $V_i$ is discrete, it more accurately takes values in sets $\{I_k\}_{k=1}^{K}$ as described above, but we refer to these as "intervals" to underscore the similarity to the typical setting in the literature, in which a continuously distributed regressor is only observed as belonging to some interval. While there are many cases in which the true regressor may indeed be of a discrete nature, this restriction would unduly limit the scope of application, since almost all continuous distributions can be approximated by a discrete distribution (Chamberlain, 1987, section 3).[2]

What distinguishes our setting from those in the prior literature is information on the distribution of the latent regressor. We will alternatively treat $\pi$ as known, as estimated from the same sample $i = 1, \ldots, N_1$, or as estimated from a second independent sample

[2]Although non-constructive, Lemma 3 in Chamberlain (1987) shows that the approximating multinomial distribution requires only finite support, as is assumed here for $\text{Supp}(V_i)$.

drawn from the same population, $j = 1, \ldots, N_2$, for which we observe only $V_j \overset{iid}{\sim} \pi$. These alternatives reflect the different ways in which the particular structure of the data can arise. In the following, we will omit indices whenever this does not lead to ambiguity. We will use bold-face font for a data vector, and define $\mathbf{X} = (\mathbf{Y}, \mathbf{D}, \mathbf{V})$ as the data matrix that contains all realizations of $Y$ with corresponding interval indicators $\{D_k\}_{k=1}^K$ as well as the realizations of $V$ based on the same sample or an auxiliary sample.

We are interested in linear functionals of the conditional expectation function $\mathbb{E}[Y \mid V = \cdot]$. Let the parameter space be given by $\Theta$ and assume it is convex. Since $\text{Supp}(V) = \{\gamma_l\}_{l=1}^L$, we can represent this function simply as a vector $r \in \Theta \subseteq \mathbb{R}^L$, where $r_l = \mathbb{E}[Y \mid V = \gamma_l]$, with $r^{(k)}$ the subvector of $r$ corresponding to the support points in $I_k$. Any linear functional $L : \mathbb{R}^L \to \mathbb{R}$ can then be written as $L(r) = c'r$ for an appropriate choice of $c \in \mathbb{R}^L$. Examples of such functionals include the expected earnings at a given age, or wage growth, which is simply the difference in earnings for two age groups. We will refer to the identified set for the conditional regression function vector as $\Theta_r^I \ni r$. Extending the definition of $L$ to subsets $A \subseteq \mathbb{R}^L$ such that $L(A) = \{c'r : r \in A\}$, the identification region for a linear functional of the conditional regression function is given by $L\left(\Theta_r^I\right)$. In addition, we write $\underline{L}(A) = \min_{r \in A} c'r$ and $\bar{L}(A) = \max_{r \in A} c'r$ .

The closure, interior, and boundary of a set $A \subseteq \mathbb{R}^L$ are denoted by $\text{cl}(A)$, $\text{int}(A)$, and $\text{bnd}(A) = \text{cl}(A) \setminus \text{int}(A)$, respectively. We write $\text{conv}(A)$ for the convex hull of a set $A$, $K_{kc}\left(\mathbb{R}^d\right)$ for the space of compact convex subsets in $\mathbb{R}^d$, and $\mathcal{S}^{d-1}$ for the unit sphere in $\mathbb{R}^d$, given by $\left\{x \in \mathbb{R}^d : \|x\| = 1\right\}$ for the Euclidean norm $\|\cdot\|$. For points $a$ and $b$, and sets $A$ and $B$, all in $\mathbb{R}^d$, we define the point-to-set distance $d(a, B) \equiv \inf_{b \in B} \|a - b\|$. The Hausdorff distance $H(A, B)$ of two sets $A$ and $B$ is given by $H(A, B) \equiv \max\{\sup_{a \in A} d(a, B), \sup_{b \in B} d(b, A)\}$. Convergence of a set estimator $\hat{A}$ to a set $A$, defined as $H\left(\hat{A}, A\right) \xrightarrow{p} 0$, is denoted by $\hat{A} \xrightarrow{H} A$. Convergence in total variation of a sequence of random variables $U_n$ to a random variable $U$ is denoted by $U_n \xrightarrow{tv} U$. For a Borel set $A$ and a distribution $P$, we write $\Pr_P(A) \equiv \Pr(U \in A)$, where $U \sim P$. We define the quantile function $Q_U(\tau) \equiv \inf\{r : F_U(r) \geq \tau\}$ for any random variable $U$ with distribution function $F_U(\cdot)$.

## 3.2 Identification

### 3.2.1 General result

The marginal distribution of the regressor, $\pi$, can be used to derive bounds which are tighter than those obtained without this information, with the latter briefly discussed in the next subsection. Before we can present the main identification result, which follows directly from a result in Cross and Manski (2002), we need to introduce the concept of stacked distributions. For any $\tau_1$, $\tau_2 \in [0,1]$ with $\tau_2 > \tau_1$, and $k = 1, \ldots, K$, define $\bar{F}^{\tau_1:\tau_2}_{Y|V\in I_k} \equiv \min\left(\max\left(\frac{1}{\tau_2-\tau_1}F_{Y|V\in I_k} - \tau_1, 0\right), 1\right)$ to be the conditional distribution that results from removal of the lowest $\tau_1$ and highest $1-\tau_2$ fractions from the conditional distribution of $Y$ given $V \in I_k$.[3] The support points of $V$ in $I_k$ can be ordered according to $\mid I_k \mid!$ different permutations. To this end, define the functions $\phi^m_k(\cdot)$, $k = 1, \ldots, K$, $m = 1, \ldots, \mid I_k \mid!$, which take as argument an element of the set $I_k$ and return its position under permutation $m$. For any support point $\gamma_l$ in $I_k$, define the quantiles $\tau_1^{k,\delta(l),m} \equiv \sum_{l':\gamma_{l'}\in I_k,\, \phi^m_k(\gamma_{l'})<\phi^m_k(\gamma_l)} \bar{\pi}^{(k)}_{\delta(l')}$ and $\tau_2^{k,\delta(l),m} \equiv \tau_1^{k,\delta(l),m} + \bar{\pi}^{(k)}_{\delta(l)}$, such that $\tau_1^{k,\delta(l),m}$ gives the mass placed by $\bar{\pi}(k)$ on support points appearing before $\gamma_l$ under permutation $m$, and $\tau_2^{k,\delta(l),m}$ gives the mass before and including $\gamma_l$. For any permutation $m$, a stacked distribution is now given by the $\mid I_k \mid$-vector of distributions that has $\bar{F}^{\tau_1^{k,\delta(l),m}:\tau_2^{k,\delta(l),m}}_{Y|V\in I_k}$ as its $\delta(l)$-th element.

Each such stacked distribution, which we will refer to as $P^m_k$, is a candidate set of distributions $\left\{F_{Y|V=\gamma_l}\right\}_{l:\,\gamma_l\in I_k}$ that is consistent with the identified distribution $F_{Y|V\in I_k}$. Cross and Manski choose the term "stacked distribution" because for a given permutation, the outcome distribution conditional on the support point in position $j$ lies weakly to the left of that corresponding to position $j+1$.[4]

**Proposition 1.** *Assume that $\pi$ and $F_{Y|V\in I_k}$, $k = 1, \ldots, K$, are identified from the distribution*

[3] We avoid the term "truncation." Under truncation, the removal of density occurs at points of the underlying support, whereas here, it is based on values of the distribution function. If $Y_i \mid V_i \in I_k$ has no mass points, then we can say that $\bar{F}^{\tau_1:\tau_2}_{Y|V\in I_k}$ is based on left-truncation at the $\tau_1$-quantile of $Y_i \mid V_i \in I_k$ and right-truncation at its $\tau_2$-quantile. The identification results, however, allow for point mass.

[4] Figure 1 of their paper provides a very helpful example of a set of stacked distributions based on three support points, and further discussion can be found in Ridder and Moffitt (2007, pp. 5486-91).

*of the observable data. For every* $k = 1, \ldots, K$, *let* $E_k^m \equiv \int_{\mathbb{R}^{|I_k|}} y dP_k^m(y)$ *for all permutations* $m = 1, \ldots, |I_k|!$ *of stacked distributions. Then,* $r \in \Theta_r^I = \tilde{\Theta}_r^I \cap \Theta$, *where*

$$\tilde{\Theta}_r^I = \times_{k=1}^K \text{conv}\{E_k^m, m = 1, \ldots, |I_k|!\}.$$

*These bounds are sharp.*

*Proof.* Adapt arguments in Cross and Manski (2002) and Molinari and Peski (2006).[5] □

*Note.* Geometrically, $\tilde{\Theta}_r^I$, the identified set in the absence of any restrictions on the parameter space, takes the shape of a convex polytope. Since $\Theta$ is assumed to be convex, their intersection $\Theta_r^I$ is also convex. All the interval extreme points $\{E_k^m\}_{m=1}^{|I_k|}$ lie in a hyperplane $H_k$ given by $\left\{r^{(k)} : \sum_{l:\gamma_l \in I_k} \bar{\pi}_{\delta(l)}^{(k)} r_{\delta(l)}^{(k)} = \mathbb{E}[Y \mid V \in I_k]\right\}$ by the law of total expectation. If only $\mathbb{E}[Y \mid V \in I_k]$ but not $F_{Y|V \in I_k}$ is identified from the data (possibly because only aggregate statistics on the outcome are available), the sharp identified set is given by $\times_{k=1}^K H_k$.

The intuition is best explained focusing on just one interval $I_k$. Feasible conditional outcome distributions $\left\{F_{Y|V=\gamma_l}\right\}_{l:\gamma_l \in I_k}$ (a $|I_k|$-vector of distribution functions) solve a finite mixture problem derived from the law of total probability:

$$F_{Y|V \in I_k} = \sum_{l:\gamma_l \in I_k} \bar{\pi}_{\delta(l)}^{(k)} F_{Y|V=\gamma_l}$$

All feasible conditional expectation functions are obtained from integration with respect to feasible conditional outcome distributions. Since the set of such distributions is convex, and because the expectation operator is an affine mapping and therefore convexity-preserving, the set of feasible conditional expectation functions is also convex. In addition, it is well known that any closed bounded convex set is the convex hull of its extreme points (Rockafellar, 1970, Cor. 18.5.1). Cross and Manski show that these extreme points are obtained by integration with respect to permutations of stacked distributions, resulting in points $\{E_k^m\}_{m=1}^{|I_k|}$. Molinari and Peski show that the set of feasible expectation functions is indeed closed, so that the

[5] For the random variables $X$, $Y$, and $Z$, with $Z$ discrete, Cross and Manski (2002) derive bounds on the "long" regression $\mathbb{E}[Y \mid X, Z]$ when only the "short" conditional distributions $F_{Y|X}$ and $F_{Z|X}$ are known. The random variables $V$ and $D_k$ in our paper correspond to $Z$ and $X$, respectively.

identified set proposed by Cross and Manski is sharp for conditional outcome distributions with both finite and infinite support and conditioning variables with a finite number of support points.

The identification region for a linear functional $L(r) = c'r$ is given by $L\left(\Theta_r^I\right) = \left\{c'r : r \in \Theta_r^I\right\} = \left[\underline{L}\left(\Theta_r^I\right), \bar{L}\left(\Theta_r^I\right)\right]$, which is a closed interval due to linearity of $L(\cdot)$ and the fact that $\Theta_r^I \in K_{kc}\left(\mathbb{R}^L\right)$.

### 3.2.2 Shape constraints

In many cases, restrictions on the shape of the conditional distribution function can be motivated by economic theory. Any kind of shape constraint is a restriction directly on the parameter space $\Theta$. When $\pi$ is identified, and shape restrictions are placed on the conditional expectation function, the sharp identified set is therefore simply the intersection of the restricted parameter space $\Theta$ and $\tilde{\Theta}_r^I$, the identified set in the absence of any such restrictions.

Common restrictions on the conditional expectation function, often motivated by economic theory, are monotonicity and concavity/convexity in the conditioning variable. We only consider weak monotonicity and concavity/convexity, so that the parameter space $\Theta$ is closed and furthermore convex. The same is true with positivity restrictions such as $r \in \mathbb{R}_+^L$.

Without knowledge of the distribution of the interval-censored regressor, meaningful bounds on functionals of the conditional expectation function can only be derived under restrictions on the parameter space. Manski and Tamer (2002) derive bounds under monotonicity of the regression function, which we refer to as Manski-Tamer bounds in our application:

**Proposition 2.** *Manski and Tamer (2002, Prop. 1): Assume that $\pi$ is not identified and that $Supp\,(Y \mid V \in I_k) = \mathbb{R}$, $k = 1, \ldots, K$. Assume further that $\mathbb{E}[Y \mid V = v]$ is weakly increasing in $v$. Then, $\mathbb{E}[Y|V \in I_{k-1}] \leq \mathbb{E}[Y|V = v] \leq \mathbb{E}[Y|V \in I_{k+1}]$ for $v \in I_k$, $k = 1, \ldots, K$, and these bounds are sharp.*

Initially, it may be surprising that the bounds on $\mathbb{E}[Y|V = v]$, $v \in I_k$, are not informed

by $\mathbb{E}[Y|V \in I_k]$,[6] which is only useful in combination with information on the marginal distribution of the latent regressor.

## 3.3 Estimation and inference

### 3.3.1 Representation and computation

For any interval $I_k$, the convex hull of Cross-Manski extreme points can be represented by a system of linear equalities and inequalities, which helps with estimation. The interval mean constraints discussed in the note to Proposition 1 define hyperplanes. In addition, there is a system of $2^{|I_k|} - 2$ inequalities, each of which defines a half-space. Specifically, for any set of support points $A \subsetneq I_k$,

$$\sum_{l:\,\gamma_l \in A} \bar{\pi}^{(k)}_{\delta(l)} r_l \quad \geq \quad \int_0^{\sum_{l:\,\gamma_l \in A} \bar{\pi}^{(k)}_{\delta(l)}} Q_{Y|V \in I_k}(\tau)\, d\tau,$$

where $Q_{Y|V \in I_k}(\tau)$ is the quantile function of $Y$ conditional on $V \in I_k$. We assume that all integrals exist, implying $\|E^m_k\| < \infty$ for $k = 1, \ldots, K$ and all $m$. At a Cross-Manski interval extreme point $E^m_k$, the inequality above holds with equality for any set $A = \{\phi^m_k(1), \ldots, \phi^m_k(|A|)\}$ consisting of the support points in the first $|A|$ positions of permutation $m$. The inequalities provide additional intuition for the sharp identified set: if we take any subset of the support points, we cannot rationalize the corresponding values of the conditional expectation function to be any lower than what would result if we were to assign these support points the left-most portion of $F_{Y|V \in I_k}$ which is of corresponding mass.[7] We will denote the set of required sets for interval $I_k$ by $\bar{A}^{(k)}$, and write $\bar{A} = \cup_{k=1}^K \bar{A}^{(k)}$ with sets $\bar{A}_s$ as elements.[8]

[6] The expectation $\mathbb{E}[Y|V \in I_k]$ is informative in the sense that it can be used to test the shape restrictions placed on $\mathbb{E}[Y|V = \cdot]$.

[7] There are also reverse inequalities from assigning a set of support points at best the right-most commensurate portion, but for any such set $A$, this inequality is implied by the interval mean constraint (3.1) together with the inequality corresponding to its complement $I_k \backslash A$.

[8] The number of inequalities grows at an exponential rate in the number of support points contained in an interval. This can quickly become computationally infeasible due to the large amount of memory required once $|I_k|$ exceeds about ten. When monotonicity is imposed, only a much smaller number of sets of support points needs to be considered. When the conditional expectation function is assumed to be increasing, the

The identified set can thus be written as

$$\Theta_r^I = \left\{ r \in \Theta : \, g^{IM}(r) = 0 \text{ and } g^{CM}(r) \geq 0 \right\},$$

where $g^{IM} : \Theta \to \mathbb{R}^K$ such that

$$g_k^{IM}(r) \;=\; \sum_{l:\, \gamma_l \in I_k} \bar{\pi}_{\delta(l)}^{(k)} r_l - \mathbb{E}\left[ Y \mid V \in I_k \right], \tag{3.1}$$

and $g^{CM} : \Theta \to \mathbb{R}^{|\bar{A}|}$ such that

$$g_s^{CM}(r) \;=\; \sum_{k:\, \bar{A}_s \subset I_k,\, l:\, \gamma_l \in \bar{A}_s} \bar{\pi}_{\delta(l)}^{(k)} r_l - \int_0^{\sum_{l:\, \gamma_l \in \bar{A}_s} \bar{\pi}_{\delta(l)}^{(k)}} Q_{Y|V \in I_k}(\tau)\, d\tau. \tag{3.2}$$

Denoting sample analogues as $\hat{g}^{IM}(\cdot)$ and $\hat{g}^{CM}(\cdot)$, a very natural and computationally efficient analogue estimator is given by $\hat{\Theta}_r^I = \hat{\tilde{\Theta}}_r^I \cap \Theta$, $\hat{\tilde{\Theta}}_r^I = \left\{ r \in \mathbb{R}^L : \, \hat{g}^{IM}(r) = 0 \text{ and } \hat{g}^{CM}(r) \geq 0 \right\}$. Endpoints for the identified set of any functional $L(r) = c'r$ can be obtained from a linear program with $c$ as coefficient vector, the elements of the parameter space as arguments,[9] and feasible region given by $\hat{\Theta}_r^I$. The equalities (3.1) and inequalities (3.2) are straightforward to implement as constraints, and monotonicity and concavity/convexity restrictions can easily be imposed (e.g., Freyberger and Horowitz, 2015).

### 3.3.2 Consistency for $\Theta_r^I$ and $L\left(\Theta_r^I\right)$

For each interval, we estimate the distribution vector $\bar{\pi}^{(k)}$ and the quantile function $Q_{Y|V \in I_k}(\tau)$ by the empirical frequency distribution $\hat{\pi}$ and the empirical distribution function $\hat{F}_{Y|V \in I_k}(\cdot)$, yielding consistent estimates under iid sampling, as assumed throughout. These estimates can be used to construct consistent sample analogue estimates $\hat{g}^{IM}(\cdot)$ and $\hat{g}^{CM}(\cdot)$ and, by

---

inequality (3.2) needs to be satisfied only for sets $\{\gamma_l\}_{l:\, \gamma_l \in I_k,\, \delta(l) \in \{1,\ldots,s\}}$, $s = 1, \ldots, |\, I_k \,| - 1$, in other words, for all subsets of support points that contain the $s$ smallest elements of $I_k$.

[9] For example, if interested in an upper bound on $\mathbb{E}\{Y \mid V = \gamma_l\}$, the econometrician would specify $c$ as the $l$-th column of the identity matrix $I_L$ and obtain as bound the value of the linear program $\max_x c'x$ s.t. $x \in \Theta_r^I$, while a lower bound can be obtained by using its negative as coefficient vector. Bounds on a difference such as $\mathbb{E}\left\{Y \mid V = \gamma_{l_2}\right\} - \mathbb{E}\left\{Y \mid V = \gamma_{l_1}\right\}$ can be obtained by specifying $c$ such that $c_{l_1} = -1$, $c_{l_2} = 1$, and $c_l = 0$ if $l \notin \{l_1, l_2\}$.

implication, consistent estimates for the extreme points, $\left\{\hat{E}_k^m\right\}_{k=1,\,m=1}^{K,\,|I_k|!}$, as so-called location (or L-)estimators (e.g., Koenker and Portnoy, 1987).

Even in correctly specified problems, the sample identified set based on an analogue estimator can be empty while the population identified set is not. This problem carries through in the asymptotic limit whenever the population identified set is of a particular lower-dimensional kind. For the proposed estimator $\hat{\Theta}_r^I$ to be consistent, we require that $\tilde{\Theta}_r^I$ cannot be separated from $\Theta$.[10] In our setting, separability implies that the convex polytope $\tilde{\Theta}_r^I$ is exactly tangent to the parameter set $\Theta$.[11] We therefore make the interiority assumption below.

**Assumption 3.** *Suppose that* $\tilde{\Theta}_r^I \cap \text{int}\,(\Theta) \neq \emptyset$.

**Proposition 4.** *Under assumption 3,* $\hat{\Theta}_r^I \xrightarrow{H} \Theta_r^I$.

*Note.* Yildiz (2012) requires $\Theta_r^I \subseteq \text{int}\,(\Theta)$ for consistency of analogue estimators. The convexity of $\Theta_r^I$ allows us to weaken this assumption.

The non-separability restriction implies that the estimator will not be uniformly consistent over all data-generating processes which (just) satisfy possible shape restrictions.[12] While unfortunate, this restriction allows us to present a coherent framework for practically simple estimation and inference.

Since $\Theta_r^I$ is convex, it is fully described by its support function $s\left(\cdot, \Theta_r^I\right)$. Bounds on any linear functional follow from its definition and homogeneity property.[13] For two compact

[10] Rockafellar and Wets (1998, p. 129) define that "two convex sets $C_1$ and $C_2$ in $\mathbb{R}^n$ can't be *separated* (even improperly) [...] [if] there is no hyperplane $H$ such that $C_1$ lies in one of the closed half-spaces associated with $H$ while $C_2$ lies in the other." For deterministic sequences of convex sets, Exercise 4.33 in the same text similarly requires non-separability for convergence.

[11] Shape constraints (and other constraints on the parameter space) can result in point identification if they restrict the parameter space in such a way that the only vector in $\Theta$ satisfying constraints (3.1) and (3.2) is a point on its boundary. Constraints (3.1) and (3.2) by themselves yield point identification only in pathological cases.

[12] A simple point-identified example for failure is provided by the $K = 2$ case in which $F_{Y|V\in I_1} = F_{Y|V\in I_2}$ and monotonicity is assumed. In this case, the more computationally difficult methods in Chernozhukov *et al.* (2007) can be applied using our representation.

[13] Any compact convex set $A \in K_{kc}\left(\mathbb{R}^L\right)$ is fully characterized by its support function $s\,(\cdot, A)$ defined as

convex sets $A$, $B \in K_{kc}\left(\mathbb{R}^L\right)$, the Hausdorff distance $H\left(A,B\right)$ can be obtained from the support function as $H\left(A,B\right) = \sup_{p\in\mathcal{S}^{L-1}}|s\left(p,A\right) - s\left(p,B\right)|$ (Li *et al.*, 2002, cor. 1.1.10). Therefore, consistent estimation of $\Theta_r^I$ implies that bounds on $L\left(r\right)$ can be consistently estimated by $L\left(\hat{\Theta}_r^I\right)$.

### 3.3.3 Inference

We present an inference procedure with both Bayesian and frequentist interpretation, based on Kline and Tamer. Two cases are important to distinguish: $\pi$, the marginal distribution of $V$, the interval-censored regressor, may be estimated from the same sample of individuals or units as the conditional outcome distribution $F_{Y|V\in I_k}$, $k = 1,\ldots,K$, or from a different sample. For instance, the researcher might use an auxiliary dataset with negligible or without any overlap, or the data provider might collect the precise value of a regressor of interest, censor it for confidentiality reasons, but make available an estimate of the marginal distribution based on the original data from the same sample. Such overlap results in a covariance term which cannot be estimated. Conservative inference, however, can be based on a simple upper bound of the covariance matrix.

Kline and Tamer develop a Bayesian approach to inference for models in which the identified set for a partially identified object of interest, which we momentarily call $\theta$, is given by a mapping from some point-identified statistic of the data, $\mu$. While the previous literature typically sets up a prior likelihood directly for $\theta$, Kline and Tamer instead specify a prior for $\mu$, implying a prior over identified sets for $\theta$ rather than the parameter value itself. In our example, $\theta$ is the value of some linear functional $L\left(r\right)$, and it is easiest to choose $\mu$ as the stacked vector of all extreme points given all permutations. We then have $d_\mu \equiv \dim\left(\mu\right) = \sum_{k=1}^K \mid I_k \mid \cdot \mid I_k \mid !$, and at its true value, which we refer to as $\mu_0$, $\mu_0 = \{E_k^m\}_{k,m}$. In addition, we conveniently define $\mu^{(k,m)}$ to refer to its subvector

$s\left(p,A\right) \equiv \sup_{a\in A}\langle p,a\rangle$, where $p \in \mathbb{R}^L$. The support function characterizes the location of hyperplanes tangent to the set by providing the signed distance between the origin and a hyperplane orthogonal to $p$ and tangent to $A$. The support function is homogeneous of degree one: $\forall p \in \mathbb{R}^L$, $\lambda \in \mathbb{R}$, $s\left(\lambda p,A\right) = \lambda s\left(p,A\right)$. See, for instance, Mas-Colell *et al.* (1995, pp. 64-65).

corresponding to $E_k^m$, so that $\tilde{\Theta}_r^I = \text{conv}\left(\times_k \{E_k^m\}_m\right) = \text{conv}\left(\times_k \left\{\mu_0^{(k,m)}\right\}_m\right)$. We denote the sample analogue based on the data $\mathbf{X}$ by $\hat{\mu} = \left\{\hat{E}_k^m\right\}_{k,m}$.

Rather than specify a parametric prior for $\mu$, Kline and Tamer use the Bayesian bootstrap, which places a prior on the multinomial distribution of the observed data (Rubin, 1981, Chamberlain and Imbens, 2003, Parzen and Lipsitz, 2007). Given a limiting uninformative Dirichlet prior for $\mu$, draws from its posterior can be approximated by weighting each observation by a random variable $W_{\cdot,b} \overset{iid}{\sim} Exp(1)$ in each bootstrap replication $b$. A $(1-\alpha)$-level credible set $\mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}(\mathbf{X})$ for the identified set is then constructed as a set covering a fraction $1-\alpha$ of the bootstrap draws on the identified set.[14] Unlike conservative projection methods, this inference procedure is therefore exact, at least in the two-sample case. In many settings, the Bernstein-von Mises Theorem allows the interpretation of Bayesian credible sets as frequentist confidence intervals, at least for large samples. For the procedure outlined above to allow valid frequentist inference, the data statistic $\mu$ needs to have a large-sample normal posterior, $\sqrt{N_1}(\mu - \hat{\mu}) \mid \mathbf{X} \approx \mathcal{N}(0, \Sigma_0)$, with the same covariance matrix as in the limiting distribution of the sample analogue, $\sqrt{N_1}(\hat{\mu} - \mu_0) \approx \mathcal{N}(0, \Sigma_0)$. We verify this for the case in which $\left\{\hat{F}_{Y|V\in I_k}\right\}_{k=1}^K$ and $\hat{\pi}$ are based on non-overlapping samples of size $N_1$ and $N_2$, respectively, where we assume that $N_1 \to \infty$, $N_2 \to \infty$ s.t. $\sqrt{\frac{N_1}{N_2}} \to \kappa \in [0, \infty)$.[15] When instead, both estimates come from the same sample of size $N_1$ (or samples with significant overlap), this condition will generally not be satisfied.

We first state our results on the asymptotic distribution of the estimate $\hat{\mu}$ (Prop. 6) and the large-sample posterior of $\mu$ (Prop. 7) provided that the outcome is continuously distributed (Asm. 5). Together with Lemma 8 and under Assumption 3, the Bayesian credible sets are asymptotically valid frequentist confidence intervals. Finally, we show how to perform conservative inference when outcome and regressor distribution are estimated from the same

[14] We take draws directly from the posterior for the identified set of the functional of interest by computing bounds for every draw from the posterior of $\mu$. The resulting estimates are for credible sets and confidence intervals of the identified set rather than the parameter value itself.

[15] Hellerstein and Imbens (1999) study similar asymptotics for wage regressions from a combination of the National Longitudinal Survey Young Men's Cohort and a 1% Census sample.

sample.

**Assumption 5.** *On all intervals $I_k$, $k = 1, \ldots, K$, $Y \mid V \in I_k$ has a continuous strictly positive density on some convex subset of $\mathbb{R}$ with bounded first derivative. Furthermore, $\mathbb{E}\left[Y^2 \mid V \in I_k\right] < \infty$.*

**Proposition 6.** *Under assumption 5, $\sqrt{N_1}\left(\hat{\mu} - \mu_0\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma\right)$ with $\Sigma = \Sigma_{0,1}$ and $\Sigma = \Sigma_{0,2}$ in the one-sample and two-sample case, respectively, for some covariance matrices $\Sigma_{0,1}$ and $\Sigma_{0,2}$.*

*Note.* $\Sigma_0$ is generally positive definite, except under $\kappa = 0$ asymptotics when $\bar{\pi}_s^{(k)} = \bar{\pi}_{s'}^{(k)}$ for some $k$ and $s \neq s'$. In that case, the corresponding support points get the same "slices" of $\hat{F}_{Y|V \in I_k}$ (up to lower-order terms) when swapping one for the other in a given permutation, and so the correlation of these elements of $\hat{\mu}$, centered and multiplied by $\sqrt{N_1}$, is one.

**Proposition 7.** *Suppose that posterior draws on $\mu$ are based on the Bayesian bootstrap. Under assumption 5, $\sqrt{N_1}\left(\mu - \hat{\mu}\right) \mid \mathbf{X} \xrightarrow{tv} \mathcal{N}\left(0, \Sigma\right)$ with $\Sigma = \tilde{\Sigma}_{0,1} \neq \Sigma_{0,1}$ and $\Sigma = \Sigma_{0,2}$ in the one-sample and two-sample case, respectively, for some covariance matrix $\tilde{\Sigma}_{0,1}$.*

Kline and Tamer establish the frequentist coverage properties of $\mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}\left(\mathbf{X}\right)$ (see Thm. 2 in their paper) using convergence in total variation of $\hat{\mu}$ (Asm. 6, ibid) to show that for all $\epsilon > 0$, in sufficiently large sample, $\Pr\left(\sqrt{N_1}\left(\hat{\mu} - \mu_0\right) \in A\right) \in \Pr_{\mathcal{N}(0,\Sigma_0)}\left(A\right) + \left[-\epsilon, \epsilon\right]$ for all Borel sets $A$. While convergence in total variation is not satisfied here,[16] and Proposition 6 only establishes the weaker convergence in distribution, we can apply Proposition 6 and the Portmanteau Lemma (van der Vaart, 2000, Lemma 2.2), which states that for any random vectors $U_n$ and $U$, $U_n \xrightarrow{d} U$ is equivalent to $\Pr\left(U_n \in A\right) \to \Pr\left(U \in A\right)$ for all Borel sets $A$ with $\Pr\left(U \in \text{bnd}\left(A\right)\right) = 0$. This requires us to show that the relevant Borel sets are continuity sets, meaning they have zero probability on the boundary, under $\mathcal{N}\left(0, \Sigma_0\right)$. For this, define $\mu^I\left(r\right) \equiv \left\{\mu \in \mathbb{R}^{d_\mu} : r \in \left(\text{conv}\left(\times_k \left\{\mu^{(k,m)}\right\}_m\right) \cap \Theta\right)\right\}$ to be the set of values of $\mu$ (yielding a set of Cross-Manski-type extreme points) for which $r$ is contained in the identified set.

[16] We cannot apply a central limit theorem in total variation (e.g., van der Vaart, 2000, Prop. 2.31) because the relevant (transformed) random variables follow a discrete distribution.

**Lemma 8.** *There exists an* $\bar{N}$ *such that for* $N_1 \geq \bar{N}$,

$$\Pr_{\mathcal{N}(0,\Sigma_0)}\left(\text{bnd}\left(\sqrt{N_1}\left(\cap_{\delta\in\left(\mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}(\mathbf{X})\right)}{}^{C}\cap_{\{r:\,L(r)=\delta\}}\mu_I\left(r\right)^C-\hat{\mu}\right)\right)\right) \;=\; 0.$$

*Note.* The above is a frequentist statement for a random sample $\mathbf{X}$ and given our rule for constructing $\mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}(\mathbf{X})$. When the parameter space is unrestricted, $\Theta=\mathbb{R}^L$, this only requires that none of the extreme points in $\times_k\{E_k^m\}_m$ lie in the two level planes generated by $\left\{r\in\mathbb{R}^L:\,L\left(r\right)\in\text{bnd}\left(\mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}(\mathbf{X})\right)\right\}$, i.e., $c'r$ equal to either the lower or upper endpoint of the interval $\mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}(\mathbf{X})$. With shape restrictions, such that $\Theta\subsetneq\mathbb{R}^L$, we need to show that the probability that the convex hull of extreme points is tangent to $\Theta\cap\left\{r\in\mathbb{R}^L:\,L\left(r\right)\in\mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}(\mathbf{X})\right\}$ is zero under the limiting distribution $\mathcal{N}\left(0,\Sigma_0\right)$.

In addition, Kline and Tamer require that the construction of the credible sets across datasets satisfies an asymptotic independence condition (Asm. 5, ibid), which is satisfied whenever there is a unique solution to the linear program.[17]

**Assumption 9.** *For the linear functional* $L\left(r\right)$*, the sets of minimizers and maximizers* $\arg\min_{r\in\Theta_r^I}L\left(r\right)$ *and* $\arg\max_{r\in\Theta_r^I}L\left(r\right)$ *are each a singleton.*

*Note.* Multiplicity can arise when the coefficient vector $c$ of the linear functional $L\left(r\right)=c'r$ is exactly orthogonal to one of the (flat) faces of $\Theta_r^I$.

Furthermore, asymptotic independence requires interiority as in Assumption 3. Otherwise, even in large sample, the non-emptiness condition in equation (B.4) (see proof of Lemma 8) will not necessarily be satisfied with probability one, and the credible set will thus depend on the sample even asymptotically.

[17]The failure of the Bayesian bootstrap in case of multiplicity is related to the inconsistency of the generic nonparametric bootstrap in such settings: the behavior of the limiting bootstrap distribution, as the number of draws grows, is discontinuous in the underlying distribution, and so, draws from the sample distribution (with a unique solution almost surely) do not mimic draws from the population distribution (with multiple solutions). See Freyberger and Horowitz (2015) for further discussion and a modification to the nonparametric bootstrap. In their simulation study, both modified and unmodified bootstrap perform well.

An interesting issue arises when any of the Bayesian bootstrap draws return an empty set due to infeasibility of the linear program. Dropping these draws corresponds to a limiting Dirichlet prior with additional shape constraints. While it is difficult to write down such a truncated prior, it is very simple to draw from the corresponding posterior. In the frequentist interpretation, under Assumption 3, the empty draws are a finite-sample problem. With an otherwise smooth prior, the draws on the identified set which are kept come from a truncated bootstrap distribution and may on average be "too large."[18]

When both the estimate of the outcome distribution, $\left\{\hat{F}_{Y|V\in I_k}\right\}_{k=1}^{K}$, and the estimate of the marginal distribution of the regressor, $\hat{\pi}$, are based on the same sample, the generic Bayesian bootstrap does not yield consistent estimates for confidence intervals. The problem is that the two terms in the asymptotic representation of $\hat{\mu}$ related to the estimation of these objects are potentially correlated, yet when it is unknown due to censoring which outcome and regressor values belong to the same unit, we cannot assign them identical draws for the bootstrap weights. As a result, the covariance term is lost in our estimate of the posterior distribution. However, by the simple observation that for any two random variables $X$ and $Y$ with finite variances, $Var\left(X+Y\right) \leq 2\left(Var\left(X\right)+Var\left(Y\right)\right)$ (Lemma 14), we can construct a very simple upper bound $\bar{\Sigma} \geq \Sigma_0$. For any solution to the linear program, equal to a weighted sum of the elements in $\mu$, we can thus provide an upper bound on its variance as well, and so, provided that said solution is unique, we can bound the asymptotic variance of the endpoint estimates. For a confidence interval containing the identified set with probability no less than $1-\alpha$ (in the frequentist sense), we construct "worst-case" intervals $\left[\underline{L}\left(\hat{\Theta}_r^I\right) - \sqrt{2}\left(\underline{L}\left(\hat{\Theta}_r^I\right) - \underline{L}_{\alpha/2}\right),\ \bar{L}\left(\hat{\Theta}_r^I\right) + \sqrt{2}\left(\bar{L}_{1-\alpha/2} - \bar{L}\left(\hat{\Theta}_r^I\right)\right)\right]$, where $\underline{L}_{\alpha/2}$ and $\bar{L}_{1-\alpha/2}$ are equal to the $\alpha/2$ and $1-\alpha/2$ quantile of the estimated posterior distribution of $\underline{L}\left(\Theta_r^I\right)$ and $\bar{L}\left(\Theta_r^I\right)$, respectively.[19,20]

---

[18] See Gelman (1996) for a similar problem in point-identified shape-restricted models with a uniform prior on the parameters.

[19] The "worst-case" character comes from the fact that the covariance between the two endpoints cannot be estimated, and the proposed interval has coverage at least $1-\alpha$ for any value of the correlation. In our application, the endpoints are positively correlated, so the excess length that results may not be great.

[20] Tighter bounds could possibly be constructed based on the Cauchy-Schwarz inequality, but this would

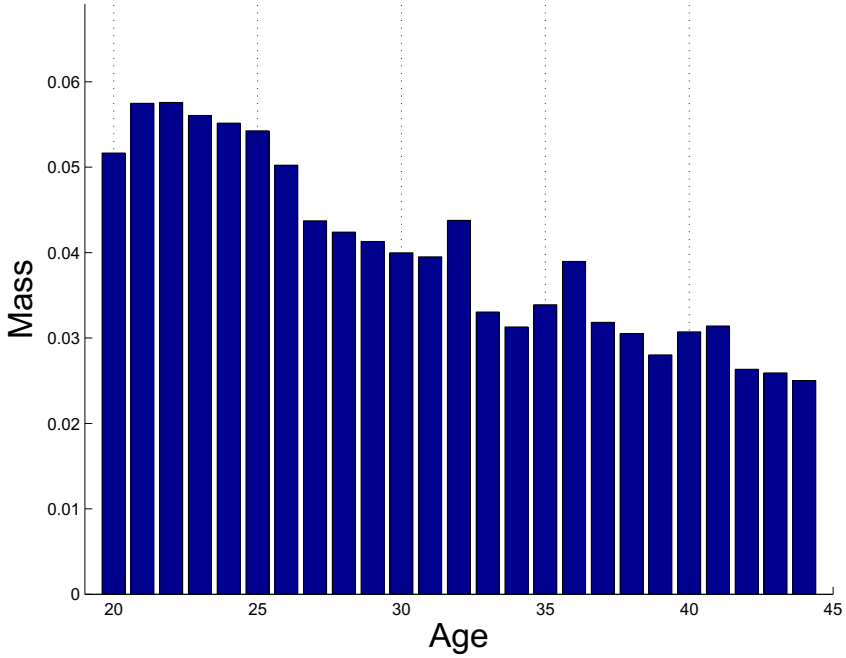


**Figure 3.1:** *Age histogram of white male full-time workers with 12 years of schooling.*

## 3.4 Application to age-earnings profiles

In this section, we illustrate the application and potential usefulness of the methods discussed above by means of a wage regression example. Specifically, we will look at age-earnings profiles, such that $Y$ is earnings, and $V$ is age. From the outgoing rotation group of the 1979 full-year Current Population Survey, we select white male full-time workers and employees aged 20 to 44 with 12 years of schooling, yielding a sample of 18,913 observations.[21] Our earnings measure is the log of usual weekly earnings.[22] Age is provided in years, so that we can use the sample histogram (figure 3.1) as our estimate $\hat{\pi}$. For illustration, we artificially censor age into five-year intervals $\{20,\ldots,24\},\ldots,\{40,\ldots,44\}$, so that $K = 5$ and $L = 25$.

require estimation of the different elements in the asymptotic representation in the proof of Proposition 6.

[21] The data files for the Merged Outgoing Rotations Groups are available at http://www.nber.org/morg/annual/.

[22] We choose the 1979 wave mostly because only a negligible fraction of our sample have top-coded earnings (below 0.4%), which we simply use at their top-coded value ($999). While CPS respondents report total annual household income in intervals, they are asked to provide an exact number for weekly earnings in a given job.

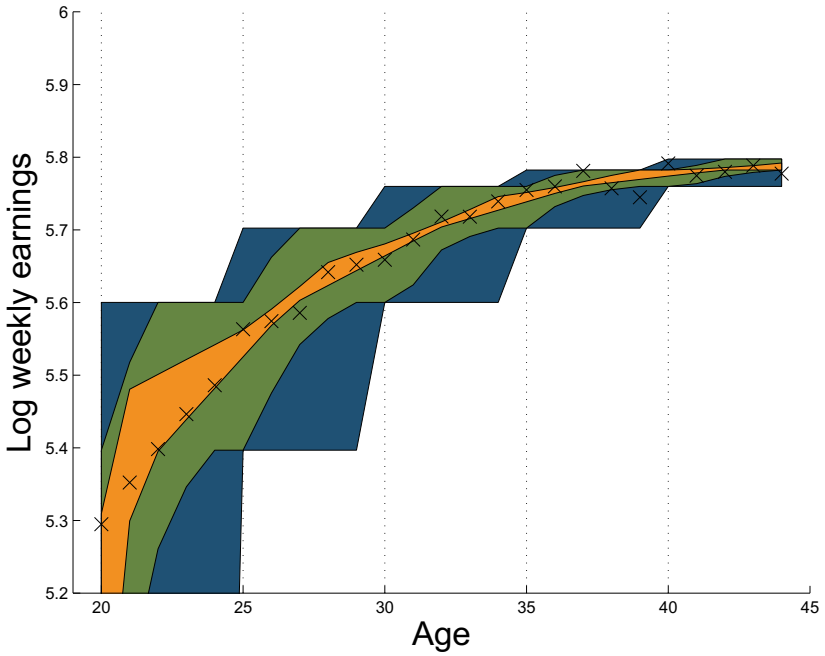


**Figure 3.2:** *Average log weekly earnings by age. Cross-Manski bounds with (orange) and without (green) concavity; Manski-Tamer bounds (blue).*

Figure 3.2 shows estimated bounds on the age-earnings profile using monotonicity (green) and monotonicity and concavity (orange) inside the Manski-Tamer bounds (blue), which only use monotonicity but not information on the distribution of the regressor (blue).[23] Both monotonicity and concavity are motivated by standard human capital models. The ticks give the actual earnings mean for each age based on the original uncensored sample. Since both restrictions are close to satisfied in our sample, almost all of them are contained in the identified set. Especially under concavity, but also under monotonicity, the slices become very narrow as we move to the right of the support, and are substantially narrower than the Manski-Tamer bounds.

For additional comparison, figure 3.3 shows the bounds from the previous figure using both restrictions and information on the marginal distribution of the regressor $\pi$ (orange) along with bounds which do not use $\pi$ but only the shape restrictions of monotonicity (blue) and

[23]Figure 2 does not plot the identified set $\Theta_r \subset \mathbb{R}^{25}$ for the earnings function $r$, but rather slices of the identified set along every dimension corresponding to a given age.

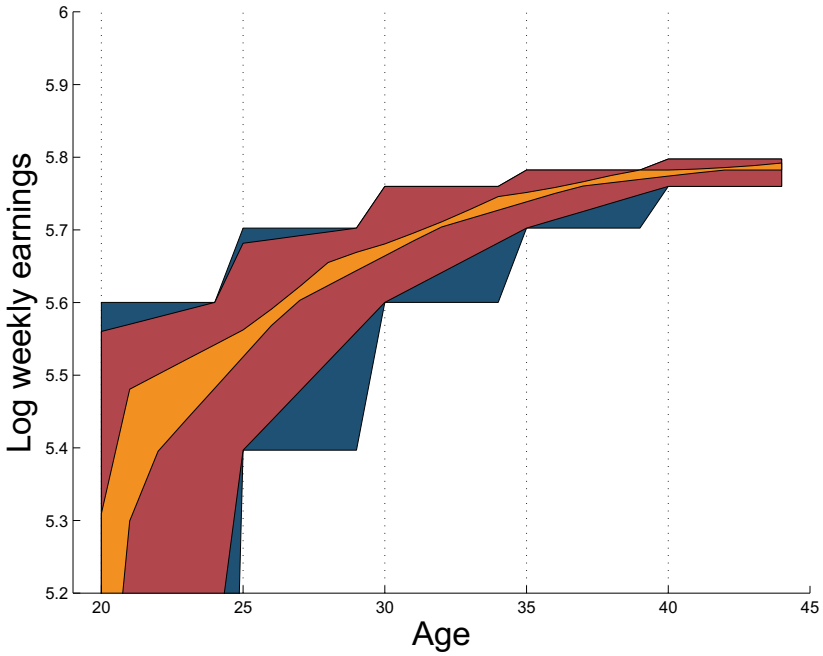


**Figure 3.3:** *Average log weekly earnings by age. Manski-Tamer bounds without (blue) and with concavity (red); Cross-Manski bounds with concavity (orange).*

additionally concavity (red). Again, it is immediately clear that incorporating information on the marginal distribution substantially narrows the bounds. At least visually, the distribution information makes a greater contribution to this compression than does concavity as an additional shape constraint.

A non-sharp superset of the identified set can be estimated using only the interval mean equality constraints (3.1) but not the inequalities (3.2), which reduces the time cost for the programmer as well as the memory burden if $\max_k \mid I_k \mid$ is large; computation time is less of a concern with modern linear programming techniques. The estimated set is actually unchanged except for a slight difference on the left end of the age distribution. The constraints that derive from the outcome distribution as opposed to only its first moment are rather unrestrictive and therefore typically do not bind. They may be more informative when the linear functional of interest is of a different form.

At each age, we would like to provide a 95% confidence interval for the identified set of mean earnings. Since we are in the one-sample case (with $N_1 = N_2 = 18,913$), we can only

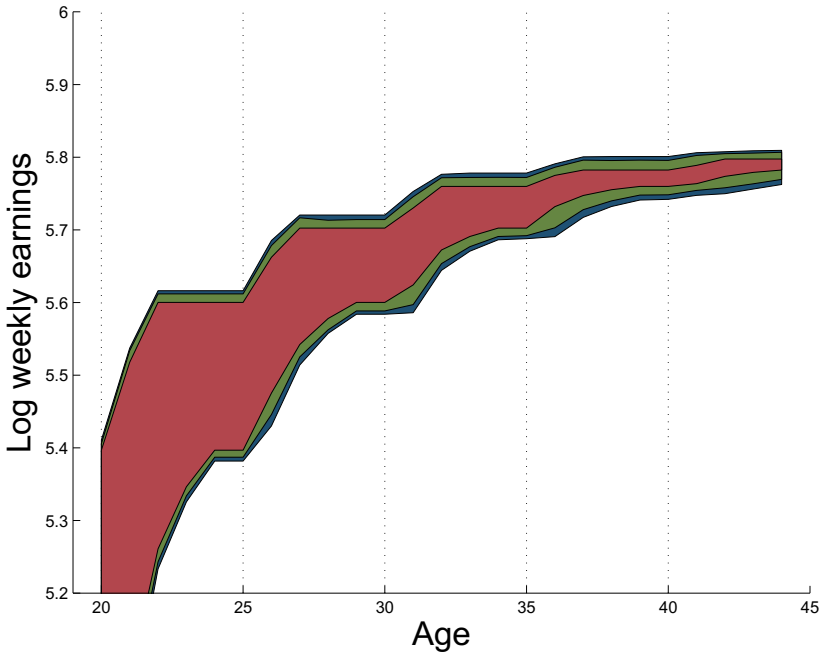


**Figure 3.4:** *Average log weekly earnings by age. Estimate under monotonicity (red), asymptotically exact 95% CI (green), conservative 95% CI (blue).*

give a conservative approximation to that set. However, the artificial nature of our censoring allows us to estimate the ordinarily infeasible asymptotically exact confidence interval by applying the same bootstrap weight to a given unit when used to estimate $F_{Y|V\in I_k}$ and $\pi$. We can compare this to both the conservative approximation and an invalid estimate which ignores the one-sample problem by falsely assuming that the estimates for $F_{Y|V\in I_k}$ and $\pi$ come from separate samples.

The invalid and the infeasible confidence interval are straightforward to estimate by constructing an interval which contains 95% of the respective interval bootstrap draws. We select the shortest such intervals. The conservative approximation is obtained as described in section 3.3.3. The results, based on $B = 1,000$ bootstrap draws, are shown in figures 3.4 (monotonicity) and 3.5 (monotonicity and concavity) with the estimate from the previous subsection in red, the infeasible asymptotically exact confidence interval in green, and the conservative approximation in blue. The estimates which assume that the data come from different samples are indistinguishable from the exact interval and therefore not plotted.

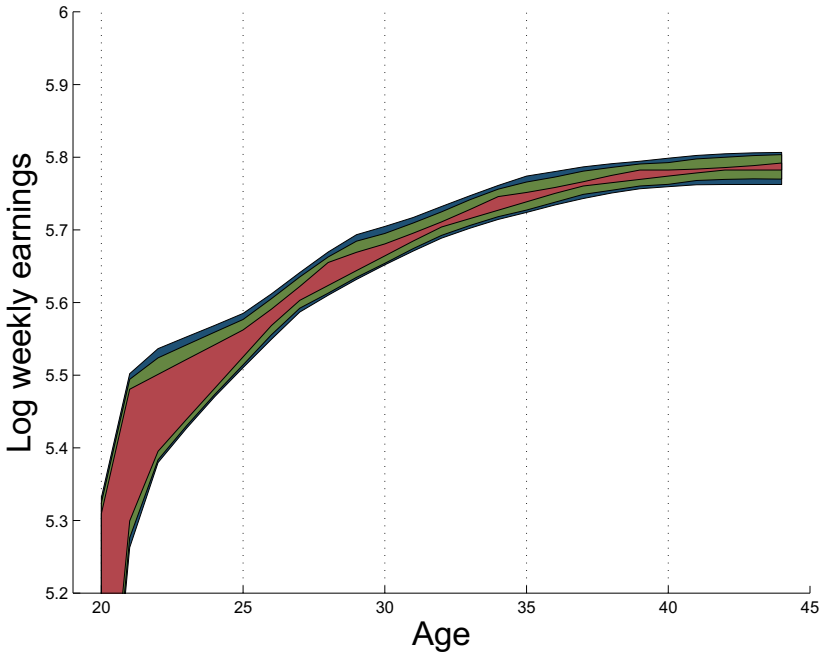


**Figure 3.5:** *Average log weekly earnings by age. Estimate under monotonicity and concavity (red), asymptotically exact 95% CI (green), conservative 95% CI (blue).*

Under the monotonicity restriction alone, we throw out around 6% of observations, with concavity around 10%.[24] The respective computation times on a 3.40 GHz CPU with 8 GB RAM are around 10 and 30 minutes. Recall that one of the requirements for asymptotically valid inference is uniqueness (Asm. 9). Duality implies that the optimal value of each linear program is given by a weighted sum of the expectations and integrals in (3.1) and (3.2) with weights given by the optimal dual variables equal to the Lagrange multipliers on the constraints in the primal problem. Only the interval mean constraints tend to bind, corresponding to a non-zero multiplier. With only the monotonicity restriction, the Lagrange multipliers are fairly stable between different bootstrap draws, while with concavity, they appear to be more volatile, and for bounds at some of the support points, there are two different solutions that each appear around 50% of the time. In finite sample, this does not imply multiplicity. The conservative approximation also requires uniqueness, and departures could push into the direction of under-

[24]We start with an interior-point algorithm, and switch to a simplex algorithm in case the former does not successfully return an estimate.

as well as overcoverage, depending on the sign of that part of the correlation between the multiple solutions which is due to the estimation of outcome and regressor distribution on the same sample. This may be a second-order concern in our application, since the first-order effect of the covariance term makes virtually no difference, as the comparison between the infeasible and the invalid confidence interval shows.

## Conclusion

We show how to obtain the sharp identified set for the expectation function when the regressor is interval-censored but its marginal distribution is estimable. To this end, we establish the form of the sharp identified set and provide an estimator which is consistent under a set of conditions. An application to age-earnings profiles in the CPS show that these methods can be very useful.

Inference with valid Bayesian and frequentist interpretation can be performed using the Bayesian bootstrap. We verify the necessary conditions for a Bernstein-von Mises-type result, one of which, on the mode of convergence, is non-standard and may be useful in other work. When the outcome distribution and the marginal distribution of the regressor are estimated on the same sample, inference becomes more complicated and we provide a conservative approximation to confidence intervals with no less than the desired coverage.

# References


Agrawal, S. and Goyal, N. (2012). Analysis of Thompson Sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*.

— and — (2013). Further optimal regret bounds for Thompson Sampling. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Andersen, E. B. (1970). Asymptotic Properties of Conditional Maximum-Likelihood Estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, **32** (2), 283–301.

Athey, S. and Imbens, G. W. (2015). *Machine Learning for Estimating Heretogeneous Casual Effects*. Tech. rep.

Berry, S., Levinsohn, J. and Pakes, A. (1995). Automobile Prices in Market Equilibrium. *Econometrica*, **63** (4), 841–890.

Berry, S. T. (1994). Estimating Discrete-Choice Models of Product Differentiation. *RAND Journal of Economics*, **25** (2), 242–262.

Besanko, D., Doraszelski, U., Kryukov, Y. and Satterthwaite, M. (2010). Learning-by-Doing, Organizational Forgetting, and Industry Dynamics. *Econometrica*, **78** (2), 453–508.

Bruegel (2016). Big data, digital platforms and market competition.

Bundorf, M. K., Levin, J. and Mahoney, N. (2012). Pricing and Welfare in Health Plan Choice. *American Economic Review*, **102** (7), 3214–3248.

Bureau, U. C. (2015). U.S. Electronic Shopping and Mail-Order Houses (NAICS 4541) - Total and E-commerce Sales by Merchandise Line (1999-2014).

Chamberlain, G. (1980). Analysis of Covariance with Qualitative Data. *The Review of Economic Studies*, **47** (1), 225–238.

— (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, **34** (3), 305–334.

— and Imbens, G. W. (2003). Nonparametric Applications of Bayesian Inference. *Journal of Business & Economic Statistics*, **21** (1), 12–18.

Chapelle, O. and Li, L. (2012). An Empirical Evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems 24 (NIPS-11)*, Neural Information Processing Systems Foundation, pp. 2249–2257.

Chernozhukov, V., Hong, H. and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, **75** (5), 1243–1284.

Cho, J., Lee, K., Shin, E., Choy, G. and Do, S. (2016). *How Much Data is Needed to Train a Medical Image Deep Learning System to Achieve Necessary High Accuracy?* Tech. rep., Department of Radiology, Massachusetts General Hospital and Harvard Medical School.

Conlon, C. T. and Mortimer, J. H. (2013). Demand Estimation under Incomplete Product Availability. *American Economic Journal: Microeconomics*, **5** (4), 1–30.

Cross, P. J. and Manski, C. F. (2002). Regressions, Short and Long. *Econometrica*, **70** (1), 357–368.

de Fortuny, E. J., Martens, D. and Provost, F. (2013). Predictive Modeling with Big Data: Is Bigger Really Better? *Big Data*, **4** (1), 215–226.

Department of Justice (2010). Statement of the Department of Justice Antitrust Division on Its Decision to Close Its Investigation of the Internet Search and Paid Search Advertising Agreement Between Microsoft Corporation and Yahoo! Inc.

Dickstein, M. J. (2014). *Efficient Provision of Experience Goods: Evidence from Antidepressant Choice*. Tech. rep.

Dinerstein, M., Einav, L., Levin, J. and Sundaresan, N. (2014). *Consumer Price Search and Platform Design in Internet Commerce*. Tech. rep.

Eckles, D. and Kaptein, M. (2014). Thompson sampling with the online bootstrap. *CoRR*, **abs/1410.4009**.

Einav, L., Finkelstein, A., Ryan, S. P., Schrimpf, P. and Cullen, M. R. (2013). Selection on Moral Hazard in Health Insurance. *American Economic Review*, **103** (1), 178–219.

—, Jenkins, M. and Levin, J. (2012). Contract Pricing in Consumer Credit Markets. *Econometrica*, **80** (4), 1387–1432.

European Commission (2010). Case No COMP/M.5727 - Microsoft/Yahoo! Search Business.

Feenberg, D., Ganguli, I., Gaulé, P. and Gruber, J. (2017). It's Good to Be First: Order Bias in Reading and Citing NBER Working Papers. *Review of Economics and Statistics*, **99** (1), 32–39.

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2008). *Longitudinal Data Analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press.

Fradkin, A. (2015). *Search Frictions and the Design of Online Marketplaces*. Tech. rep.

Freyberger, J. and Horowitz, J. L. (2015). Identification and shape restrictions in nonparametric instrumental variables estimation. *Journal of Econometrics*, **189** (1), 41–53.

Gelman, A. (1996). Bayesian model-building by pure thought: some principles and examples. *Statistica Sinica*, **6**, 215–232.

General Insurance Research Organisation (2009). Winner's Curse: The Unmodelled Impact of Competition.

Gittins, J., Glazebrook, K. and Weber, R. (2011). *Multi-armed Bandit Allocation Indices*. 2nd edn.

— and Jones, D. (1974). A Dynamic Allocation Index for the Sequential Design of Experiments. In J. Gani (ed.), *Progress in Statistics*, Amsterdam, NL: North-Holland, pp. 241–266.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, **41** (2), 148–177.

Goldman, M. and Rao, J. M. (2014). *Experiments as Instruments: Heterogeneous Position Effects in Sponsored Search Auctions*. Tech. rep.

Graepel, T., Candela, J. Q., Borchert, T. and Herbrich, R. (2010). Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine. In *Proceedings of the 27th International Conference on Machine Learning ICML 2010, Invited Applications Track (unreviewed, to appear)*.

Greene, W. (2004). The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *Econometrics Journal*, **7** (1), 98–119.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis.

Hausman, J. A. (1994). *Valuation of New Goods under Perfect and Imperfect Competition*. NBER Working Papers 4970, National Bureau of Economic Research, Inc.

Hellerstein, J. K. and Imbens, G. W. (1999). Imposing Moment Restrictions From Auxiliary Data By Weighting. *The Review of Economics and Statistics*, **81** (1), 1–14.

Hsieh, C.-C., Neufeld, J., King, T. and Cho, J. (2015). Efficient Approximate Thompson Sampling for Search Query Recommendation. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, SAC '15, New York, NY, USA: ACM, pp. 740–746.

Jeziorski, P. and Moorthy, S. (2015). *Advertiser prominence effects in search advertising*. Tech. rep.

— and Segal, I. (2015). What Makes Them Click: Empirical Analysis of Consumer Demand for Search Advertising. *American Economic Journal: Microeconomics*, **7** (3), 24–53.

Kaufmann, E., Korda, N. and Munos, R. (2012). *Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 199–213.

Kline, B. and Tamer, E. (2016). Bayesian inference in a class of partially identified models. *Quantitative Economics*, **7** (2), 329–366.

Koenker, R. and Portnoy, S. (1987). L-Estimation for Linear Models. *Journal of the American Statistical Association*, **82** (399), 851–857.

Kolhatkar, R. (2004). *Grassmann Varieties*. Master's thesis, McGill University.

Lahaie, S. and McAfee, R. P. (2011). Efficient Ranking in Sponsored Search. In *Proceedings of the Seventh Workshop on Ad Auctions*.

Li, S., Ogura, Y. and Kreinovich, V. (2002). *Limit Theorems and Applications of Set-Valued and Fuzzy Set-Valued Random Variables*. Theory and Decision Library B, Springer.

Magnac, T. and Maurin, E. (2008). Partial identification in monotone binary models: Discrete regressors and interval data. *The Review of Economic Studies*, **75** (3), 835–864.

Manski, C. F. and Tamer, E. (2002). Inference on Regressions with Interval Data on a Regressor or Outcome. *Econometrica*, **70** (2), 519–546.

Mas-Colell, A., Whinston, M. D. and Green, J. R. (1995). *Microeconomic Theory*. Oxford University Press.

Molinari, F. and Peski, M. (2006). Generalization Of A Result On 'Regressions: Short and Long'. *Econometric Theory*, **22** (1), 159–163.

Narayanan, S. and Kalyanam, K. (2015). Position Effects in Search Advertising and Their Moderators: A Regression Discontinuity Approach. *Marketing Science*, **34** (3), 388–407.

Pandey, S., Chakrabarti, D. and Agarwal, D. (2007). Multi-armed Bandit Problems with Dependent Arms. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, New York, NY, USA: ACM, pp. 721–728.

Parzen, M. and Lipsitz, S. R. (2007). Perturbing the minimand resampling with gamma(1,1) random variables as an extension of the bayesian bootstrap. *Statistics & Probability Letters*, **77** (6), 654–657.

Perlich, C., Provost, F. and Simonoff, J. S. (2003). Tree Induction vs. Logistic Regression: A Learning-curve Analysis. *Journal of Machine Learning Research*, **4**, 211–255.

Ridder, G. and Moffitt, R. (2007). The Econometrics of Data Combination. In J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, vol. 6, *75*, Elsevier.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

— and Wets, R. J.-B. (1998). *Variational Analysis*, *Grundlehren der mathematischen Wissenschaften*, vol. 317. Springer.

Rubin, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics*, **9** (1), 130–134.

Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly Parameterized Bandits. *Mathematics of Operations Research*, **35** (2), 395–411.

Russo, D. and Van Roy, B. (2014). Learning to Optimize via Posterior Sampling. *Mathematics of Operations Research*, **39** (4), 1221–1243.

Schneider, R. (1993). *Convex Bodies: The Brunn-Minkowski Theory*. Cambridge University Press.

Scott, S. L. (2010). A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, **26**, 639–658.

— (2012). Google Analytics Help: Multi-armed bandit experiments.

Stiratelli, R., Laird, N. and Ware, J. H. (1984). Random-Effects Models for Serial Observations with Binary Response. *Biometrics*, **40** (4), 961–971.

Stucke, M. and Grunes, A. (2016). *Big Data and Competition Policy*. Oxford University Press.

Sundaram, R. K. (2005). *Generalized Bandit Problems*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 131–162.

Tambe, P. (2014). Big Data Investment, Skills, and Firm Value. *Management Science*, **60** (6), 1452–1469.

Tang, L., Rosales, R., Singh, A. and Agarwal, D. (2013). Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, New York, NY, USA: ACM, pp. 1587–1594.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**, 285–294.

Ursu, R. (2016). *The Power of Rankings: Quantifying the Effect of Rankings on Online Consumer Search and Purchase Decisions*. Tech. rep.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Wall Street Journal (2015). Memorandum from FTC Bureau of Competition Staff to the Commission on Google Inc., File No. 111-0163.

Whittle, P. (1980). Multi-Armed Bandits and the Gittins Index. *Journal of the Royal Statistical Society. Series B (Methodological)*, **42** (2), 143–149.

— (1988). Restless Bandits: Activity Allocation in a Changing World. *Journal of Applied Probability*, **25** (A Celebration of Applied Probability), 287–298.

Yildiz, N. (2012). Consistency Of Plug-In Estimators Of Upper Contour And Level Sets. *Econometric Theory*, **28** (2), 309–327.

# Appendix A

# Appendix to Chapter 2

## A.1 Proofs

### A.1.1 Click-through rate maximized by assortative matching

**Proposition 10.** *For the click model given in (2.2), with* $Y_j = Y_j^{(1)} \wedge Y_j^{(2)}$, $Y_j^{(1)} \perp\!\!\!\perp Y_j^{(2)}$, $\mathbb{E}\left[Y_j^{(1)}\right] = \gamma_{\rho(j)}$ *for an ordering* $\rho$ *and* $\gamma \in \mathbb{R}_+^J$ *s.t.* $\gamma_{j'} > \gamma_{j''} \Leftrightarrow j' < j''$, *and* $\mathbb{E}_\mu\left[Y_j^{(2)}\right] = \mathbb{E}_\mu\left[\mu_j\right]$, *if additionally* $Y_{j'}^{(r_1)} \perp\!\!\!\perp Y_{j''}^{(r_2)}$ *for all* $j' \neq j''$ *and* $r_1,\ r_2 \in \{1,2\}$, *the expected click-through rate* $CTR(\rho) = \mathbb{E}_\mu\left[\mathbf{1}\left\{\sum_j Y_j > 0\right\}\right]$ *is maximized by sorting products in descending order of quality beliefs, i.e.,* $\rho(j') < \rho(j'') \Leftrightarrow \mathbb{E}_\mu\left[\mu_{j'}\right] > \mathbb{E}_\mu\left[\mu_{j''}\right]$.

*Proof.* The proof proceeds by contradiction. The click-through rate $CTR(\rho)$ of a product ordering $\rho$ is given by

$$\begin{aligned} CTR(\rho) &= 1 - \Pr\left(Y_j = 0 \forall j\right) \\ &= 1 - \prod_j \left(1 - \gamma_{\rho(j)}\mu_j\right) \end{aligned}$$

Now, suppose that $CTR(\rho)$ is maximized at $\rho = \rho^*$ where $\rho^*(j') > \rho^*(j'')$ even though

$\gamma_{j'} > \gamma_{j''}$ for some $j'$, $j''$. Then,

$$CTR\left(\rho^*\right) = 1 - \left(1 - \gamma_{\rho^*(j')}\mu_{j'}\right)\left(1 - \gamma_{\rho^*(j'')}\mu_{j''}\right)\left[\prod_{j \notin \{j', j''\}}\left(1 - \gamma_{\rho^*(j)}\mu_j\right)\right]$$

$$= 1 - \left[1 - \underbrace{\left(\gamma_{\rho^*(j')}\mu_{j'} + \gamma_{\rho^*(j'')}\mu_{j''}\right)}_{=\mathbb{E}\left[Y_{j'} + Y_{j''}\right]} + \gamma_{\rho^*(j')}\mu_{j'}\gamma_{\rho^*(j'')}\mu_{j''}\right]\left[\prod_{j \notin \{j', j''\}}\left(1 - \gamma_{\rho^*(j)}\mu_j\right)\right]$$

We know that $\mathbb{E}\left[Y_{j'} + Y_{j''}\right]$ is larger under an ordering $\rho^{**}$ that sets $\rho^{**}\left(j\right) = \rho^*\left(j\right)$ for all $j \notin \{j', j''\}$, $\rho^{**}\left(j'\right) = \rho^*\left(j''\right)$, and $\rho^{**}\left(j''\right) = \rho^*\left(j'\right)$. It is also obvious that $CTR\left(\rho^*\right)$ and $CTR\left(\rho^{**}\right)$ are different only through $\mathbb{E}\left[Y_{j'} + Y_{j''}\right]$, and that thus, $CTR\left(\rho^*\right) < CTR\left(\rho^{**}\right)$. $\square$

# Appendix B

# Appendix to Chapter 3

## B.1 Proofs

### B.1.1 Additional notation

For $C \in K_{kc}\left(\mathbb{R}^d\right)$, let $\text{reg}\,(C)$ denote the set of regular (or smooth) boundary points of $C$ (Schneider, 1993, p. 73). Let $\text{Gr}\,(k,d)$ denote the Grassmannian which parametrizes all linear subspaces of dimension $k$ in $\mathbb{R}^d$.

### B.1.2 Propositions and lemmata stated in the text

*Proof of Proposition 4.* The proof establishes that $\hat{\tilde{\Theta}}_r^I$ is a consistent estimator for $\tilde{\Theta}_r^I$. The claim for $\hat{\Theta}_r^I = \hat{\tilde{\Theta}}_r^I \cap \Theta$ follows from Lemma 13 and Assumption 3. By Lemma 11, $\tilde{\Theta}_r^I = \times_{k=1}^K \text{conv}\left(\{E_k^m\}_{m=1}^{|I_k|!}\right) = \text{conv}\left(\times_{k=1}^K \{E_k^m\}_{m=1}^{|I_k|!}\right)$ with sample counterpart $\hat{\tilde{\Theta}}_r^I = \text{conv}\left(\times_{k=1}^K \left\{\hat{E}_k^m\right\}_{m=1}^{|I_k|!}\right)$. Denote the elements of $\left\{\times_{k=1}^K \{E_k^m\}_{m=1}^{|I_k|!}\right\}$ by $E_s$ with sample counterparts $\hat{E}_s$. By Lemma 12, $\hat{E}_s \xrightarrow{p} E_s$ for all $s$. For all $a \in \tilde{\Theta}_r^I$, there exists a vector $\lambda \geq 0$ s.t. $\sum_s \lambda_s = 1$ and $\sum_s \lambda_s E_s = a$. Then, $\|\sum_s \lambda_s E_s - \sum_s \lambda_s \hat{E}_s\| = \|\sum_s \lambda_s \left(\hat{E}_s - E_s\right)\| \leq \sum_s \lambda_s \|\hat{E}_s - E_s\| \leq \max_s \|\hat{E}_s - E_s\|$. Hence, $d\left(a, \hat{\tilde{\Theta}}_r^I\right) \leq \max_s \|\hat{E}_s - E_s\|$, and so, $\sup_{a \in \tilde{\Theta}_r} d\left(a, \hat{\tilde{\Theta}}_r^I\right) \leq \max_s \|\hat{E}_s - E_s\|$. By a similar argument, $\sup_{a \in \hat{\tilde{\Theta}}_r} d\left(a, \tilde{\Theta}_r^I\right) \leq \max_s \|\hat{E}_s - E_s\|$. Since $\{E_s\}_s$ is a finite set, $\max_s \|\hat{E}_s - E_s\| \xrightarrow{p} 0$, and so $\hat{\tilde{\Theta}}_r^I \xrightarrow{H} \tilde{\Theta}_r^I$.

*Proof of Proposition* 6. Every (scalar) element of the stacked vector of permutations of

extreme points, $\mu$, can be written as $E = \left(\underbrace{\alpha_2 - \alpha_1}_{\equiv \Delta}\right)^{-1} \int_{\alpha_1}^{\alpha_2} Q(\tau)\, d\tau$ for some $0 \leq \alpha_1 < \alpha_2 \leq 1$ and a quantile function $Q : [0,1] \to \mathbb{R}$. (In our problem, fractions $\alpha_1$ and $\alpha_2$ are given by $\tau_1^{k,\delta(l),m}$ and $\tau_2^{k,\delta(l),m}$ under permutation $m$, and $Q(\cdot)$ is given by $Q_{Y|V\in I_k}(\cdot)$.) All of these objects have direct sample analogues, which leads to our estimator $\hat{E}$. We can write $E = \int_0^1 J(\tau) Q(\tau)\, d\tau$ for $J : [0,1] \to \mathbb{R}_+$ given by $J(\tau) = \Delta^{-1}\mathbf{1}\{\alpha_1 \leq \tau \leq \alpha_2\}$, again with straightforward sample analogue. Then,

$$\begin{aligned}\hat{E} - E &= \int_0^1 \hat{J}(\tau)\hat{Q}(\tau)\, d\tau - \int_0^1 J(\tau) Q(\tau)\, d\tau \\ &= \int_0^1 \left[\left(\hat{J}(\tau) - J(\tau)\right)\left(\hat{Q}(\tau) - Q(\tau)\right) + J(\tau)\left(\hat{Q}(\tau) - Q(\tau)\right) + \left(\hat{J}(\tau) - J(\tau)\right) Q(\tau)\right] d\tau.\end{aligned}$$

For the first term, we have that $\sqrt{N_1}\int_0^1 \underbrace{\left(\hat{J}(\tau) - J(\tau)\right)}_{o(1)}\underbrace{\left(\hat{Q}(\tau) - Q(\tau)\right)}_{o(1)} d\tau = o(1)$. For the second term, under Assumption 5, Koenker and Portnoy (1987, Thm. 2.1) provide the Bahadur representation

$$\begin{aligned}&\sqrt{N_1} J(\tau)\left(\hat{Q}(\tau) - Q(\tau)\right) \\ &= \frac{1}{\sqrt{N_1}} \frac{J(\tau)}{f_Y(Q(\tau))} \sum_{i=1}^{N_1} \left(\tau - \mathbf{1}\{Y_i \leq Q(\tau)\}\right) + O\left(N_1^{-1/4}\log N_1\right)\end{aligned} \tag{B.1}$$

uniformly for $\tau \in [\epsilon, 1-\epsilon]$ for any $\epsilon > 0$. Note that $(\tau - \mathbf{1}\{Y_i \leq Q(\tau)\})$ is bounded and has zero mean. For the tail quantiles, $\tau \in [0,\epsilon) \cup (1-\epsilon, 1]$, a standard truncation argument applies under finite variance. The integral of (B.1) over the unit interval has mean zero and finite variance as given in Koenker and Portnoy (1987, Thm. 3.1). For the third term,

$$\begin{aligned}&\int_0^1 \left(\hat{J}(\tau) - J(\tau)\right) Q(\tau)\, d\tau \\ &= \left(\Delta\hat{\Delta}\right)^{-1}\left(\Delta - \hat{\Delta}\right)\int_{\alpha_1}^{\alpha_2} Q(\tau)\, d\tau + \hat{\Delta}^{-1}\left(\int_{\alpha_2}^{\hat{\alpha}_2} Q(\tau)\, d\tau + \int_{\hat{\alpha}_1}^{\alpha_1} Q(\tau)\, d\tau\right).\end{aligned}$$

We next define a discrete random variable $Z_j$ with support $\{a, b, c\}$ and distribution $(\alpha_1, \Delta, 1-\alpha_2)'$. Specifically, $Z_j$ maps the realization $V_j$ into three events, depending on the permutation considered: looking at support point $\gamma_l$ under permutation $m$ of the corresponding interval $I_k$, $Z_j = a$ if $\phi_k^m(V_j) < \phi_k^m(\gamma_l)$, i.e., $V_j$ belongs to the predecessors of $\gamma_l$, $Z_j = b$ if $V_j = \gamma_l$, and $Z_j = c$ if $\phi_k^m(V_j) > \phi_k^m(\gamma_l)$. Then, $\hat{\alpha}_1 = \frac{1}{N_2}\sum_{j=1}^{N_2} 1\{Z_j = a\}$, $\hat{\alpha}_2 = \frac{1}{N_2}\sum_{j=1}^{N_2} 1\{Z_j \in \{a,b\}\}$, and $\hat{\Delta} = \frac{1}{N_2}\sum_{j=1}^{N_2} 1\{Z_j = b\}$. Noting that $Z_j$ is iid since $V_j$ is

iid, this yields

$$\sqrt{N_2}\left(\Delta\hat{\Delta}\right)^{-1}\left(\Delta-\hat{\Delta}\right)\int_{\alpha_1}^{\alpha_2}Q\left(\tau\right)d\tau=\frac{1}{\sqrt{N_2}}\sum_{j=1}^{N_2}\Delta^{-2}\left(\int_{\alpha_1}^{\alpha_2}Q\left(\tau\right)d\tau\right)\left(\Delta-1\left\{Z_j=b\right\}\right)+o\left(1\right)$$
$$\sqrt{N_1}\left(\Delta\hat{\Delta}\right)^{-1}\left(\Delta-\hat{\Delta}\right)\int_{\alpha_1}^{\alpha_2}Q\left(\tau\right)d\tau=\frac{\kappa}{\sqrt{N_2}}\sum_{j=1}^{N_2}\Delta^{-2}\left(\int_{\alpha_1}^{\alpha_2}Q\left(\tau\right)d\tau\right)\left(\Delta-1\left\{Z_j=b\right\}\right)+o\left(1\right),\tag{B.2}$$

where $\left(\Delta-1\left\{Z_j=b\right\}\right)$ is bounded and has zero mean, and the second line defines the asymptotics in $N_1$ with $\sqrt{\frac{N_1}{N_2}}\to\kappa$. By the mean value theorem, $\int_{\alpha_2}^{\hat{\alpha}_2}Q\left(\tau\right)d\tau=\left(\hat{\alpha}_2-\alpha_2\right)Q\left(\alpha\right)$ for some $\alpha$ s.t. $\left(\alpha-\hat{\alpha}_2\right)\left(\alpha-\alpha_2\right)\leq 0$, and similarly for the second integral. Hence,

$$\begin{aligned}&\sqrt{N_1}\hat{\Delta}^{-1}\left(\int_{\alpha_2}^{\hat{\alpha}_2}Q\left(\tau\right)d\tau+\int_{\hat{\alpha}_1}^{\alpha_1}Q\left(\tau\right)d\tau\right)\\=&\frac{\kappa}{\sqrt{N_2}}\sum_{j=1}^{N_2}\Delta\left[Q\left(\alpha_1\right)\left(\alpha_1-1\left\{Z_j=a\right\}\right)+Q\left(\alpha_2\right)\left(1\left\{Z_j\in\left\{a,b\right\}\right\}-\alpha_2\right)\right]+o\left(1\right),\end{aligned}\tag{B.3}$$

where each element of the iid sum is bounded and has zero mean. In case $\alpha_1=0$ or $\alpha_2=1$, the respective quantiles are not necessarily well-defined, but their respective coefficients are zero with probability one. Adding up the integral of (B.1) over all quantiles and terms (B.2) and (B.3), the claim follows from the central limit theorem.

In the one-sample case, we can write $i$ and $N_1$ instead of $j$ and $N_2$ everywhere, with $\kappa=1$. Convergence in total variation generally fails because unless $\left(\alpha_1,\alpha_2\right)'=\left(0,1\right)'$, each of the three elements has a distribution with point mass, and so the characteristic function does not vanish.

*Proof of Proposition 7.* The same decomposition obtains as in the proof of Proposition 6, only now, each term of the sum in (B.1) is weighted by $\left(W_{i,b}-1\right)$, while each term of the sums in (B.2) and (B.3) is weighted by $\left(W_{j,b}-1\right)$, where $W_{i,b}$ and $W_{j,b}$ are the bootstrap weights on $i$ and $j$ in bootstrap draw $b$, all iid draws from an $Exp\left(1\right)$ distribution independent of $\mathbf{X}$, so that $\left(W_{\cdot,b}-1\right)$ has mean zero and variance one. In the two-sample case, the variances and covariances of all terms are therefore identical to those in the derivation for the former proof. In the one-sample case, however, since $W_{i,b}\perp W_{j,b}$ even when $i=j$, i.e., the indices refer to the same unit, the covariance between (B.1) on the one hand and (B.2) and (B.3) on the other does not feature in the large-sample posterior, and therefore $\tilde{\Sigma}_{0,1}\neq\Sigma_{0,1}$. Since

the exponential draws provide smoothing, the characteristic function now vanishes, and so, a central limit theorem in total variation (van der Vaart, 2000, Prop. 2.31) applies.

*Proof of Lemma 8.* We first explicitly characterize the structure of the set, solving for $\mu \in \mathbb{R}^{d_\mu}$ as the coordinate vector of the set:

$$
\begin{aligned}
&\sqrt{N_1}\left(\cap_{\delta \in \left(\mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}(\mathbf{X})\right)^C} \cap_{\{r:\, L(r)=\delta\}} \mu_I\left(r\right)^C - \hat{\mu}\right) \\
=&\sqrt{N_1}\left(\left\{\mu : \max_{r \in \operatorname{conv}\left(\times_k \left\{\mu^{(k,m)}\right\}_m\right)} c'r \leq \sup \mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}(\mathbf{X}),\right.\right. \\
&\qquad \min_{r \in \operatorname{conv}\left(\times_k \left\{\mu^{(k,m)}\right\}_m\right)} c'r \geq \inf \mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}(\mathbf{X}), \\
&\qquad \left.\left.\operatorname{conv}\left(\times_k \left\{\mu^{(k,m)}\right\}_m\right) \cap \Theta \neq \emptyset\right\} \cup \left\{\mu :\, \operatorname{conv}\left(\times_k \left\{\mu^{(k,m)}\right\}_m\right) \cap \Theta = \emptyset\right\} - \hat{\mu}\right) \\
=&\left\{\mu : \max_{r \in \operatorname{conv}\left(\times_k \left\{\mu^{(k,m)}\right\}_m\right)} c'r \leq \sqrt{N_1}\left[\sup \mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}(\mathbf{X}) - c'\hat{\mu}\right],\right. \\
&\qquad \min_{r \in \operatorname{conv}\left(\times_k \left\{\mu^{(k,m)}\right\}_m\right)} c'r \geq \sqrt{N_1}\left[\inf \mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}(\mathbf{X}) - c'\hat{\mu}\right], \\
&\qquad \left.\operatorname{conv}\left(\times_k \left\{\mu^{(k,m)}\right\}_m\right) \cap \sqrt{N_1}\left(\Theta - \hat{\mu}\right) \neq \emptyset\right\} \\
&\cup \left\{\mu :\, \operatorname{conv}\left(\times_k \left\{\mu^{(k,m)}\right\}_m\right) \cap \sqrt{N_1}\left(\Theta - \hat{\mu}\right) = \emptyset\right\}
\end{aligned}
\tag{B.4}
$$

Hence, we have that

$$
\begin{aligned}
&\Pr_{\mathcal{N}(0,\Sigma_0)}\left(\operatorname{bnd}\left(\sqrt{N_1}\left(\cap_{\delta\in\left(\mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}(\mathbf{X})\right)^C\cap_{\{r:\,L(r)=\delta\}}\mu_I\left(r\right)^C}-\hat{\mu}\right)\right)\right)\\
&=\Pr_{\mathcal{N}(0,\Sigma_0)}\left(\left\{\mu:\ \max_{r\in\operatorname{conv}\left(\times_k\left\{\mu^{(k,m)}\right\}_m\right)}c'r=\sqrt{N}_1\left[\sup\mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}\left(\mathbf{X}\right)-c'\hat{\mu}\right],\right.\right.\\
&\qquad\min_{r\in\operatorname{conv}\left(\times_k\left\{\mu^{(k,m)}\right\}_m\right)}c'r=\sqrt{N}_1\left[\inf\mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}\left(\mathbf{X}\right)-c'\hat{\mu}\right],\\
&\qquad\left.\operatorname{conv}\left(\times_k\left\{\mu^{(k,m)}\right\}_m\right)\cap\sqrt{N}_1\left(\Theta-\hat{\mu}\right)\neq\emptyset\right\}\\
&\qquad\cup\left\{\mu:\ \operatorname{conv}\left(\times_k\left\{\mu^{(k,m)}\right\}_m\right)\cap\operatorname{bnd}\left(\sqrt{N}_1\left(\Theta-\hat{\mu}\right)\right)\neq\emptyset,\right.\\
&\qquad\left.\left.\operatorname{conv}\left(\times_k\left\{\mu^{(k,m)}\right\}_m\right)\cap\operatorname{int}\left(\sqrt{N}_1\left(\Theta-\hat{\mu}\right)\right)=\emptyset\right\}\right)\\
&\leq\Pr_{\mathcal{N}(0,\Sigma_0)}\left(\left\{\mu:\ \max_{r\in\operatorname{conv}\left(\times_k\left\{\mu^{(k,m)}\right\}_m\right)}c'r=\sqrt{N}_1\left[\sup\mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}\left(\mathbf{X}\right)-c'\hat{\mu}\right],\right.\right.\\
&\qquad\min_{r\in\operatorname{conv}\left(\times_k\left\{\mu^{(k,m)}\right\}_m\right)}c'r=\sqrt{N}_1\left[\inf\mathcal{C}_{1-\alpha}^{L\left(\Theta_r^I\right)}\left(\mathbf{X}\right)-c'\hat{\mu}\right],\\
&\qquad\left.\left.\operatorname{conv}\left(\times_k\left\{\mu^{(k,m)}\right\}_m\right)\cap\sqrt{N}_1\left(\Theta-\hat{\mu}\right)\neq\emptyset\right\}\right)\\
&\quad+\Pr_{\mathcal{N}(0,\Sigma_0)}\left(\left\{\mu:\ \operatorname{conv}\left(\times_k\left\{\mu^{(k,m)}\right\}_m\right)\cap\operatorname{bnd}\left(\sqrt{N}_1\left(\Theta-\hat{\mu}\right)\right)\neq\emptyset,\right.\right.\\
&\qquad\left.\left.\operatorname{conv}\left(\times_k\left\{\mu^{(k,m)}\right\}_m\right)\cap\operatorname{int}\left(\sqrt{N}_1\left(\Theta-\hat{\mu}\right)\right)=\emptyset\right\}\right)\\
&=\Pr_{\mathcal{N}(0,\Sigma_0)}\left(\left\{\mu:\ \operatorname{conv}\left(\times_k\left\{\mu^{(k,m)}\right\}_m\right)\cap\operatorname{bnd}\left(\sqrt{N}_1\left(\Theta-\hat{\mu}\right)\right)\neq\emptyset,\right.\right.\\
&\qquad\left.\left.\left(\operatorname{conv}\left(\times_k\left\{\mu^{(k,m)}\right\}_m\right)\cap\operatorname{int}\left(\sqrt{N}_1\left(\Theta-\hat{\mu}\right)\right)=\emptyset\right)\right\}\right)
\end{aligned}
$$

The final line follows from the fact that the probability for any of the extreme points to lie in the level plane is zero, since they individually follow a non-singular normal distribution. Next, Lemma 15 for absolutely continuous distributions establishes the claim for the convex hull of a set of extreme points in which no pair of extreme points corresponds to the same permutation on any interval, and further, $\kappa > 0$ or $\bar{\pi}_s^{(k)} \neq \bar{\pi}_{s'}^{(k)}$ for all $s \neq s'$ and $k = 1, \ldots, K$,

since then, the joint distribution of the extreme points is non-singular.

Finally, we discuss the cases in which the joint distribution of the extreme points is singular. Choose a set of $L$ extreme points from $\times_k \left\{\mu^{(k,m)}\right\}_m$. Carathéodory's theorem (Rockafellar, 1970, Thm. 17.1) implies that for a plane in $\mathbb{R}^L$, which is of dimension $L-1$ itself, we only need to consider the convex hull of at most $L$ points.

It is possible that two or more of them are identical on some interval because they correspond to the same permutation of support points on that interval. If all $L$ points have an identical subvector corresponding to some interval, then the convex hull of these points has this subvector fixed and lies in the corresponding plane. The distance of this plane from the origin in each of the directions which are fixed follows a normal distribution, so it will be tangent to $\sqrt{N_1}\left(\Theta - \hat{\mu}\right)$ with probability zero. Now, suppose that only a subset of extreme points shares a permutation. Then, only that subset of the convex hull of the $L$ extreme points which puts weight only on the subset of extreme points with shared permutation has a singular distribution, but the same argument just given applies.

If $\kappa = 0$ and $\bar{\pi}_s^{(k)} = \bar{\pi}_{s'}^{(k)}$ for some $s \neq s'$, $k \in \{1, \ldots, K\}$, there will be an $m \neq m'$, $r \neq r'$ such that $\mu_r^{(k,m)}$ and $\mu_{r'}^{(k,m')}$ are equal with probability one. Similar to above, this fixes a subvector of the normal vector of the hyperplane defined by the extreme points, but the location remains random and follows an absolutely continuous distribution.

### B.1.3 Additional lemmata

**Lemma 11.** *Let $A = conv(\{a_m\}_m)$ and $B = conv(\{b_n\}_n)$ for $a_m$, $b_n \in \mathbb{R}^d$ for all $m$, $n$. Then, $A \times B = conv(\{a_m\}_m \times \{b_n\}_n)$.*

*Proof.* One direction of inclusion, $A \times B \supseteq \text{conv}\left(\{a_m\}_m \times \{b_n\}_n\right)$, is obvious. The other direction, $A \times B \subseteq \text{conv}\left(\{a_m\}_m \times \{b_n\}_n\right)$, can be shown as follows. If $a \in A$ and $b \in B$, then $a = \sum_m \lambda_m a_m$ and $b = \sum_n \tilde{\lambda}_n b_n$ for some $\lambda$, $\tilde{\lambda}$ with $\lambda$, $\tilde{\lambda} \geq 0$ and $\sum_m \lambda_m = 1$, $\sum_n \tilde{\lambda}_n = 1$. Then, $(a', b')' = \sum_{m,n} \lambda_m \tilde{\lambda}_n \left(a_m', b_n'\right)' \in \text{conv}\left(\{a_m\}_m \times \{b_n\}_n\right)$, where $\lambda_m \tilde{\lambda}_n \geq 0$ for all $m$, $n$ and $\sum_{m,n} \lambda_m \tilde{\lambda}_n = 1$. □

**Lemma 12.** *Suppose $a_n \xrightarrow{p} a$ and $b_n \xrightarrow{p} b$. Then, $(a_n', b_n')' \xrightarrow{p} (a', b')'$.*

*Proof.* Note that $\|(a_n', b_n')' - (a', b')'\| \leq 2 \cdot \max(\|a_n - a\|, \|b_n - b\|)$ by the triangle inequality. Now, $\Pr(\max(\|a_n - a\|, \|b_n - b\|) \geq \epsilon) \leq \Pr(\|a_n - a\| \geq \epsilon/2) + \Pr(\|b_n - b\| \geq \epsilon/2)$, and the two elements of the sum each converge to zero as $n \to \infty$. □

**Lemma 13.** *Let $A, B \in K_{kc}(\mathbb{R}^d)$ be compact and convex, and suppose $A \cap int(B) \neq \emptyset$. Then, for a consistent estimator $\hat{A} \xrightarrow{H} A$, we have that $\hat{A} \cap B \xrightarrow{H} A \cap B$.*

*Proof.* Fix $\epsilon > 0$. Consider a $\delta$-contraction of $A \cap B$, $(A \cap B)^{-\delta}$, where $\delta$ is chosen s.t. $(A \cap B)^{-\delta} \subseteq \text{int}(B)$ and for all $a \in A \cap B$, $d\left(a, (A \cap B)^{-\delta}\right) < \epsilon/2$. Since $B$, by assumption, has a non-empty interior and thus, $\dim(B) = d$, there exists an $\eta < \epsilon/2$ such that for an $\eta$-neighboorhood of this contraction, $N_\eta\left((A \cap B)^{-\delta}\right) \subseteq \text{int}(B)$. Since $\hat{A} \xrightarrow{H} A$, $\lim_{n\to\infty} \Pr\left(\hat{A} \cap N_\eta\left((A \cap B)^{-\delta}\right) = \emptyset\right) = 0$, implying $\lim_{n\to\infty} \Pr\left(\hat{A} \cap B = \emptyset\right) = 0$. Using the fact that $\hat{A} \xrightarrow{H} A$ along with the triangle inequality, $\lim_{n\to\infty} \Pr\left(\sup_{c \in N_\eta\left((A\cap B)^{-\delta}\right)} d\left(c, \hat{A} \cap B\right) > \epsilon/2\right) = 0$, and by the choice of $\delta$ and the triangle inequality again, $\lim_{n\to\infty} \Pr\left(\sup_{c\in(A\cap B)} d\left(c, \hat{A} \cap B\right) > \epsilon\right) = 0$. In addition, $\hat{A} \xrightarrow{H} A$ implies that $\lim_{n\to\infty} \Pr\left(\sup_{c \in \left(\hat{A}\cap B\right)} d(c, A \cap B) > \epsilon\right) = 0$. Hence, $\hat{A} \cap B \xrightarrow{H} A \cap B$. □

**Lemma 14.** *Suppose $X, Y \in \mathbb{R}^d$, $d \geq 1$, are random variables with $\mathbb{E}[X^2] < \infty$, $\mathbb{E}[Y^2] < \infty$. Then, $2(Var(X) + Var(Y)) \geq Var(X + Y)$.*

*Proof.* For two matrices $A$ and $B$ of equal size, $A - B \geq 0 \Leftrightarrow A \geq B$ denotes that the matrix $A - B$ is positive semi-definite. Now, suppose, w.l.o.g., that $\mathbb{E}(X) = \mathbb{E}(Y) = 0$. The covariance matrix of $X - Y$ is positive semi-definite:

$$\begin{aligned}
\mathbb{E}\left[(X - Y)(X - Y)'\right] &\geq 0 \\
\mathbb{E}[XX'] + \mathbb{E}[YY'] &\geq \mathbb{E}[XY'] + \mathbb{E}[YX'] \\
2\{\mathbb{E}[XX'] + \mathbb{E}[YY']\} &\geq \mathbb{E}[XX'] + \mathbb{E}[XY'] + \mathbb{E}[YX'] + \mathbb{E}[YY'] \\
2\left\{\underbrace{\mathbb{E}[XX'] + \mathbb{E}[YY']}_{Var(X)+Var(Y)}\right\} &\geq \underbrace{\mathbb{E}\left[(X + Y)(X + Y)'\right]}_{Var(X+Y)}
\end{aligned}$$

□

**Lemma 15.** *Let $Z \sim P$ be a random variable in $\mathbb{R}^{d \cdot s}$, $d \in \mathbb{Z} > 0$, $s \in \mathbb{Z} > 0$. Suppose $P$ is an absolutely continuous distribution such that $\|\mathbb{E}(Z)\|_1 < \infty$, $\|Var(Z)\|_1 < \infty$, and $Var(Z)$ is non-singular. Decompose $Z$ as $Z = (E_1', \ldots, E_s')'$, where $\dim(E_r) = d$ for $r = 1 \ldots, s$, and let $C \in K_{kc}\left(\mathbb{R}^d\right)$. Then, $\Pr(conv(E_1, \ldots, E_s) \cap C \neq \emptyset,\ conv(E_1, \ldots, E_s) \cap int(C) = \emptyset) = 0$.*

*Proof.* Carathéodory's theorem (Rockafellar, 1970, Thm. 17.1) implies that for a plane in $\mathbb{R}^d$, which is of dimension $d-1$ itself, we only need to consider the convex hull of at most $d$ points, so w.l.o.g., suppose that $s \leq d$. The joint distribution of $s$ points in $\mathbb{R}^d$ is defined on $\mathbb{R}^{s \cdot d}$, in which objects of dimension less than $s \cdot d$ have Lebesgue measure zero, so that we can appeal to absolute continuity to conclude that the corresponding event has probability zero.

Define $C' = \text{reg}(C)$, the set of points in $\text{bnd}(C)$ with unique tangent hyperplane. In other words, $C'$ is the boundary of $C$ minus the edges and vertices, which we treat separately below. Note that $\dim(C') \leq d-1$.

For $s = 1$, the result is obvious since $C'$ is of dimension at most $d-1$.

For $s = d$, the union of all tangent hyperplanes is isomorphic to $C' \times \mathbb{R}^{\dim(C')}$, and all convex hulls of $s$ points which are tangent to $C$ lie in a manifold of dimension $(d+1)\dim(C') < d^2$: for every point in $C'$, there is a unique hyperplane tangent to $C$ of dimension $\dim(C')$, in which we place $d$ points, resulting in a total of $\dim(C') + d \cdot \dim(C') = (d+1)\dim(C')$ dimensions; this holds even when $C$ has flat faces.

For $1 < s < d$, we again have a tangent hyperplane of dimension $\dim(C')$ at every point. We further suppose that $\dim(C') > s-1$, which is always true if $\dim(C) = d$; lower-dimensional cases are nested. Each hyperplane contains the Grassmannian $\text{Gr}(s-1, \dim(C'))$ recentered at the tangency point, which is of dimension $(s-1)[\dim(C') - (s-1)]$ (e.g., Kolhatkar, 2004, cor. 1.15). We place $s$ points of dimension $s-1$ in each of the planes in

$\mathrm{Gr}\left(s-1, \dim\left(C'\right)\right)$. Adding up dimensions,

$$\begin{aligned}
\dim\left(C'\right) + \left[\dim\left(C'\right) - (s-1)\right](s-1) + s\,(s-1) &= \dim\left(C'\right) + \left[\dim\left(C'\right) + 1\right](s-1) \\
&= \left[\dim\left(C'\right) + 1\right] s - 1 \\
&\leq s \cdot d - 1 \\
&< s \cdot d.
\end{aligned}$$

Consider now the non-smooth singular edges and vertices of $C$, at which the tangent hyperplane is not unique. Since the set of $r$-singular points of $C$ can be covered by countably many sets of finite $r$-dimensional measure (Schneider, 1993, Thm. 2.2.4), and every $r$-singular point has a space of tangent planes of dimension $d-r-1$, we get $r + (d-r-1) = d-1$, so that the above arguments apply one-to-one. □